Perfect samplers for mixtures of distributions

G. Casella,
University of Florida, Gainesville, USA

K. L. Mengersen,
Queensland University of Technology, Brisbane, Australia

C. P. Robert
Centre de Recherche en Economie et Statistique, Paris, and Université Dauphine, Paris, France

and D. M. Titterington
University of Glasgow, UK

[Received September 1999. Final revision March 2002]

Summary. We consider the construction of perfect samplers for posterior distributions associated with mixtures of exponential families and conjugate priors, starting with a perfect slice sampler in the spirit of Mira and co-workers. The methods rely on a marginalization akin to Rao–Blackwellization and illustrate the duality principle of Diebolt and Robert. A first approximation embeds the finite support distribution on the latent variables within a continuous support distribution that is easier to simulate by slice sampling, but we later demonstrate that the approximation can be very poor. We conclude by showing that an alternative perfect sampler based on a single backward chain can be constructed. This alternative can handle much larger sample sizes than the slice sampler first proposed.

Keywords: Coupling; Gibbs sampling; Marginalization; Monotonicity; Slice sampling

1. Introduction

Perfect sampling has been developed in recent years to take advantage of Markov chain Monte Carlo (MCMC) algorithms without suffering from the drawback of MCMC methods, namely that the distribution of interest is only the asymptotic distribution of the Markov chain generated. See Dimakos (2001) for an excellent introduction, as well as Wilson (1998), whose Web site http://dimacs.rutgers.edu/~dbwilson/exact.html is constantly updated, and Green and Murdoch (1999) and Møller and Nicholls (1999) for recent statistical applications.

When considering realistic statistical models like those involving finite mixtures of distributions (Titterington et al., 1985), with densities of the form

\[
\sum_{i=1}^{k} p_i f(x \mid \theta_i),
\]

\[
\sum_{i=1}^{k} p_i = 1,
\]

Address for correspondence: C. P. Robert, Ceremade, Université Paris-Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France.
E-mail: Christian.Robert@ceremade.dauphine.fr

© 2002 Royal Statistical Society 1369–7412/02/64777
where the $p_i$s are all positive and the $f(\cdot|\theta_i)$ are density functions, MCMC algorithms are necessary for processing the posterior distribution of the parameters $\xi_i = (p_i, \theta_i)$ (see, for example, Celeux et al. (2000)). It is, however, quite a delicate exercise to produce a perfect sampling version, as shown by Hobert et al. (1999), who could only process a mixture like expression (1) when the parameters $\theta_i$ are known and when $k \leq 3$.

The reasons for this difficulty are that perfect sampling techniques, although not requiring monotonicity structures in the Markov transition, work better using monotonicity assumptions, i.e. the existence of an ordering ‘$\leq$’ on the state space that is preserved by the Markov transition (see Propp and Wilson (1996)). It requires hard work to establish monotonicity in the mixture model. One of the key features of the solution of Hobert et al. (1999) is to exploit the duality principle established by Diebolt and Robert (1994) for latent variable models. In set-ups where the chain of interest ($\xi^{(i)}$) is generated conditionally on a second chain ($Z^{(i)}$) with finite support, the probabilistic properties of ($\xi^{(i)}$) can be derived from the properties of ($Z^{(i)}$), whose finiteness facilitates theoretical study. For instance, if we assume that the component distributions in mixture (1) are from an exponential family, the latent variable that is associated with mixture model (1) and a sample $(X_1, \ldots, X_n)$ is the allocation vector $Z = (Z_1, \ldots, Z_n)$ such that, for $j = 1, \ldots, n$,

$$X_j, Z_j \sim p_z f(x|\theta_z) \mathbb{I}_{\{1, \ldots, k\}}(z). \quad (2)$$

This finite latent structure is not of direct practical relevance for perfect sampling, since the support of $Z$ is of size $k^n$ for $k$-component mixtures with $n$ observations, and thus precludes a direct implementation of the coupling from the past (CFTP) algorithm (Propp and Wilson, 1996). However, monotonicity structures can often be observed on the $Z^{(i)}$ chain.

This paper extends the result of Hobert et al. (1999) to the case of general finite mixtures of exponential family distributions, under conjugate priors, by proposing a different and more generic approach to the problem. The foundation of the technique used here relies on the facts that, under conjugate priors, the marginal posterior distribution of the latent variables $Z$ is known in closed form, up to a constant, as exhibited and exploited for importance sampling in Casella et al. (1999). Moreover, minimum and maximum points can be found for this distribution. The results of Mira et al. (2001) on perfect slice samplers can then be adapted to this setting. Although its practical implementation is limited to small sample sizes, we show that a coupling strategy of Breyer and Roberts (2001) can be easily implemented on the same principle for much larger sample sizes.

In Section 2, we use the marginalization argument of Casella et al. (1999) to show that the marginal posterior of $Z$ is available in closed form and allows for a minimum and a maximum point in common cases. Section 3 then provides a detailed description of the perfect slice sampling techniques, first via a naive uniform proposal and then through an approximating augmented distribution. Section 4 introduces a generic perfect sampler based on a single backward chain and evaluates the validity of the approximation of Section 3.

2. Marginalization arguments

Consider a random sample from a $k$-component exponential family mixture, with density

$$X_1, \ldots, X_n \sim \sum_{i=1}^{k} p_i \exp\{x \theta_i - \psi(\theta_i)\}. \quad (3)$$

Assume in addition that the $\theta_i$s are associated with conjugate priors, i.e.
\[ \theta_i \sim \pi(\theta_i) \propto \exp\{x_0\theta_i - \lambda_i \psi(\theta_i)\}, \]  

where \((x_0, \lambda_i)\) is the hyperparameter, and that the \(p_i\)s, when unknown, are associated with a Dirichlet \(D(\delta_1, \ldots, \delta_k)\) prior.

Model (3) can be represented as a latent variable model, i.e. it can be interpreted as the marginal distribution of the joint distribution (2). As previously noted in Casella et al. (1999), the corresponding marginal posterior distribution on the \(Z_j\)s is then

\[ Z \mid x \sim \prod_{i=1}^k \int p_i^{\delta_i+n_i-1} \exp\{(x_0i + s_i)\theta_i - (\lambda_i + n_i) \psi(\theta_i)\} \, dp_i \, d\theta_i, \]

where \(n_i\) denotes the number of \(z_j\)s equal to \(i\) and \(s_i\) is the sum of the corresponding \(x_j\)s,

\[ n_i = \sum_{j=1}^{n} I_{z_j=i}, \]

\[ s_i = \sum_{j=1}^{n} I_{z_j=i} x_j. \]

Taking advantage of the conjugate priors, we can then solve the above integrals and obtain

\[ Z \mid x \sim \prod_{i=1}^k \Gamma(\delta_i + n_i) \, K(\lambda_i + n_i, x_0i + s_i), \]

where \(K(\lambda, s)\) is the normalizing constant associated with expression (4).

The marginal posterior distribution of the latent variables \((Z_1, \ldots, Z_n)\) is therefore available in closed form. Since this is also true for the parameters of interest \(\xi_i\), one may wonder where the advantage of this representation is. The crucial points are as follows.

(a) Posterior expectations of functions of the \(\xi\)s can often be approximated from the distribution of the \(Z_j\)s by using Rao–Blackwellization.

(b) The \(Z_j\)s have a finite support, with size \(k^n\), and it is often possible to determine the minimum and maximum of expression (5). Moreover, the distribution factors through the (latent) sufficient statistics \((n_j, s_i)\).

(c) When possible, perfect simulation from expression (5) leads to perfect simulation from the posterior distribution of the \(\xi_j\)s by a single call to the conditional distribution \(\pi(\xi \mid x, z)\).

The constraint on this closed form representation is obviously that the prior distributions must be conjugate, but this is often the case in the literature (see Diebolt and Robert (1994) or Richardson and Green (1997)).

In what follows, two running examples will be presented.

### 2.1. Example 1: exponential mixtures

Consider the case of the two-component mixture

\[ p\lambda_0 \exp(-\lambda_0 x) + (1 - p)\lambda_1 \exp(-\lambda_1 x), \quad \lambda > 0, \]

where \(0 < p < 1\) is known. A conjugate prior distribution on \(\lambda_i\) is the \(Ga(\alpha_i, \beta_i)\) distribution. The marginal posterior distribution on the \(Z_j\)s is then
\[ Z \mid x \sim p_n^0 (1 - p)^n_1 \int_{-\lambda_0}^{\lambda_1} \left\{ \exp \left\{ -\lambda_0 (\beta_0 + s_0) \right\} \lambda_0^{\alpha_0 + n_0 - 1} \times \lambda_1^{\alpha_1 + n_1 - 1} \exp \left\{ -\lambda_1 (\beta_1 + s_1) \right\} d\lambda_0 d\lambda_1 \right\} \sim p_n^0 (1 - p)^n_1 \frac{\Gamma(\alpha_0 + n_0) \Gamma(\alpha_1 + n_1)}{(\beta_0 + s_0)^{\alpha_0 + n_0} (\beta_1 + s_1)^{\alpha_1 + n_1}}. \]

This distribution thus factors through the sufficient statistic \((n_0, s_0)\), since \(n_1 = n - n_0\) and \(s_1 = S - s_0\), where \(S\) is the sum of all observations. Therefore, the entire simulation problem is reduced to the distribution of the pair \((n_0, s_0)\). (Note that a similar expression can be obtained for a Poisson mixture with the same factorization.)

2.2. **Example 2: normal mixture**

We now consider a two-component normal mixture density with common variance \(\sigma^2\) and known weight \(p\),

\[ p \mathcal{N}(\mu_0, \sigma^2) + (1 - p) \mathcal{N}(\mu_1, \sigma^2), \]

under a conjugate prior distribution in which

\[
\begin{align*}
\mu_0 | \sigma^2 &\sim \mathcal{N}(\xi_0, \tau_0^2 \sigma^2), \\
\mu_1 | \sigma^2 &\sim \mathcal{N}(\xi_1, \tau_1^2 \sigma^2), \\
\sigma^{-2} &\sim \mathcal{G}(\alpha, \beta).
\end{align*}
\]

The marginal posterior distribution of \(Z\) is proportional to

\[
p_n^0 (1 - p)^n_1 \left\{ T - n_0 \bar{x}_0^2 - n_1 \bar{x}_1^2 + \beta + \frac{n_0 (\bar{x}_0 - \xi_0)^2}{1 + n_0 \tau_0^2} + \frac{n_1 (\bar{x}_1 - \xi_1)^2}{1 + n_1 \tau_1^2} \right\}^{-(\alpha+n)/2},
\]

where \(T\) denotes the total sum of squares and \(\bar{x}_j\) is the average of the observations allocated to component \(j\).

3. **The perfect slice sampler**

3.1. **Background**

We recall that the slice sampler, as developed in Besag and Green (1993) or Damien et al. (1999), consists of simulating a density \(\pi(\xi)\) by introducing an auxiliary variable \(\varepsilon \sim \mathcal{U}([0, \pi(\xi)])\), and generating a Markov chain according to the following transition kernel.

**Step 1:** generate \(U \sim \mathcal{U}([0, 1])\) and take \(\varepsilon = U \pi(\xi)\).

**Step 2:** generate \(\xi \sim \mathcal{U}(\{\xi; \pi(\xi) \geq \varepsilon\})\).

Step 1 of this algorithm is straightforward but step 2 can be quite complex. For instance, in the mixture problem, generating uniformly a point in the region

\[
\left\{ (p, \theta); \prod_{j=1}^k \sum_{i=1}^p p_i f(x_j | \theta_i) \theta_i \geq \varepsilon \right\},
\]

where \(\theta = (\theta_1, \ldots, \theta_k)\) and \(p = (p_1, \ldots, p_k)\), is impossible in practice given that this region is usually not connected. As detailed in Section 3.2, it is also difficult to generate a point at random in \(\{z; \pi(z) \geq \varepsilon\}\), given the finite support of \(Z\).
A very appealing feature of slice samplers is that, as pointed out in Mira et al. (2001), there is a natural stochastic ordering, which is induced by $\pi(\cdot)$, and hence monotonicity arguments can be invoked, as we now detail.

First, if $\pi(\xi_1) \leq \pi(\xi_2)$, the corresponding slices satisfy

$$A_2 = \{\xi; \pi(\xi) \geq u \pi(\xi_2)\} \subset A_1 = \{\xi; \pi(\xi) \geq u \pi(\xi_1)\},$$

and, therefore, simulation from a uniform distribution on $A_2$ can first proceed by acceptance–rejection of a uniform sampling on $A_1$. From a perfect sampling point of view, this property induces a natural propensity for coupling: if $\xi'_1 \sim U(A_1)$ belongs to $A_2$, this realization is also acceptable as a simulation from $U(A_2)$, and both chains couple; if it does not belong to $A_2$, the value $\xi'_2$ simulated directly from the uniform distribution on $A_2$ preserves the ordering $\pi(\xi'_1) \leq \pi(\xi'_2)$.

Secondly, and most importantly, in the case of expression (5), there is a maximal and a minimal element, $\tilde{1}$ and $\tilde{0}$, for the order induced by $\pi(n_0, s_0)$, since the space of the $(n_0, s_0)$s is finite. Moreover, these maximal and minimal elements can be identified for standard mixtures. Thus, monotone CFTP (Propp and Wilson, 1996) applies, i.e. it is sufficient to run two chains starting from $\tilde{1}$ and $\tilde{0}$, and to check whether both chains coalesce at time 0. Following a monotonicity argument, the collection of all chains in between the extreme chains has coalesced by the time that those two coalesce.

### 3.1.1. Example 3: exponential mixtures (continued)

Consider distribution (7). For a given value $n_0$ such that $0 \leq n_0 \leq n$, $s_0$ belongs to the interval $[s_0(n_0), \overline{s}_0(n_0)]$, where

$$s_0(n_0) = \sum_{j=1}^{n_0} x(j),$$
$$\overline{s}_0(n_0) = \sum_{j=1}^{n_0} x(n_0+1-j),$$

in which the $x(i)$s denote the order statistics of the sample $x_1, \ldots, x_n$, with $x(1) \leq \ldots \leq x(n)$. The minimum and maximum of $\pi(n_0, s)$ for a given value of $n_0$ can then be obtained in closed form, given the monotonicity in $s$ of $\pi(n_0, s)$. This function is decreasing and then increasing, with a minimum at

$$\hat{s}_0(n_0) = \frac{(n_0 + \alpha_0)(\beta_1 + S) - (n - n_0 + \alpha_1)\beta_0}{n + \alpha_0 + \alpha_1},$$

provided that this value is in $[s_0(n_0), \overline{s}_0(n_0)]$. The maximum is obviously attained at one of the two extremes. The points $\tilde{1}$ and $\tilde{0}$ are then obtained by $n + 1$ comparisons over all $n_0$s. Note that $\hat{s}_0(n_0)$ and hence $\tilde{0}$ are not necessarily possible points, i.e. sums of $n_0$ observations from the sample. This is not a problem for perfect sampling: starting from $\tilde{0}$ instead of the true minimum only slows down coalescence but does not modify the validity of CFTP. (This is the envelope argument of Kendall (1998).)

### 3.1.2. Example 4: normal mixture (continued)

Since, for a given $n_0$, $\overline{s}_0$ is between

$$\overline{x}(n_0) = (x(1) + \ldots + x(n_0))/n_0$$
and
\[ \overline{x}(n_0) = (x(n) + \ldots + x(n - n_0 + 1))/n_0, \]
and, because the weight of \( x^2_0 \) in expression (8) is negative, expression (8) attains its maximum at one of the end points and its minimum either inside the interval \( (\overline{x}(n_0), \overline{x}(n_0)) \) or at the other end point.

Note again the crucial appeal of running the slice sampler on the latent variable chain rather than on the original parameter chain. It is indeed nearly impossible to find \( \hat{0} \) and \( \hat{1} \) for the latter, since this is equivalent to finding the maximum likelihood estimator (for \( \hat{1} \)) and a ‘minimum likelihood estimator’ (for \( \hat{0} \)). In non-compact cases, these extrema may not even exist when the likelihood is unbounded.

As noted in Mira et al. (2001), only knowledge of the maximal element \( \hat{1} \) is necessary to run the monotone perfect slice sampler, given that the minimal element \( \hat{0} \) is never really used: for the chain starting from \( \hat{0} \), the next value is selected at random from the entire statespace of the \( \omega_s \), since \( \pi(\xi) \geq \pi(\hat{0}) \) does not impose any constraint on \( \xi \).

### 3.2. Exact slice sampling

In principle, it is possible to implement a perfect slice sampler on the discrete chain. For simplicity, we study only the exponential mixture case for \( k = 2 \), although the method formally extends to larger values of \( k \).

Consider the two extremes \( \hat{0} \) and \( \hat{1} \) in the \( (n_0, s_0) \) state space. Then, values can be generated from the uniform distribution
\[ U(\{(n_0, s_0); \pi(n_0, s_0) \geq \varepsilon\}) \]
by a crude accept–reject scheme which picks out points at random from
\[ \left\{(n_0, s_0); s_0 = \sum_{j=1}^{n_0} x_{i_j}, 1 \leq i_1 < \ldots < i_{n_0} \leq n \right\} \tag{9} \]
until they satisfy the constraint \( \pi(n_0, s_0) \geq \varepsilon \). (Note that this random sampling can be decomposed as the selection of \( n_0 \) with probability proportional to \( \binom{n}{n_0} \) and the subsequent generation of a random permutation of \( \{1, \ldots, n\} \).)

The first difficulty with this implementation is that the uniform sequence from expression (9) used to generate the points \( (n_0^{(t)}, s_0^{(t)}) \) must be preserved for the next time that \( t \) is visited in the CFTP scheme, to build the deterministic mapping that validates CFTP. (As rightly pointed out by a referee, a practical implementation of this requirement is to store only the initial random seed used at time \( t \), given that the infinite sequence of uniform points on expression (9) will always be the same.)

The second problem with this perfect slice sampler is more down to earth but cannot be overcome: since the accept–reject step is based on a uniform proposal on the set of all possible values, acceptance may require many proposals. It is obviously reduced to be implemented by eliminating values of \( n_0 \) for which the maximum (in \( s_0 \)) posterior is below the bound \( \varepsilon \), but this only removes a tiny portion of the \( k^n \)-sized parameter space. The poor match between set (9) and \( \{(n_0, s_0); \pi(n_0, s_0) \geq \varepsilon\} \) is such a serious problem that we failed to achieve a single coalescence after several days of monitoring for a two-component exponential mixture with 60 observations.

For smaller sample sizes, we could, however, achieve coalescence in a ‘human scale’ (rather than geological) time. For instance, Fig. 1 represents the 32 steps of a successful CFTP for 35
The fundamental difficulty behind the brute force implementation of the slice sampler via the uniform proposal in expression (9) is that the range of the log-posterior values \( \log \{ \pi(n_0, s_0) \} \) increases with \( n \). For instance, in the case of 40 observations, it ranges from \(-102\) to \(-69\). Therefore, when considering the maximum point \( \hat{1} \), it is increasingly unlikely to find a point at random that satisfies \( \pi(n_0, s_0)s \geq u \pi(\hat{1}) \) if \( u \) is quite close to 1. A way to alleviate this partially is to introduce intermediate chains, i.e. new starting-points between \( \hat{0} \) and \( \hat{1} \). The motivation for this multiplication of chains is that, by providing a better coverage of the range \( [\pi(\hat{0}), \pi(\hat{1})] \), we increase the chances that the upper chain (starting from \( \hat{1} \)) collapses into lower log-posterior chains, thus reducing the odds that the algorithm becomes stuck with a large \( u \) and almost no \( (n_0, s_0) \) satisfying \( \pi(n_0, s_0)s \geq u \pi(\hat{1}) \). However, this modification does not improve the speed of the algorithm, as shown by an experiment with the same data set as that used in Fig. 1 with 99 intermediate chains, since coalescence still took approximately 40 min of computation.

3.3. A continuous approximation

Since the difficulty with the above perfect slice sampler resides in the discrete nature of the parameter space, we may be tempted to embed distribution (7) in a continuous distribution, for which slice sampling is much easier. We first show why this is so, before discussing the validity of the approximation.

We thus temporarily consider distribution (7) as a density function in \( s_0 \), conditionally on
\( n_0 \), where \( s_0 \) varies on the interval \([\bar{s}_0(n_0), \tilde{s}_0(n_0)]\). The maximum and minimum values \( \tilde{1} \) and \( \tilde{0} \) obtained previously are still valid in this continuous set-up. Moreover, it is straightforward for standard exponential families to generate points uniformly from

\[
\{(n_0, s_0); \ n_0 = 0, \ldots, n, \ s_0 \in [\bar{s}_0(n_0), \tilde{s}_0(n_0)], \ \pi(n_0, s_0) \geq \varepsilon \},
\]

given that this set is the union over \( n_0 \) of the sets

\[
\mathcal{R}(n_0) = \{s_0(n_0) \leq s_0 \leq \tilde{s}_0(n_0); \ \pi(n_0, s_0) \geq \varepsilon\},
\]

which can be represented in closed form.

### 3.3.1. Example 5: exponential mixtures (continued)

The set of \( s_0 \)s such that

\[
p^{\alpha_0} (1 - p)^{\alpha_1} \frac{\Gamma(\alpha_0 + n_0) \Gamma(\alpha_1 + n_1)}{(\beta_0 + s_0)^{\alpha_0 + n_0} (\beta_1 + S - s_0)^{\alpha_1 + n_1}} \geq \varepsilon
\]

is the entire interval \([\bar{s}_0(n_0), \tilde{s}_0(n_0)]\) if the value at the minimum \( s_0^* \) is larger than \( \varepsilon \); otherwise, it is easily derived by dichotomous trial-and-error search.

The joint uniform generation of \((n_0, s_0)\) in expression (10) thus depends on two uniform random variables: one to select \( n_0 \) with probability proportional to \( l(n_0) \), where \( l(n_0) \) is the length of the region \( \mathcal{R}(n_0) \), and one to select \( s_0 \) uniformly in the region \( \mathcal{R}(n_0) \). We can thus represent the slice sampling procedure as a deterministic mapping and derive from this representation a straightforward CFTP algorithm, which is a true perfect sampling algorithm for the continuous distribution but only an approximation for the discrete distribution (7).

Fig. 2 provides an illustration of the paths of the two chains started at \( \tilde{0} \) and \( \tilde{1} \) for a two-component exponential mixture with \( n = 73 \) observations. It also provides the corresponding range of the log-posteriors \( \log \{\pi(\omega_0^{(t)})\} \) and \( \log \{\pi(\omega_1^{(t)})\} \). Note that, contrary to the exact case of Section 3.2, coalescence is fast, even for large data sets.

The above theory extends to the case \( k > 2 \), as long as we can find a maximum starting value \( \tilde{1} \), although we are then forced to settle for simple but slow accept–reject methods.

### 3.4. Evaluation of the approximation

At first glance, we may consider that, for large values of \( n \), the continuous interval \([\bar{s}_0(n_0), \tilde{s}_0(n_0)]\) is a good approximation to the set of possible values of \( s_0 \). This is not so, however, because, as pointed out to us by R. M. Neal, there will be many more \( s_0 \)s in the centre of the interval \([\bar{s}_0(n_0), \tilde{s}_0(n_0)]\) than close to its boundaries and the uniform assumption used in the continuous approximation is therefore invalid. The true distribution of the \( s_0 \)s will thus tend to have a binomial or normal distribution shape. The second reason for a discrepancy is that the first and last values of \( s_0 \) are well separated, regardless of the sample size, given their dependence on extreme values.

The continuous extension thus leads to an approximation of the genuine distribution of interest. One may then wonder how accurate this approximation can be. The answer is, unfortunately, that it can be very poor. Given that the approach of Section 4 does provides a true sample from the posterior distribution associated with a mixture, we can assess the effect of the continuous approximation. For instance, in the case of a two-component exponential mixture with \( n = 81 \) and \( n_0 = 59 \), Fig. 3 shows how disastrous the approximation can be: whereas the density of \( s_0 \)
Fig. 2. Sample paths similar to Fig. 1 for a simulated sample of 73 observations from a mixture of two exponential distributions, when using the continuous CFTP approximation

given by distribution (7) decreases quite rapidly towards 0, as reflected by the histogram on the left-hand side, the main bulk of the possible values of \( s_0 \) is centred much more to the right and the reweighting of these values by the probability mass function (7) only brings the histogram slightly to the left, still leaving the supports of both histograms almost disjoint.

4. Perfect sampling via automatic coupling

As well as slice sampling, there is another approach to perfect sampling, based on Breyer and Roberts’s (2001) automatic coupling: given two Markov chains \((x_0^{(t)})\) and \((x_1^{(t)})\), the first chain \((x_0^{(t)})\) is run following a standard Metropolis–Hasting scheme:

\[
x_0^{(t+1)} = \begin{cases} 
  y_t \sim q(y|x_0^{(t)}) & \text{if } u_t \leq \frac{\pi(y_t) q(x_0^{(t)}|y_t)}{\pi(x_0^{(t)}) q(y_t|x_0^{(t)})}, \\
  x_0^{(t)} & \text{otherwise.}
\end{cases}
\]

The transition for the second chain \((x_1^{(t)})\) is then based on the same proposed value \(y_t\), with the difference that the Metropolis–Hastings scheme is now an independent scheme, and hence the transition mechanism is

\[
x_1^{(t+1)} = \begin{cases} 
  y_t & \text{if } u_t \leq \frac{\pi(y_t) q(x_1^{(t)}|x_0^{(t)})}{\pi(x_1^{(t)}) q(y_t|x_0^{(t)})}, \\
  x_1^{(t)} & \text{otherwise.}
\end{cases}
\]
We can use the same uniform variable $u_t$ for both transitions, thus ensuring that the chains coalesce once they both accept a proposal $y_t$.

The implications of this rudimentary coupling strategy on perfect sampling are quite interesting. Consider an independent proposal distribution $h(y)$ for the first chain $(x_0^{(t)})$. The Metropolis–Hastings acceptance probabilities for both chains are then

$$
\rho_0^{(t)} = \frac{\pi(y_t) h(x_0^{(t)})}{\pi(x_0^{(t)}) h(y_t)} \wedge 1,
$$

$$
\rho_1^{(t)} = \frac{\pi(y_t) h(x_1^{(t)})}{\pi(x_1^{(t)}) h(y_t)} \wedge 1.
$$

If the two chains are ordered according to the order induced by $\pi/h$, i.e. if

$$
\frac{\pi(x_0^{(t)})}{h(x_0^{(t)})} \leq \frac{\pi(x_1^{(t)})}{h(x_1^{(t)})},
$$

coupling scheme (11) preserves the ordering. Indeed, since $\rho_0^{(t)} \geq \rho_1^{(t)}$ when the chains are ordered and the same uniform variable $u_t$ is used to update both chains, there are three possible cases:

(a) $y_t$ is rejected for both chains;
(b) $y_t$ is accepted for the first chain but rejected by the second chain;
(c) $y_i$ is accepted for both chains.

That the order is preserved for cases (a) and (c) is trivial. For case (b), the fact that $y_i$ is rejected for the second chain implies that $\rho_1^{(i)} < 1$, and thus that

$$\frac{\pi(y_i)}{h(y_i)} < \frac{\pi(x_1^{(i)})}{h(x_1^{(i)})}.$$ 

This property, namely that the coupling proposal preserves the ordering, was first obtained by Corcoran and Tweedie (2000).

Consider a monotone setting where there are a minimal and a maximal element, $\tilde{0}$ and $\tilde{1}$, for the order induced by $\pi/h$. We can then run a perfect sampler based on the two chains starting from $\tilde{0}$ and $\tilde{1}$. Note that, since the choice of $h$ is arbitrary, it can be made to give an easy derivation of $\tilde{0}$ and $\tilde{1}$. None-the-less, this cannot be achieved in every setting since the existence of $\tilde{1}$ implies that the corresponding MCMC chain is uniformly ergodic, as shown by Mengersen and Tweedie (1996). In the special case where the state space $\mathcal{X}$ is compact, $h$ can be chosen as the uniform distribution on $\mathcal{X}$ and the extremal elements $\tilde{0}$ and $\tilde{1}$ are then those induced by $\pi$.

This principle applies to the mixture problem, since the pair $(n_0, s_0)$ evolves in a compact (finite) state. To illustrate its implementation in the exponential case, we fixed the value of $n_0$ and simulated an exact sample of $s_0$ from the distribution proportional to

$$\frac{\Gamma(\alpha_0 + n_0)}{(\beta_0 + s_0)^{\alpha_0 + n_0}} \quad \frac{\Gamma(\alpha_1 + n - n_0)}{(\beta_1 + S - s_0)^{\alpha_1 + n - n_0}}$$

on the finite set of the possible values of $s_0 = \sum_{j=1}^{n_0} x_{ij}$.

This means that, starting at time $-T$ from $\tilde{0}$ and $\tilde{1}$, both of which can be obtained in closed form on the continuous interval, we generated the chain $(x_0^{(i)})$ forwards in time according to a regular Metropolis–Hastings scheme with a uniform proposal on the set of possible values of $s_0$. The chain $(x_1^{(i)})$ was then coupled to the first chain by mechanism (11), i.e. by using the current proposal for $x_0^{(i)}$ as a proposal for $x_1^{(i)}$. If coalescence occurred by time 0, this meant that CFTP was successful. Otherwise, we increased $T$.

Fig. 4(a) illustrates the rudimentary nature of this coupling: for a long time (almost 1000 iterations), coupling is rejected by $(x_1^{(i)})$, which remains equal to the original value 1, whereas $(x_0^{(i)})$ moves with a good mixing rate. Once coupling has occurred, both chains are identical and keep mixing with the same rate till time 0. So, in practice, this CFTP fundamentally relies on a single chain, the second chain only being used through its reference value $\tilde{1}$ to delay coalescence. In that sense, this perfect sampling algorithm is simply a hidden accept–reject algorithm.

Fig. 4(b) provides additional insight into the coupling phenomenon by plotting the corresponding log-posteriors. As can be seen from this graph, the chain $(x_0^{(i)})$ explores almost from the start a given range of log-posteriors, whereas $(x_1^{(i)})$ remains, until coupling, at this impossibly high value. Once the proposed value of $x_0^{(i)}$ is accepted by $(x_1^{(i)})$, the (single) chain never again visits such high values of the log-posteriors.

A drawback with this approach is obviously that the difference between the two extremes $\tilde{0}$ and $\tilde{1}$ may prevent coupling for a long run of iterations. Fig. 5 presents an illustration of this problem for 81 observations from a mixture of two exponential distributions through the
5. Conclusion

We have studied two exact samplers for mixture posterior distributions, one of which can handle reasonably large sample sizes. We also introduced a continuous approximation which allows us to run a much faster slice sampler but we demonstrated that this approximation can be poor.

For brevity we have not considered examples where the weights are unknown, but the
Perfect Samplers

integration of the weights against a conjugate prior does not modify the complexity of the posterior distribution in $Z$, since it only depends on the $n_i$s, as shown by expression (5).

We have thus obtained a general sampling method for data that are independent and identically distributed for mixture posterior distributions. As pointed out in the discussion of Green and Murdoch (1999), mixture models are a good showcase for establishing that perfect sampling can be achieved for realistic statistical models. Green and Murdoch (1999) 'surmise that it might be possible to couple MCMC methods for mixtures’ and the present paper somehow achieves this goal, although more work remains to be done to handle larger sample sizes and more general mixtures.

Acknowledgements

The first author was supported by National Science Foundation grant DMS-9971586. The second author was supported by an Australian Science Council ‘Large grant’ and by the Centre National de la Recherche Scientifique during a visit to the Centre de Recherche en Economie et Statistique in August 1999. The third author was supported by the European Union travel and mobility of researchers network ERB-FMRX-CT96-0095 on ‘Computational and statistical methods for the analysis of spatial data’. He thanks the Statistics Department of the University of Glasgow for its hospitality and support in both 1999 and 2000. He is grateful to Laird Breyer, Peter Green, Antonietta Mira and Gareth Roberts for helpful discussions, and to Radu Craiu and Radford Neal for pointing out difficulties with the continuous approximation. Supporting comments from the referees on earlier versions are gratefully acknowledged.

References

Besag, J. and Green, P. J. (1993) Spatial statistics and Bayesian computation (with discussion). *J. R. Statist. Soc. B*, 55, 25–38.
Breyer, L. and Roberts, G. O. (2001) Catalytic perfect sampler. *J. Comput. Math.*, to be published.

Casella, G., Robert, C. P. and Wells, M. T. (1999) Partitioned importance sampling. Department of Statistics, Cornell University, Ithaca.

Celeux, G., Hurn, M. and Robert, C. P. (2000) Computational and inferential difficulties with mixture posterior distributions. *J. Am. Statist. Ass.*, 95, 957–970.

Corcoran, J. N. and Tweedie, R. L. (2000) Perfect sampling from independent Metropolis–Hastings chains. *Technical Report*. Colorado State University, Fort Collins.

Damien, P., Wakefield, J. and Walker, S. (1999) Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. R. Statist. Soc. B*, 61, 331–344.

Diebolt, J. and Robert, C. P. (1994) Estimation of finite mixture distributions through Bayesian sampling. *J. R. Statist. Soc. B*, 56, 363–375.

Dimakos, X. (2001) A guide to exact simulation. *Int. Statist. Rev.*, 69, 27–48.

Green, P. J. and Murdoch, D. J. (1999) Exact sampling (with discussion). In *Bayesian Statistics 6* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 301–321. Oxford: Oxford University Press.

Hobert, J. P., Robert, C. P. and Titterington, D. M. (1999) On perfect simulation for some mixtures of distributions. *Statist. Comput.*, 9, 287–298.

Kendall, W. S. (1998) Perfect simulation for the area-interaction point process. In *Probability Towards 2000* (eds L. Accardi and C. C. Heyde), pp. 218–234. New York: Springer.

Mengersen, K. L. and Tweedie, R. L. (1996) Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, 24, 101–121.

Mira, A., Møller, J. and Roberts, G. O. (2001) Perfect slice samplers. *J. R. Statist. Soc. B*, 63, 593–606; corrigendum, 64 (2002), 581.

Møller, J. and Nicholls, G. (1999) Perfect simulation for sample-based inference. *Statist. Comput.*, to be published.

Propp, J. G. and Wilson, D. B. (1996) Exact sampling with coupled Markov chains and applications to statistical mechanics. *Rand. Struct. Alg.*, 9, 223–252.

Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Statist. Soc. B*, 59, 731–792.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985) *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

Wilson, D. B. (1998) Annotated bibliography of perfectly random sampling with Markov chains. In *Microsurveys in Discrete Probability* (eds D. Aldous and J. Propp), pp. 209–220. Providence: American Mathematical Society.