Asymptotically Independent Markov Sampling: a new MCMC scheme for Bayesian Inference

Konstantin M. Zuev[1] and James L. Beck[2]

[1]Institute for Risk and Uncertainty, Centre for Engineering Sustainability, University of Liverpool, UK, email: zuev@liverpool.ac.uk
[2]Departments of Computing and Mathematical Sciences, and Mechanical and Civil Engineering, California Institute of Technology, USA, email: jimbeck@caltech.edu

ABSTRACT

In Bayesian inference, many problems can be expressed as the evaluation of the expectation of an uncertain quantity of interest with respect to the posterior distribution based on relevant data. Standard Monte Carlo method is often not applicable because the encountered posterior distributions cannot be sampled directly. In this case, the most popular strategies are the importance sampling method, Markov chain Monte Carlo, and annealing. In this paper, we introduce a new scheme for Bayesian inference, called Asymptotically Independent Markov Sampling (AIMS), which is based on the above methods. The efficiency of AIMS is demonstrated with an example that involves a multi-modal target distribution.

INTRODUCTION

In Bayesian statistics, many problems can be expressed as the evaluation of the expectation of a quantity of interest with respect to the posterior distribution. Standard Monte Carlo simulation, where expectations are estimated by sample averages based on samples drawn independently from the posterior, is often not applicable because the encountered posterior distributions are multi-dimensional distributions that cannot be explicitly normalized. In this case, the most popular strategies are importance sampling and Markov chain Monte Carlo methods. We briefly review these two methods first because they play an important role in the new MCMC method introduced in this paper.

Importance sampling: This is nearly as old as the Monte Carlo method (see, for instance, (Kahn and Marshal, 1953)), and works as follows. Suppose we want to evaluate $\mathbb{E}_\pi[h]$ that is an expectation of a function of interest $h : \Theta \to \mathbb{R}$ under distribution $\pi(\cdot)$ defined on a parameter space $\Theta \subseteq \mathbb{R}^d$,

$$\mathbb{E}_\pi[h] = \int_\Theta h(\theta)\pi(\theta)d\theta. \tag{1}$$

Suppose also that we are not able to sample directly from $\pi(\cdot)$, although we can compute $\pi(\theta)$ for any $\theta \in \Theta$ to within a proportionality constant. Instead, we sample

from some other distribution $q(\cdot)$ on Θ which is readily computable for any $\theta \in \Theta$. Let $\theta^{(1)}, \ldots, \theta^{(N)}$ be N i.i.d. samples from $q(\cdot)$, and $w^{(i)} = \pi(\theta^{(i)})/q(\theta^{(i)})$ denote the *importance weight* of the i^{th} sample, then we can estimate $\mathbb{E}_\pi[h]$ by

$$\hat{h}_N = \frac{\sum_{i=1}^N w^{(i)} h(\theta^{(i)})}{\sum_{i=1}^N w^{(i)}}. \tag{2}$$

The estimator $\hat{h}_N$ converges almost surely as $N \to \infty$ to $\mathbb{E}_\pi[h]$ by the Strong Law of Large Numbers for any choice of distribution $q(\cdot)$, provided $\mathrm{supp}(\pi) \subseteq \mathrm{supp}(q)$. Note that the latter condition automatically holds in Bayesian updating using data $\mathcal{D}$ where $q(\theta) = \pi_0(\theta)$ is the prior density and $\pi(\theta) \propto \pi_0(\theta)L(\theta)$ is the posterior $p(\theta|\mathcal{D})$, where L stands for the likelihood function $p(\mathcal{D}|\theta)$.

The accuracy of $\hat{h}_N$ depends critically on the choice of the *importance sampling distribution* (ISD) $q(\cdot)$, which is also called the *instrumental* or *trial* distribution. If $q(\cdot)$ is chosen carelessly such that the the importance weights $w^{(i)}$ have a large variation, then $\hat{h}_N$ is essentially based only on the few samples $\theta^{(i)}$ with the largest weights, yielding generally a very poor estimate. Hence, for importance sampling to work efficiently, $q(\cdot)$ must be a good approximation of $\pi(\cdot)$ so that the variance $\mathrm{var}_q[w]$ is not large.

MCMC Sampling: Instead of generating independent samples from an ISD, we could generate dependent samples by simulating a Markov chain whose state distribution converges to the posterior distribution $\pi(\cdot)$ as its stationary distribution. *Markov chain Monte Carlo* sampling (MCMC) originated in statistical physics, and now is widely used in solving statistical problems (Robert and Casella, 2004).

The Metropolis-Hastings algorithm, the most popular MCMC technique, works as follows. Let $q(\cdot|\theta)$ be a distribution on Θ, which may or may not depend on $\theta \in \Theta$. Assume that $q(\cdot|\theta)$ is easy to sample from and it is either computable (up to a multiplicative constant) or symmetric, i.e. $q(\xi|\theta) = q(\theta|\xi)$. The sampling distribution $q(\cdot|\theta)$ is called the *proposal distribution*. Starting from essentially any $\theta^{(1)} \in \mathrm{supp}(\pi)$, the Metropolis-Hastings algorithm proceeds by iterating the following two steps. First, generate a *candidate* state ξ from the proposal density $q(\cdot|\theta^{(n)})$. Second, either accept ξ as the next state of the Markov chain, $\theta^{(n+1)} = \xi$, with probability $\alpha(\xi|\theta^{(n)}) = \min\left\{1, \frac{\pi(\xi)q(\theta^{(n)}|\xi)}{\pi(\theta^{(n)})q(\xi|\theta^{(n)})}\right\}$; or reject ξ and set $\theta^{(n+1)} = \theta^{(n)}$ with the remaining probability $1 - \alpha(\xi|\theta^{(n)})$. It can be shown (see, for example, (Robert and Casella, 2004)), that under fairly weak conditions, $\pi(\cdot)$ is the stationary distribution of the Markov chain $\theta^{(1)}, \theta^{(2)}, \ldots$ and

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^N h(\theta^{(i)}) = \int_\Theta h(\theta)\pi(\theta)d\theta. \tag{3}$$

The two main special cases of the Metropolis-Hastings algorithm are Independent Metropolis-Hastings (IMH), where the proposal distribution $q(\xi|\theta) = q_g(\xi)$ is independent of θ (so q_g is a *global proposal*), and Random Walk Metropolis-Hastings (RWMH), where the proposal distribution is of the form $q(\xi|\theta) = q_l(\xi - \theta)$, i.e. a candidate state

is proposed as $\xi = \theta^{(n)} + \epsilon_n$, where $\epsilon_n \sim q_l(\cdot)$ is a random perturbation (so q_l is a *local proposal*). In both cases, the choice of the proposal distribution strongly affects the efficiency of the algorithms.

Annealing: The concept of *annealing* (or *tempering*), which involves moving from an easy-to-sample distribution to the target distribution via a sequence of intermediate distributions, is one of the most effective methods of handling multiple isolated modes.

In Bayesian inference problems, the idea of annealing is typically employed in the following way. First, we construct (in advance or adaptively), a sequence of distributions $\pi_0(\cdot), \ldots, \pi_m(\cdot)$ interpolating between the prior distribution $\pi_0(\cdot)$ and the posterior distribution $\pi(\cdot) \equiv \pi_m(\cdot)$. Next, we generate i.i.d. samples $\theta_0^{(1)}, \ldots, \theta_0^{(N)}$ from the prior, which is assumed to be readily sampled. Then, at each annealing level j, using some MCMC algorithm and samples $\theta_{j-1}^{(1)}, \ldots, \theta_{j-1}^{(N)}$ from the previous level $j-1$, we generate samples $\theta_j^{(1)}, \ldots, \theta_j^{(N)}$ which are approximately distributed according to $\pi_j(\cdot)$. We proceed sequentially in this way, until the posterior distribution has been sampled. The rationale behind this strategy is that sampling from the multi-modal and, perhaps, high-dimensional posterior in such a way is likely to be more efficient than a straightforward MCMC sampling of the posterior.

In this paper we introduce a new MCMC scheme for Bayesian inference, called *Asymptotically Independent Markov Sampling* (AIMS), which combines the three approaches described above — importance sampling, MCMC, and annealing — in the following way. Importance sampling with $\pi_{j-1}(\cdot)$ as the ISD is used for a construction of an approximation $\hat{\pi}_j^N(\cdot)$ of $\pi_j(\cdot)$, which is based on samples $\theta_{j-1}^{(1)}, \ldots, \theta_{j-1}^{(N)} \sim \pi_{j-1}(\cdot)$. This approximation is then employed as the independent (global) proposal distribution for sampling from $\pi_j(\cdot)$ by the IMH algorithm. Intermediate distributions $\pi_0(\cdot), \ldots, \pi_m(\cdot)$ interpolating between prior and posterior are constructed adaptively, using the essential sample size (ESS) to measure how much $\pi_{j-1}(\cdot)$ differs from $\pi_j(\cdot)$. When the number of samples $N \to \infty$, the approximation $\hat{\pi}_j^N(\cdot)$ converges to $\pi_j(\cdot)$, providing the optimal proposal distribution. In other words, when $N \to \infty$, the corresponding MCMC sampler produces independent samples, hence the name of the algorithm.

In this introductory section, we have described all the main ingredients that we will need in the subsequent sections. In the rest of the paper we describe the AIMS algorithm and illustrate its efficiency with an example that involves a multi-modal target distribution.

ASYMPTOTICALLY INDEPENDENT MARKOV SAMPLING

Let $\pi_0(\cdot)$ and $\pi(\cdot)$ be the prior and the posterior distributions defined on a parameter space Θ, respectively, so that, according to Bayes' Theorem, $\pi(\theta) \propto \pi_0(\theta)L(\theta)$, where L denotes the likelihood function for data $\mathcal{D}$. Our ultimate goal is to draw samples that are distributed according to $\pi(\cdot)$.

In AIMS, we sequentially generate samples from intermediate distributions

$\pi_0(\cdot), \ldots, \pi_m(\cdot)$ interpolating between the prior $\pi_0(\cdot)$ and the posterior $\pi(\cdot) \equiv \pi_m(\cdot)$. The sequence of distributions could be specially constructed for a given problem but the following scheme (Neal, 2001) generally yields good efficiency:

$$\pi_j(\theta) \propto \pi_0(\theta) L(\theta)^{\beta_j}, \tag{4}$$

where $0 = \beta_0 < \beta_1 < \ldots < \beta_m = 1$. We will refer to j and β_j as the *annealing level* and the *annealing parameter* at level j, respectively.

AIMS at annealing level j

Our first goal is to describe how AIMS generates sample $\theta_j^{(1)}, \ldots, \theta_j^{(N_j)}$ from $\pi_j(\cdot)$ based on the sample $\theta_{j-1}^{(1)}, \ldots, \theta_{j-1}^{(N_{j-1})} \sim \pi_{j-1}(\cdot)$ obtained at annealing level $j-1$.

Let $K_j(\cdot|\cdot)$ be any transition kernel such that $\pi_j(\cdot)$ is a stationary distribution with respect to $K_j(\cdot|\cdot)$. By definition, this means that

$$\pi_j(\theta)d\theta = \int_\Theta K_j(d\theta|\xi)\pi_j(\xi)d\xi \tag{5}$$

Applying importance sampling with the ISD $\pi_{j-1}(\cdot)$ to integral (5), we have:

$$\pi_j(\theta)d\theta = \int_\Theta K_j(d\theta|\xi)\frac{\pi_j(\xi)}{\pi_{j-1}(\xi)}\pi_{j-1}(\xi)d\xi \approx \sum_{i=1}^{N_{j-1}} K_j(d\theta|\theta_{j-1}^{(i)})\bar{w}_{j-1}^{(i)} \stackrel{\text{def}}{=} \hat{\pi}_j^{N_{j-1}}(d\theta), \tag{6}$$

where $\hat{\pi}_j^{N_{j-1}}(\cdot)$ will be used as the *global proposal* distribution in the Independent Metropolis-Hastings algorithm, and

$$w_{j-1}^{(i)} = \frac{\pi_j(\theta_{j-1}^{(i)})}{\pi_{j-1}(\theta_{j-1}^{(i)})} \propto L(\theta_{j-1}^{(i)})^{\beta_j - \beta_{j-1}} \quad \text{and} \quad \bar{w}_{j-1}^{(i)} = \frac{w_{j-1}^{(i)}}{\sum_{k=1}^{N_{j-1}} w_{j-1}^{(k)}} \tag{7}$$

are the importance weights and normalized importance weights. If adjacent intermediate distributions $\pi_{j-1}(\cdot)$ and $\pi_j(\cdot)$ are sufficiently close (in other words, if $\Delta\beta_j = \beta_j - \beta_{j-1}$ is small enough), then the importance weights (7) will not vary wildly, and, therefore, we can expect that, for reasonably large N_{j-1}, approximation (6) is accurate.

From now on, we consider a special case where $K_j(\cdot|\cdot)$ is the random walk Metropolis-Hastings (RWMH) transition kernel. It can be written as follows:

$$K_j(d\theta|\xi) = q_j(\theta|\xi)\min\left\{1, \frac{\pi_j(\theta)}{\pi_j(\xi)}\right\}d\theta + (1 - a_j(\xi))\delta_\xi(d\theta), \tag{8}$$

where $q_j(\cdot|\xi)$ is a symmetric *local* proposal density, and $a_j(\xi)$ is the probability of having a proper transition ξ to $\Theta \setminus \{\xi\}$.

For sampling from $\pi_j(\cdot)$, we will use the Independent Metropolis-Hastings algorithm (IMH) with the *global* proposal distribution $\hat{\pi}_j^{N_{j-1}}(\cdot)$. This leads to the following algorithm (see details in (Beck and Zuev, 2013)):

AIMS at annealing level j

Input:

$\triangleright$ $\theta_{j-1}^{(1)}, \ldots, \theta_{j-1}^{(N_{j-1})} \sim \pi_{j-1}(\cdot)$, samples generated at annealing level $j-1$;

$\triangleright$ $\theta_j^{(1)} \in \Theta_j^* = \Theta \setminus \left\{\theta_{j-1}^{(1)}, \ldots, \theta_{j-1}^{(N_{j-1})}\right\}$, initial state of a Markov chain;

$\triangleright$ $q_j(\cdot|\xi)$, symmetric proposal density associated with the RWMH kernel;

$\triangleright$ N_j, total number of Markov chain states to be generated.

Algorithm:

for $i = 1, \ldots, N_j - 1$ **do**

1) Generate a global candidate state $\xi_g \sim \hat{\pi}_j^{N_{j-1}}(\cdot)$ as follows:

a. Select k from $\{1, \ldots, N_{j-1}\}$ with probabilities $\bar{w}_{j-1}^{(i)}$ given by (7).

b. Generate a local candidate $\xi_l \sim q_j(\cdot|\theta_{j-1}^{(k)})$.

c. Accept or reject ξ_l by setting

$$\xi_g = \begin{cases} \xi_l, & \text{with probability } \min\left\{1, \frac{\pi_j(\xi_l)}{\pi_j(\theta_{j-1}^{(k)})}\right\}; \\ \theta_{j-1}^{(k)}, & \text{with the remaining probability.} \end{cases} \tag{9}$$

2) Update $\theta_j^{(i)} \to \theta_j^{(i+1)}$ by accepting or rejecting ξ_g as follows:

if $\xi_g = \theta_{j-1}^{(k)}$

Set $\theta_j^{(i+1)} = \theta_j^{(i)}$

else

Set

$$\theta_j^{(i+1)} = \begin{cases} \xi_g, & \text{with probability } \min\left\{1, \frac{\pi_j(\xi_g)\hat{\pi}_j^{N_{j-1}}(\theta_j^{(i)})}{\pi_j(\theta_j^{(i)})\hat{\pi}_j^{N_{j-1}}(\xi_g)}\right\}; \\ \theta_j^{(i)}, & \text{with the remaining probability.} \end{cases} \tag{10}$$

end if

end for

Output:

$\blacktriangleright$ $\theta_j^{(1)}, \ldots, \theta_j^{(N_j)}$, N_j states of a Markov chain with a stationary distribution $\pi_j(\cdot)$

The proof that $\pi_j(\cdot)$ is indeed a stationary distribution for the Markov chain generated by AIMS is given in (Beck and Zuev, 2013).

The full AIMS procedure

At the zero$^{\text{th}}$ annealing level, $j = 0$, we generate prior samples $\theta_0^{(1)}, \ldots, \theta_0^{(N_0)}$, which usually can be readily drawn directly by a suitable choice of the prior distribution $\pi_0(\cdot)$. Then, using the algorithm described in the previous subsection, we generate samples $\theta_1^{(1)}, \ldots, \theta_1^{(N_1)}$, which are approximately distributed according to intermediate distribution $\pi_1(\theta) \propto \pi_0(\theta)L(\theta)^{\beta_1}$. We proceed like this until the posterior distribution

$\pi_m(\theta) \propto \pi_0(\theta)L(\theta)^{\beta_m}$ ($\beta_m = 1$) has been sampled. To make the description of AIMS complete, we have to explain how to choose the annealing parameters β_j.

In importance sampling, a useful measure of degeneracy of the method is the *effective sample size* (ESS) N^{eff} introduced in (Kong et al, 1994). The ESS measures how similar the importance sampling distribution $\pi_{j-1}(\cdot)$ is to the target distribution $\pi_j(\cdot)$. Suppose N_{j-1} independent samples $\theta_{j-1}^{(1)}, \ldots, \theta_{j-1}^{(N_{j-1})}$ are generated from $\pi_{j-1}(\cdot)$, then the ESS of these samples is defined as

$$N_{j-1}^{\text{eff}} = \frac{N_{j-1}}{1 + \text{var}_{\pi_{j-1}}[w]} = \frac{N_{j-1}}{\mathbb{E}_{\pi_{j-1}}[w^2]}, \tag{11}$$

where $w(\theta) = \pi_j(\theta)/\pi_{j-1}(\theta)$. The ESS can be interpreted as implying that N_{j-1} weighted samples $(\theta_{j-1}^{(1)}, w_{j-1}^{(1)}), \ldots, (\theta_{j-1}^{(N_{j-1})}, w_{j-1}^{(N_{j-1})})$ are worth $N_{j-1}^{\text{eff}} (\leq N_{j-1})$ i.i.d. samples drawn from the target distribution $\pi_j(\cdot)$. One cannot evaluate the ESS exactly but an estimate $\hat{N}_{j-1}^{\text{eff}}$ of N_{j-1}^{eff} is given by

$$\hat{N}_{j-1}^{\text{eff}}(\bar{w}_{j-1}) = \frac{1}{\sum_{i=1}^{N_{j-1}} (\bar{w}_{j-1}^{(i)})^2}, \tag{12}$$

where $\bar{w}_{j-1} = (\bar{w}_{j-1}^{(1)}, \ldots, \bar{w}_{j-1}^{(N_{j-1})})$ and $\bar{w}_{j-1}^{(i)}$ is the normalized importance weight.

At annealing level j, when β_{j-1} is already known, the problem is to define β_j. Let $\gamma = \hat{N}_{j-1}^{\text{eff}}/N_{j-1} \in (0, 1)$ be a prescribed threshold that characterizes the "quality" of the weighted sample (the larger γ is, the "better" the weighted sample is). Then we obtain the following equation:

$$\sum_{i=1}^{N_{j-1}} (\bar{w}_{j-1}^{(i)})^2 = \frac{1}{\gamma N_{j-1}} \tag{13}$$

Observe that this equation can be expressed as an equation for β_j by using (7):

$$\frac{\sum_{i=1}^{N_{j-1}} L(\theta_{j-1}^{(i)})^{2(\beta_j - \beta_{j-1})}}{\left(\sum_{i=1}^{N_{j-1}} L(\theta_{j-1}^{(i)})^{\beta_j - \beta_{j-1}}\right)^2} = \frac{1}{\gamma N_{j-1}} \tag{14}$$

Solving this equation for β_j gives us the value of the annealing parameter at level j.

Combining the AIMS algorithm at a given annealing level with the described adaptive annealing scheme gives rise to the following procedure.

The AIMS procedure

Input:

- $\triangleright$ γ, threshold for the effective sample size (ESS);
- $\triangleright$ $N_0, N_1, \ldots$, where N_j is the total number of Markov chain states to be generated at annealing level j;
- $\triangleright$ $q_1(\cdot|\xi), q_2(\cdot|\xi), \ldots$, where $q_j(\cdot|\xi)$ is the symmetric proposal density associated

with the RWMH kernel at annealing level j.

Algorithm:

Set $j = 0$, current annealing level.

Set $\beta_0 = 0$, current annealing parameter.

Sample $\theta_0^{(1)}, \ldots, \theta_0^{(N_0)} \overset{i.i.d}{\sim} \pi_0(\cdot)$.

Calculate $\bar{W}_0^{(i)} = \frac{L(\theta_0^{(i)})^{1-\beta_0}}{\sum_{i=1}^{N_0} L(\theta_0^{(i)})^{1-\beta_0}}$, $i = 1, \ldots, N_0$.

Calculate the ESS $\hat{N}_0^{\text{eff}} = \hat{N}_0^{\text{eff}}(\bar{W}_0)$ using (12), which measures how similar the prior distribution $\pi_0(\cdot)$ is to the target posterior distribution $\pi(\cdot)$.

while $\hat{N}_j^{\text{eff}}/N_j < \gamma$ **do**

Find β_{j+1} from equation (14).

Calculate normalized importance weights $\bar{w}_j^{(i)}$, $i = 1, \ldots, N_j$ using (7).

Generate a Markov chain $\theta_{j+1}^{(1)}, \ldots, \theta_{j+1}^{(N_{j+1})}$ with the stationary distribution $\pi_{j+1}(\cdot)$ using the AIMS algorithm at annealing level $j + 1$.

Calculate $\bar{W}_{j+1}^{(i)} = \frac{L(\theta_{j+1}^{(i)})^{1-\beta_{j+1}}}{\sum_{i=1}^{N_{j+1}} L(\theta_{j+1}^{(i)})^{1-\beta_{j+1}}}$, $i = 1, \ldots, N_{j+1}$.

Calculate the ESS $\hat{N}_{j+1}^{\text{eff}} = \hat{N}_{j+1}^{\text{eff}}(\bar{W}_{j+1})$ using (12), which measures how similar the intermediate distribution $\pi_{j+1}(\cdot)$ is to the posterior $\pi(\cdot)$.

Increment j to $j + 1$.

end while

Set $\beta_{j+1} = 1$, current annealing parameter.

Set $m = j + 1$, the total number of distributions in the annealing scheme.

Set $\bar{w}_{m-1}^{(i)} = \bar{W}_{m-1}^{(i)}$, $i = 1, \ldots, N_{m-1}$.

Generate a Markov chain $\theta_m^{(1)}, \ldots, \theta_m^{(N_m)}$ with the stationary distribution $\pi_m(\cdot) = \pi(\cdot)$ using the AIMS algorithm at annealing level m.

Output:

▶ $\theta_m^{(1)}, \ldots, \theta_m^{(N_m)} \dot{\sim} \pi(\cdot)$, samples that are approximately distributed according to the posterior distribution.

The implementation issues of AIMS are discusses in detail in (Beck and Zuev, 2013).

ILLUSTRATIVE EXAMPLE

In this section we illustrate the use of AIMS with an example that involves a mixture of Gaussian distributions in two dimensions (a multi-modal case). A high-dimensional example, an example of Bayesian updating of a neural network model, and the comparison of AIMS with Transitional Markov chain Monte Carlo (TMCMC) (Ching and Chen, 2007) are given in (Beck and Zuev, 2013).

Multi-modal mixture of Gaussians in 2D

To demonstrate the efficiency of AIMS for sampling from multi-modal distributions, consider simulation from a truncated two-dimensional mixture of M Gaussian densities:

$$\pi(\theta) \propto \pi_0(\theta) \cdot L(\theta) = \mathcal{U}_{[0,a]\times[0,a]}(\theta) \cdot \sum_{i=1}^{M} w_i \mathcal{N}(\theta|\mu_i, \sigma^2 \mathbb{I}_2), \tag{15}$$

where $\mathcal{U}_{[0,a]\times[0,a]}(\cdot)$ denotes the uniform distribution on the square $[0, a] \times [0, a]$. In this example, $a = 10$, $M = 10$, $\sigma = 0.1$, $w_1 = \ldots = w_{10} = 0.1$, and the mean vectors $\mu_1, \ldots, \mu_{10}$ are drawn uniformly from the square $[0, 10]\times[0, 10]$. Because of our interest in Bayesian updating, we refer to $\pi(\cdot)$ in (15) as a posterior distribution.

Figure 1(a) displays the scatterplot of 10^3 posterior samples obtained from AIMS. Notice there are two clusters of samples that overlap significantly near $\theta = (4, 4)$ that reflect two closely spaced Gaussian densities but the other 8 clusters are widely spaced. The parameters of the algorithm were chosen as follows: sample size $N = 10^3$ per annealing level; the threshold for the ESS $\gamma = 1/2$; the local proposal density $q_j(\cdot|\xi) = \mathcal{N}(\cdot|\xi, c^2\mathbb{I}_2)$, with $c = 0.2$. The trajectory of the corresponding posterior Markov chain, i.e. the chain generated at the last (sixth) annealing level with stationary distribution $\pi(\cdot)$, is shown in Figure 1(b). Black crosses $\times$ represent the mean vectors $\mu_1, \ldots, \mu_{10}$. As expected, the chain does not exhibit a local random walk behavior and it moves freely between well-separated modes of the posterior distribution.

Let us now compare the performance of AIMS with the Random Walk Metropolis-Hastings algorithm. For a fair comparison, the Metropolis-Hastings algorithm was implemented as follows. First, a sample of $N_0 = 10^3$ points $\theta_0^{(1)}, \ldots, \theta_0^{(N_0)}$ was drawn from the prior distribution $\pi_0(\cdot) = \mathcal{U}_{[0,a]\times[0,a]}(\cdot)$ and the corresponding values of the likelihood function $L(\theta) = \sum_{i=1}^{M} w_i \mathcal{N}(\theta|\mu_i, \sigma^2\mathbb{I}_2)$ were calculated, $L_i = L(\theta_0^{(i)})$. Then, starting from the point with the largest likelihood, $\theta^{(1)} = \theta_0^{(k)}$, $k = \arg\max L_i$, a Markov chain $\theta^{(1)}, \ldots, \theta^{(N)}$, with stationary distribution $\pi(\cdot)$ was generated using the Metropolis-Hastings algorithm. The proposal distribution used was $q(\cdot|\xi) = N(\cdot|\xi, c^2\mathbb{I}_2)$ with $c = 0.2$, and the length of the chain was $N = 5 \cdot 10^3$. Thus, the total number of samples used in both AIMS (with six annealing levels) and RWMH was $N_t = 6 \cdot 10^3$. The scatterplot of posterior samples obtained from RWMH and the trajectory of the corresponding Markov chain are show in Figures 1(c) and 1(d), respectively. While the AIMS algorithm successfully sampled all 10 modes with the approximately correct proportion of total samples, RWHM completely missed 7 modes.

Suppose that we are interested in estimating the posterior mean vector, $\mu^\pi = (\mu_1^\pi, \mu_2^\pi)$, and the components $(\sigma_1^\pi)^2, (\sigma_2^\pi)^2, \sigma_{12}^\pi$ of the posterior covariance matrix Σ^π. Their true values are given in Table 1 along with the AIMS estimates in terms of their means and coefficients of variation averaged over 50 independent simulations, all based on 10^3 posterior samples.

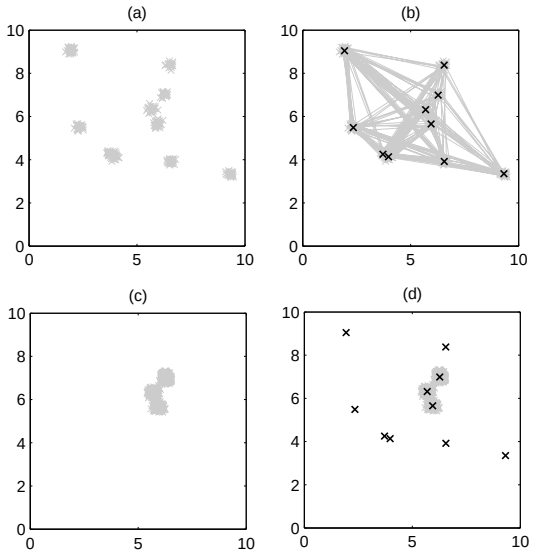

Figure 1. (a) Scatterplots of 10^3 posterior samples; (b) the trajectories of the corresponding posterior Markov chain obtained from AIMS; and (c), (d) corresponding plots from RWMH. Black crosses $\times$ represent the modes $\mu_1, \ldots, \mu_{10}$.

Parameter	μ_1^π	μ_2^π	$(\sigma_1^\pi)^2$	$(\sigma_2^\pi)^2$	σ_{12}^π
True value	5.23	5.75	4.51	3.37	-1.30
AIMS mean	5.20	5.73	4.56	3.32	-1.25
AIMS cov	2.4%	2.0%	8.2%	8.2%	27.7%

Table 1. True values of the posterior parameters and the AIMS estimates in terms of their means and coefficients of variation averaged over 50 simulations.

CONCLUDING REMARKS

In this paper, a new scheme for Bayesian inference, called Asymptotically Independent Markov Sampling (AIMS), is introduced. The algorithm is based on three widely-used simulation methods: importance sampling, MCMC, and simulated annealing. The key idea behind AIMS is to use N samples drawn from $\pi_{j-1}(\cdot)$ as an importance sampling density to construct an approximation $\hat{\pi}_j^N(\cdot)$ of $\pi_j(\cdot)$, where $\pi_0(\cdot), \ldots, \pi_m(\cdot)$ is a sequence of intermediate distributions interpolating between the prior $\pi_0(\cdot)$ and posterior $\pi(\cdot) = \pi_m(\cdot)$. This approximation is then employed as the independent proposal distribution for sampling from $\pi_j(\cdot)$ by the independent Metropolis-Hastings algorithm. When $N \to \infty$, the AIMS sampler generates independent draws from the target distribution, hence the name of the algorithm. The efficiency of AIMS is demonstrated with an example which include multi-modal target distribution.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under award number EAR-0941374 to the California Institute of Technology. This support is gratefully acknowledged. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation.

REFERENCES

Beck, J. L., and Zuev, K. M. (2013). "Asymptotically independent Markov sampling: a new Markov chain Monte Carlo scheme for Bayesian inference." *Int. J. Uncertainty Quantification*, 3(5), 445–474.

Ching, J., and Chen, Y.-C. (2007). "Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection, and model averaging." *Journal of Engineering Mechanics*, 133(7), 816–832.

Kahn, H., and Marshall, A. W. (1953). "Methods of reducing sample size in Monte Carlo computations." *Journal of the Operations Research Society of America*, 1(5), 263–278.

Kong, A., Liu, J. S., and Wong, W. H. (1994). "Sequential imputations and Bayesian missing data problems." *Journal of the American Statistical Association*, 89(425), 278-288.

Neal, R. M. (2001). "Annealed importance sampling." *Statistics and Computing*, 11, 125–139.

Robert, C. P., and Casella, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. Springer Texts in Statistics.