Technical Report No. 0101 , Department of Statistics, University of Toronto

Improving Markov chain Monte Carlo Estimators by Coupling to an Approximating Chain

Ruxandra L. Pinto * Radford M. Neal †

February 23, 2001

Abstract. We show how large improvements in the accuracy of MCMC estimates for posterior expectations can sometimes be obtained by coupling a Markov chain that samples from the posterior distribution with a chain that samples from a Gaussian approximation to the posterior. Use of this method requires a coupling scheme that produces high correlation between the two chains. An efficient estimator can then be constructed that exploits this correlation, provided an accurate value for the expectation under the Gaussian approximation can be found, which for simple functions can be done analytically. Good coupling schemes are available for many Markov chain samplers, including Gibbs sampling with standard conditional distributions. For many moderate-dimensional problems, the improvement in accuracy using this method will be much greater than the overhead from simulating a second chain.

1 Introduction

Bayesian inference problems require calculation of the expectations of various functions of the model parameters with respect to their posterior distribution. If the posterior density, $f(y)$, is easy to sample from, the expectation of $a(y)$ with respect to f can be estimated using Monte Carlo integration by $\overline{a}_y = (1/n)\sum_{i=1}^{n} a(y_i)$, where $y_1, \ldots, y_n$ is a sample of n independent points drawn from f.

Drawing samples directly from the posterior distribution is not feasible in most Bayesian inference problems because the posterior distribution, f, is usually too complicated. One of the early solutions for this problem was to find a Gaussian approximation, g, for f and use $E_g(a(y))$ as an approximation to $E_f(a(y))$. This reduces the problem to calculating the expected value of the function a with respect to a Gaussian distribution, which, depending on the function a, may be doable analytically, or by Gaussian quadrature (Thisted 1996, Section 5.3), or by efficient Monte Carlo techniques. Another possible solution is importance sampling (Tanner 1993, Section 3.3.3), perhaps using the Gaussian approximation to the posterior distribution. A sample from the Gaussian distribution is drawn and the points of the sample are reweighted to account for the fact that the sample is not from the correct distribution.

*Department of Statistics, University of Toronto, Toronto, Ontario, Canada, M5S 3G3. Email: ruxandra@utstat.toronto.edu, Web: http://www.utstat.toronto.edu/~ruxandra/

†Department of Statistics and Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, M5S 3G3. Email: radford@utstat.toronto.edu, Web: http://www.cs.toronto.edu/~radford/

In many problems, the Gaussian approximation will be close to the posterior distribution, but not close enough to provide sufficiently accurate estimates. If the posterior distribution has heavier tails than the Gaussian approximation, even importance sampling will not provide good estimates, as in this case the importance sampling weights will be very variable, and only a few points from the sample drawn from the Gaussian approximation will contribute to the estimate. For this reason, Markov chain Monte Carlo techniques are now commonly used to estimate expected values with respect to complex or high-dimensional posterior distributions.

In this paper, we will use the Gaussian approximation to the posterior distribution to improve the accuracy of Markov chain Monte Carlo estimates. The mean of the Gaussian approximation is taken to be the mode of the posterior distribution. The mode can be found using the Newton-Raphson algorithm, for example, perhaps using as an initial value the sample mean, $\overline{y} = (1/n)\sum_{i=1}^{n} y_i$, where $y_1, \ldots, y_n$ is a sample generated by simulating a Markov chain that converges to f. The variance-covariance matrix for the Gaussian approximation is chosen to be minus the inverse of the Hessian (matrix of second derivatives) of the logarithm of the posterior density, evaluated at the mode of the posterior distribution.

The Markov chain used to generate the sample $y_1, \ldots, y_n$ from f will be coupled with a chain that converges to the Gaussian approximation, g, producing a sample $x_1, \ldots, x_n$. We hope that these two samples will be highly correlated. To take advantage of this correlation, we construct new estimators for $E_f(a(y))$ that depend on both the y's and the x's and which make use of $E_g(a(x))$, which is assumed to be accurately known.

One such estimator is

$$\overline{a}_y - \alpha(\overline{a}_x - E_g(a(x))), \tag{1}$$

where $\overline{a}_x = (1/n)\sum_{i=1}^{n} a(x_i)$. For $\alpha = Cov\,(\overline{a}_y, \overline{a}_x)/Var\,(\overline{a}_x)$, this is the best unbiased linear estimator. In practice, α will have to be determined from the data, introducing some small bias. This new estimator is sometimes much more accurate than $\overline{a}_y$, due to the information provided by the sample drawn from the Gaussian approximation, which for problems of moderate dimensionality can be found with little computational effort.

In the context of simple Monte Carlo estimation from simulation data, a similar technique has been used to reduce variance using control variates (Kelijnen 1974, Section III.4; Ripley 1987). Cheng (1978) investigates the properties of estimators of type (1) when the joint distribution for x and y is Gaussian, an assumption that seems to be reasonable in their queueing system context, but perhaps not for our application. Lavenberg, Moeller and Welch (1982) discuss the loss of variance reduction due to estimating α from the data. Another use of coupling to improve estimation is due to Frigessi, Gåsemyr and Rue (2000), who use antithetic coupling of two chains sampling from the same distribution.

In Section 2, we introduce the coupling procedure and show how it can produce correlation between chains. Section 3 presents the estimator (1) and discusses its properties and efficiency. In Section 4, this estimator is seen as the simplest of a larger class of estimators that can also model more complex relationships between the two chains. Section 5 presents an example based on the data on pump failures from Gelfand and Smith (1990). We conclude, in Section 6, by discussing possible further extensions and applications.

2 The Coupling Technique

Two chains are coupled when their transitions are determined by the same random numbers. Suppose we have two distributions, g and f, from which we want to draw samples $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$, respectively. We will start with $x_0 = y_0$. At each iteration we randomly draw v_t from some distribution V and generate the updates for the two chain by $x_t = \phi_g(x_{t-1}, v_t)$ and $y_t = \phi_f(y_{t-1}, v_t)$. The transitions functions, ϕ_f and ϕ_g, take two inputs, the state at time $t-1$ and some randomness v_t, and return the state at time t. These transition functions are chosen to keep the target distributions, f and g, invariant. From the chains obtained, an initial burn-in period that depends on how fast the chains converge to their joint stationary distribution will be discarded.

Coupling is used in Propp and Wilson's (1996) coupling from the past method of exact sampling. In this context, chains are started from all possible starting points and are run using the same transition probability function. Under certain assumptions, if we start the chains far enough back in the past, they will all coalesce by time zero.

We couple two chains using different transition functions, sampling from similar but different distributions. We start both chains from the same initial state and hope the chains will stay close together for the whole run, producing high correlation between the states of the two chains. The success of coupling in producing chains that move together depends on the Markov chain Monte Carlo methods used to sample from the distributions, and on the way they are expressed in terms of ϕ functions.

To illustrate how coupling works, and later the properties of the estimators we introduce, we consider a toy example in which a gamma distribution is approximated with a Gaussian distribution. In Figure 1, we can see the effect of coupling a chain sampling from the Gamma(10,5) distribution with a chain sampling from its Gaussian approximation. If we denote by f the density of the Gamma distribution, then the parameters of the Gaussian approximation, g, are the mean μ, set to the mode of f, and the variance $\sigma^2 = -\left[d^2 \log f(x)/dx^2\right]^{-1}$ evaluated at the mode. We used the Metropolis algorithm to sample from the Gamma and Gaussian distributions. The proposal used by the Metropolis algorithm was a Gaussian distribution centered at the current point and with standard deviation three. The coupling for this example is done by using the same Gaussian random numbers for the proposals and the same uniform numbers for the accept-reject decisions. The random noise $v_t = (n_t, u_t)$ has two components, $n_t \sim N(0, 3^2)$ and $u_t \sim$ Uniform(0,1), and the two deterministic functions ϕ_f and ϕ_g are defined by

$$\phi_f(y_{t-1}, (n_t, u_t)) = \begin{cases} y_{t-1} + n_t & \text{if } u_t < f(y_{t-1} + n_t)/f(y_{t-1}) \\ y_{t-1} & \text{otherwise} \end{cases} \tag{2}$$

and similarly

$$\phi_g(x_{t-1}, (n_t, u_t)) = \begin{cases} x_{t-1} + n_t & \text{if } u_t < g(x_{t-1} + n_t)/g(x_{t-1}) \\ x_{t-1} & \text{otherwise} \end{cases} \tag{3}$$

where f is the gamma density $f(x|\alpha, \beta) = (x^{\alpha-1}e^{-x/\beta})/(\Gamma(\alpha)\beta^\alpha)$ with parameters $\alpha = 10$ and $\beta = 5$, and g is the Gaussian density with $\mu = \beta(\alpha - 1)$ and $\sigma^2 = \beta^2(\alpha - 1)$.

How high the correlation between chains will be depends on the coupling technique and on how close the Gaussian approximation is to the posterior distribution. In this example, the coupling is

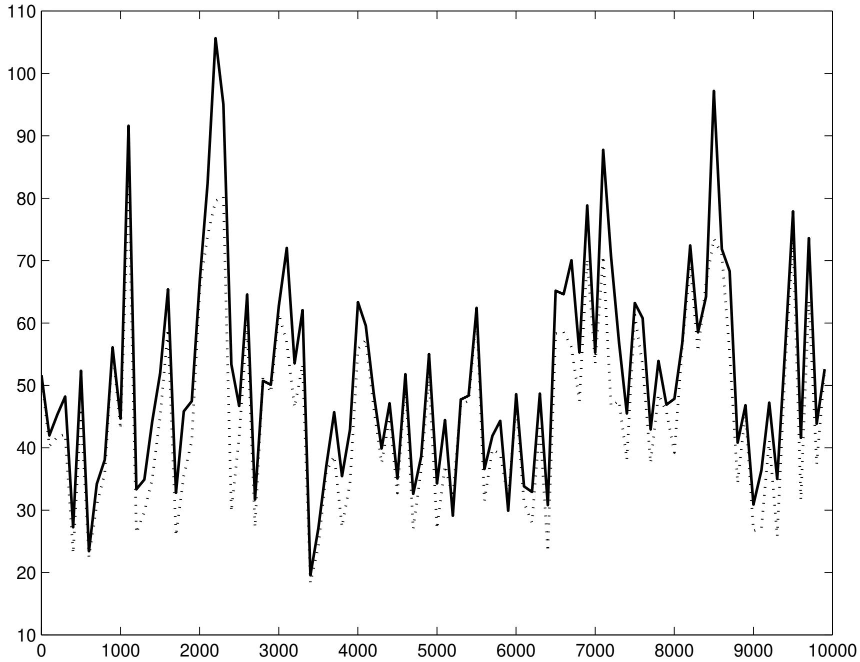

Figure 1: Coupling of chains sampling from Gamma(α, β) and the Gaussian approximation with mean $\mu = \beta(\alpha - 1)$ and variance $\sigma^2 = \beta^2(\alpha - 1)$. Here $\alpha = 10$ and $\beta = 5$. Every hundredth point of a long run of the Metropolis chains is plotted for each distribution. The solid line is the sample from the Gamma distribution and the dotted line is the Gaussian approximation.

good because the step size is small — smaller than would be optimal if only one chain were being simulated. For realistic problems, methods that produce good coupling at less cost are needed, as illustrated in the example of Section 5.

3 A Simple Estimator Exploiting Coupling between Two Chains

Assume we have samples from two coupled chains, $(y_1, \ldots, y_n)$ from the distribution f, and $(x_1, \ldots, x_n)$ from g, the Gaussian approximation to f. The usual estimator for $E_f(a(y))$ would be $\overline{a}_y = (1/n) \sum_{i=1}^n a(y_i)$. To exploit the correlation of the two chains, we can construct a new estimator for $E_f(a(y))$ of the form:

$$\overline{a}_y - \alpha(\overline{a}_x - E_g(a(x))), \tag{4}$$

where $\overline{a}_x = (1/n) \sum_{i=1}^n a(x_i)$. This estimator is unbiased for any fixed α. It has minimum variance (Ripley, 1987, Section 5.3) when

$$\alpha = \frac{Cov(\overline{a}_y, \overline{a}_x)}{Var(\overline{a}_x)}. \tag{5}$$

If the pairs of points from the two chains were independent, the appropriate estimate for α would be

$$\widehat{\alpha} = \frac{\sum_{i=1}^n (a(y_i) - \overline{a}_y)(a(x_i) - \overline{a}_x)}{\sum_{i=1}^n \left(a(x_i) - \overline{a}_x\right)^2}. \tag{6}$$

For samples of dependent pairs obtained using Markov chains, we will still use estimator (6), since it is close to optimal.

For notational simplicity, assume we are interested in $\mu_f = E_f(y)$ and we know $\mu_g = E_g(x)$. The first estimate for μ_f we will look at is:

$$\hat{\mu}_f^{(1)} = \overline{y} - \hat{\alpha}(\overline{x} - \mu_g) \tag{7}$$

where, applying (6),

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}. \tag{8}$$

Estimator (7) is unbiased if α is not estimated from the data. When α is estimated from the data, we can still prove that the estimator is consistent. If the two chains are ergodic, the ergodic theorem helps us establish the following:

$$\overline{y} \rightarrow_p \mu_f \tag{9}$$

$$\overline{x} \rightarrow_p \mu_g \tag{10}$$

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} \rightarrow_p E_g((x - \mu_g)^2) \tag{11}$$

$$\frac{\sum_{i=1}^{n} x_i y_i}{n} \rightarrow_p E_{fg}(xy) \tag{12}$$

The last statement is justified if the joint coupled chain (x, y) is ergodic. However, for the purpose of this proof, all that is required is that the joint coupled chain converges to some distribution, such that (12) converges to some constant. From (9) to (12) it follows that α converges to a constant and that

$$\hat{\mu}_f^{(1)} \rightarrow_p \mu_f. \tag{13}$$

The asymptotic efficiency of this estimator can be investigated by considering α to be constant. If we write the estimator (7) as $\hat{\mu}_f^{(1)} = (1/n)\sum_{i=1}^{n} z_i$, where $z_i = y_i - \hat{\alpha}(x_i - \mu_g)$, we can estimate its variance by:

$$\widehat{Var}\left(\hat{\mu}_f^{(1)}\right) = \frac{\sum_{i=1}^{n}\left(z_i - \hat{\mu}_f^{(1)}\right)^2}{n - \hat{\tau}} \frac{\hat{\tau}}{n}, \tag{14}$$

where $\hat{\tau}$ is the estimated autocorrelation time, which is obtained by summing the autocorrelations of $(z_1, \ldots, z_n)$ at all positive and negative lags up to the lag past which the autocorrelations seem to be nearly zero.

For the example presented in Section 2, we ran two coupled chains 100,000 iterations long and found that the autocorrelations were close to 0 past lag 350. The correlation between the two chains is 0.9466 and $\hat{\alpha} = .9926$. The parameter estimated here is the mean of Gamma(10,5), which is 50. The estimate using (7) is 50.21 with standard error of 0.22 and the traditional estimate is 49.34 with standard error of 0.63. These results show that exploiting the correlations between chains improves the efficiency of the estimator by a factor of $(0.63/0.22)^2 \approx 8$.

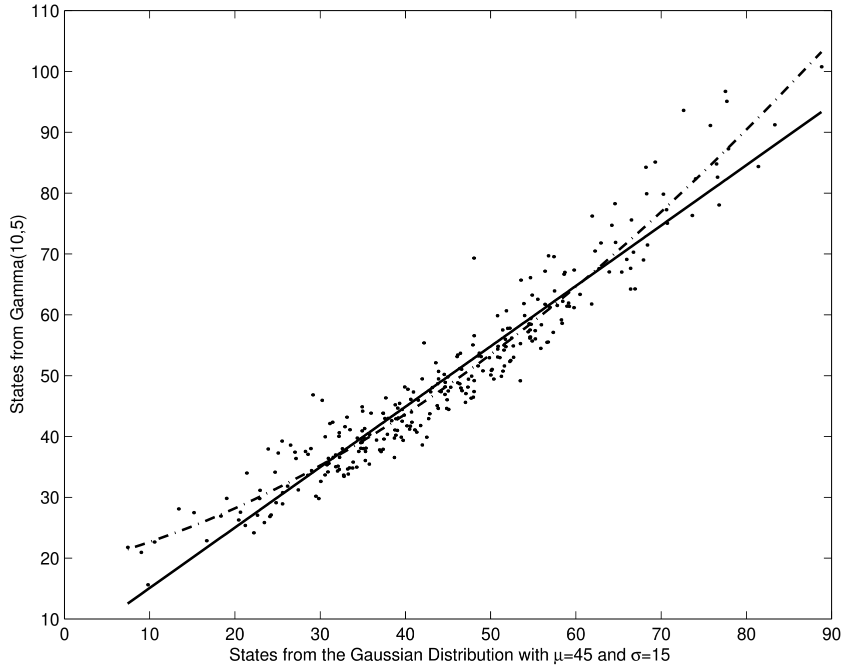

Figure 2: Every 350th point of the chains from Gamma(10,5) and the Gaussian approximation with $\mu = 45$ and $\sigma = 15$, along with the regression lines for first and third order models.

4 Coupled estimators based on regression models

The relationship between the samples $y_1, \ldots, y_n$ and $x_1, \ldots, x_n$ could be modeled by a simple linear regression:

$$y_i = \beta_0 + \beta_1(x_i - \mu_g) + \epsilon_i \tag{15}$$

The estimator (7) is exactly the least square estimator for the intercept in this simple linear regression model. This observation leads us to consider new estimators for μ_f based on higher-order regression models. The samples based on the two coupled chains in Section 2 will not be linearly related because the upper tail of the Gamma distribution is heavier than that of the Gaussian distribution. In Figure 2 we can see that a third-order regression model fits better than the linear model.

Assume we fit the following model for how $(y_1, \ldots, y_n)$ relates to $(x_1, \ldots, x_n)$:

$$y_i = \beta_0 + \beta_1(x_i - \mu_g) + \beta_2(x_i - \mu_g)^2 + \beta_3(x_i - \mu_g)^3 + \epsilon_i \tag{16}$$

If $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3$ are the least square estimates for the regression coefficients, we propose the following estimator for μ_f:

$$\widehat{\mu}_f^{(3)} = \overline{y} - \widehat{\beta}_1\overline{(x - \mu_g)} + \widehat{\beta}_2\sigma_g^2 - \widehat{\beta}_2\overline{(x - \mu_g)^2} - \widehat{\beta}_3\overline{(x - \mu_g)^3} \tag{17}$$

where $\sigma_g^2 = E\left((x - \mu_g)^2\right)$, $\overline{(x - \mu_g)} = (1/n)\sum_{i=1}^n (x_i - \mu_g)$, $\overline{(x - \mu_g)^2} = (1/n)\sum_{i=1}^n (x_i - \mu_g)^2$, and $\overline{(x - \mu_g)^3} = (1/n)\sum_{i=1}^n (x_i - \mu_g)^3$. By using the ergodic theorem, we can establish that the coefficients, $\widehat{\beta}_i$, converge to some constants, and that

$$\overline{(x-\mu_g)} \quad \rightarrow_p \quad 0 \tag{18}$$

$$\overline{(x-\mu_g)^2} \quad \rightarrow_p \quad \sigma_g^2 \tag{19}$$

$$\overline{(x-\mu_g)^3} \quad \rightarrow_p \quad 0 \tag{20}$$

$$\overline{y} \quad \rightarrow_p \quad \mu_f \tag{21}$$

It follows that $\widehat{\mu}_f^{(3)} \rightarrow_p \mu_f$.

Letting $z_i = y_i - \widehat{\beta}_1(x_i-\mu_g) + \widehat{\beta}_2\sigma_g^2 - \widehat{\beta}_2(x_i-\mu_g)^2 - \widehat{\beta}_3(x_i-\mu_g)^3$, we can write $\widehat{\mu}_f^{(3)} = (1/n)\sum_{i=1}^n z_i$, and estimate its variance by:

$$\widehat{Var}\left(\widehat{\mu}_f^{(3)}\right) \quad = \quad \frac{\sum_{i=1}^n \left(z_i - \widehat{\mu}_f^{(3)}\right)^2}{n-\widehat{\tau}} \frac{\widehat{\tau}}{n} \tag{22}$$

where $\widehat{\tau}$ is the estimated autocorrelation time for the z_i.

We expect this estimator to be better because the z_i's are the residuals of the model plus the constant $\widehat{\beta}_0 + \widehat{\beta}_1\sigma_g^2$. The better the model we fit, the smaller the variance of the residuals and hence the smaller the variance of $\widehat{\mu}_f^{(3)}$. Note, however, that the estimators based on these methods are valid even when the regression model is not correct.

For the gamma example presented in Section 2, this third-order regression estimator, $\widehat{\mu}_f^{(3)}$, gives an estimate for μ_f of 50.20 with standard error 0.18. This estimator is about 1.5 times better than the linear regression estimate, $\widehat{\mu}_f^{(1)}$, and about 12 times more efficient than the standard estimator based on one chain.

5 Pumps Data Example

To illustrate the performance of the new coupled estimators we will consider the model for data on pump failures from Gelfand and Smith (1990). The data consists of p counts, $s_1,\ldots,s_p$, that represent the number of failures for p pumps over known periods of time, $t_1,\ldots,t_p$. It is assumed that conditionally on unknown failure rates $\lambda_1,\ldots,\lambda_p$, the counts $s_1,\ldots s_p$ are independent and Poisson distributed with means $\lambda_i t_i$. Given β, the unknown $\lambda_i^{,}s$ are independent, and each has a gamma distribution with a known shape parameter α and unknown scale factor β. The hyperparameter β is assumed to have an inverse gamma distribution with a known shape parameter γ and a known scale parameter δ.

The densities for this hierarchical Bayesian model are given by:

$$P(s_i|\lambda_i) \quad = \quad e^{-\lambda_i t_i}\frac{(\lambda_i t_i)^{s_i}}{s_i!},\ i=1,\ldots,p \tag{23}$$

$$P(\lambda_i|\alpha,\beta) \quad = \quad \frac{1}{\Gamma(\alpha)\beta^\alpha}\lambda_i^{\alpha-1}e^{-\lambda_i/\beta},\ i=1,\ldots,p \tag{24}$$

$$P(\beta|\gamma,\delta) \quad = \quad \frac{\delta^\gamma}{\Gamma(\gamma)}\frac{1}{\beta^{\gamma+1}}e^{-\delta/\beta} \tag{25}$$

The joint conditional distribution of the parameters of interest, $\beta, \lambda_1, \ldots \lambda$, given the observed $s_1, \ldots, s_p$ and $t_1, \ldots, t_p$ is:

$$\begin{aligned} P(\beta, \lambda_1, \ldots, \lambda_p | s_1, \ldots s_p) &\propto P(\beta, \lambda_1, \ldots, \lambda_p, s_1, \ldots s_p) &(26)\\ &\propto P(s_1, \ldots s_p | \lambda_1, \ldots \lambda_p) P(\lambda_1, \ldots, \lambda_p | \beta) P(\beta) &(27)\\ &\propto \prod_{i=1}^{p} \left(e^{-\lambda_i t_i} \frac{(\lambda_i t_i)^{s_i}}{s_i!} \right) \prod_{i=1}^{p} \left(\frac{1}{\beta^\alpha} \lambda_i^{\alpha-1} e^{-\lambda_i/\beta} \right) \frac{1}{\beta^{\gamma+1}} e^{-\delta/\beta} &(28) \end{aligned}$$

It is more convenient to work in terms of $\theta = 1/\beta$. The joint conditional density for $\theta, \lambda_1, \ldots, \lambda_p$, which we will denote by f, is:

$$f(\theta, \lambda_1, \ldots, \lambda_p | s_1, \ldots s_p) \propto \prod_{i=1}^{p} \left(e^{-\lambda_i t_i} (\lambda_i t_i)^{s_i} \right) \prod_{i=1}^{p} \left(\theta^\alpha \lambda_i^{\alpha-1} e^{-\lambda_i \theta} \right) \theta^{\gamma-1} e^{-\delta\theta} \tag{29}$$

We will use Gibbs sampling to sample from the posterior distribution of $\theta, \lambda_1, \ldots, \lambda_p$. The conditional distribution for θ given the λ_i's is Gamma$(p\alpha + \gamma, 1/\left(\delta + \sum_{i=1}^{p} \lambda_i\right))$. The conditional distribution for each λ_i given θ and all $\{\lambda_j\}_{j \neq i}$ is Gamma$(s_i + \alpha, 1/\left(t_i + \theta\right))$. One step of Gibbs sampling at time t is done by generating $u_0^{(t)}, \ldots, u_p^{(t)}$ from the Uniform(0,1) distribution and applying the following deterministic transition functions:

$$\begin{aligned} \theta^{(t)} &= F^{-1}\left(u_0^{(t)};\ p\alpha + \gamma,\ \frac{1}{\delta + \sum_{i=1}^{p} \lambda_i^{(t-1)}} \right) &(30)\\ \lambda_i^{(t)} &= F^{-1}\left(u_i^{(t)};\ s_i + \alpha,\ \frac{1}{t_i + \theta^{(t)}} \right),\quad i = 1, \ldots, p, &(31) \end{aligned}$$

where F^{-1} is the inverse cumulative distribution function for the gamma distribution with shape and scale parameters as specified. This method of generating gamma variates is not the fastest, but it is used to produce good coupling.

The Gaussian approximation, g, for the joint posterior distribution (29) has mean $\boldsymbol{\mu}$ equal to the mode of the posterior distribution and variance-covariance matrix $\boldsymbol{\Sigma}$ equal to minus the inverse of the Hessian of the logarithm of the posterior density evaluated $\boldsymbol{\mu}$. To do Gibbs sampling for the Gaussian approximation, consider $\boldsymbol{x} = (x_0, x_1, \ldots,, x_p)$, a $p+1$ dimensional Gaussian random vector with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Denote by $\boldsymbol{\Omega}$ the inverse of the variance-covariance matrix $\boldsymbol{\Sigma}$, with elements ω_{ij}, $i, j = 0, 1, \ldots, p$. The conditional distribution of x_i given the other components $\{x_j\}_{j \neq i}$ is Gaussian with mean $\mu_i - (1/\omega_{ii}) \sum_{j \neq i} \omega_{ij}(x_j - \mu_j)$ and variance $1/\omega_{ii}$. The coupling of the two chains in this example is done by using at each step the same Uniform(0,1) random numbers. One Gibbs sampling step at time t for the Gaussian approximation is

$$x_i^{(t)} = G^{-1}\left(u_i^{(t)};\ \mu_i - \frac{1}{\omega_{ii}} \left(\sum_{j<i} \omega_{ij}(x_j^{(t)} - \mu_j) + \sum_{j>i} \omega_{ij}(x_j^{(t-1)} - \mu_j) \right),\ \frac{1}{\omega_{ii}} \right),\quad i = 0, \ldots, p \tag{32}$$

where G^{-1} is the inverse cumulative distribution function for the Gaussian distribution with mean and variance as specified.

Following Gelfand and Smith (1990), we set the known parameters to $p = 10,\ \delta = 1,\ \gamma = 0.1$, and $\alpha = \overline{\rho}^2/(s_\rho^2 - p^{-1}\overline{\rho}\sum_{i=1}^p t_i^{-1})$, where $\rho_i = s_i/t_i$, $\overline{\rho} = (1/p)\sum_{i=1}^p \rho_i$, and $s_\rho^2 = (1/p)\sum(\rho_i - \overline{\rho})^2$. We draw a sample from the posterior distribution using fixed starting values $\theta^{(0)} = 1$ and $\lambda_i^{(0)} = s_i/t_i$. Based on this sample we find Monte Carlo estimates for means of the parameters and use them as starting values for the Newton-Raphson method to find the mode, which is used as the mean of the Gaussian approximation. Estimating the parameters of the Gaussian approximation is computationally inexpensive in this example; Newton-Raphson converges to the mode of the posterior distribution in 11 steps using the starting value mentioned.

We are interested in estimating the expected value of each of the parameters $\theta, \lambda_1, \ldots, \lambda_p$ with respect to the posterior distribution. We ran coupled chains 1000 iterations long and discarded the first 100 states of each chain as burn-in, which is more than adequate for this problem, for which the autocorrelations are close to zero by lag 5. The results for three estimators and their standard errors are presented in Table 1. The estimators evaluated are $\overline{y}$, the traditional Markov chain estimator based on only one chain, estimator $\widehat{\mu}^{(1)}$ presented in Section 3, and estimator $\widehat{\mu}^{(3)}$ introduced in Section 4. The last three columns of Table 1 present the relative efficiencies of the estimators as ratios of their estimated variances. For this problem the estimates based on third-order regression are all much more efficient than the estimates based on one chain, or even those based on simple linear regression. Particularly striking is the estimate for λ_1 based on third-order regression, which is approximately 24000 times more efficient than the estimate based on one chain. This is because the relationship between the two chains is very tight, with little variation unexplained in the third-order model, as seen in Figure 4. From Table 2 we can see that the correlations between λ_1 and all 10 other components in the Gaussian approximation are close to zero. For the parameter θ, the third-order model still the provides the most efficient estimate, but as we can see in Figure 3, the relationship between the coupled chains is not as tight as for λ_1.

Table 1 shows one more estimate labeled "Precise Estimate", which is obtained by running 200 pairs of coupled chains for 1000 iterations and discarding the first 100 states. We fit a third order model for the first pair of chains and with these fixed coefficients we find 199 estimates based on the other chains using (17). Due to the fact that the coefficients are fixed, these 199 estimates are unbiased. The precise estimate is taken to be the average of these 199 unbiased estimates, and the standard error for this estimate is found using their sample variance. The result is much more accurate than any of the three estimates, and hence can be used to evaluate their accuracy.

For the 200 pairs of coupled chains we also calculated the estimates and standard errors for all the parameters based on simple linear regression and on third-order regression. For each parameter, we constructed 95% and 90% confidence intervals around these estimates by taking as margin of error the estimated standard error multiplied by the corresponding quantile for the standard normal distribution. We then found the coverage probabilities for these confidence intervals, as the proportion of intervals that contain the precise estimate. As seen in Table 3, these coverage probabilities are close to the desired values of 95% and 90%, confirming that the estimators are close to being unbiased and that their standard errors are close to being correct.

As we reduce the length of the chains, we would expect that bias may be present. Also, since the procedure for estimating the standard errors doesn't take into account the variability of the regression coefficients, the standard errors will be underestimated for short chains. Since these problems are expected to be worse when there are many regression coefficients, we recommend using the estimates based on simple linear regression when the chains are short.

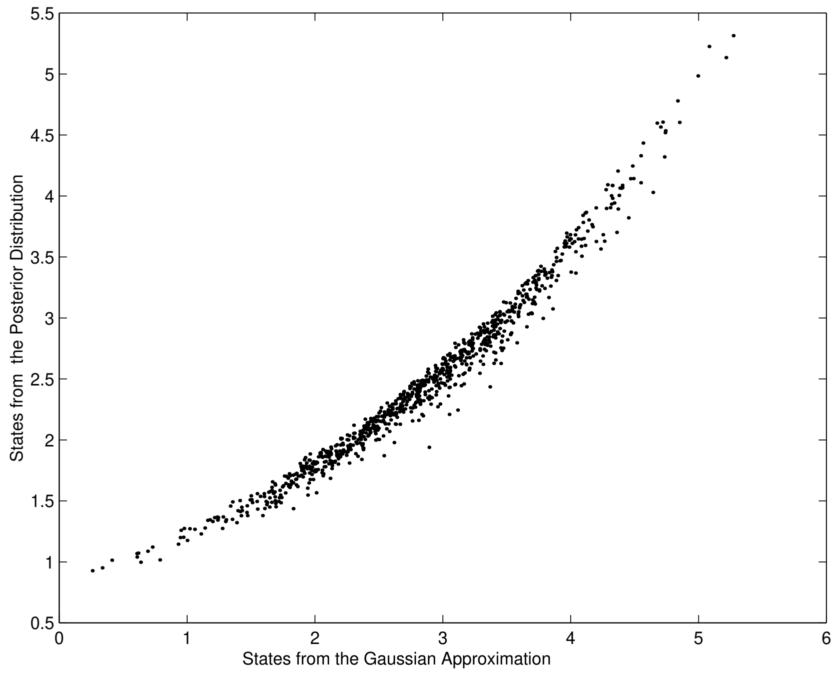

Figure 3: Plot showing the relationship between the θ values for the two coupled chains.

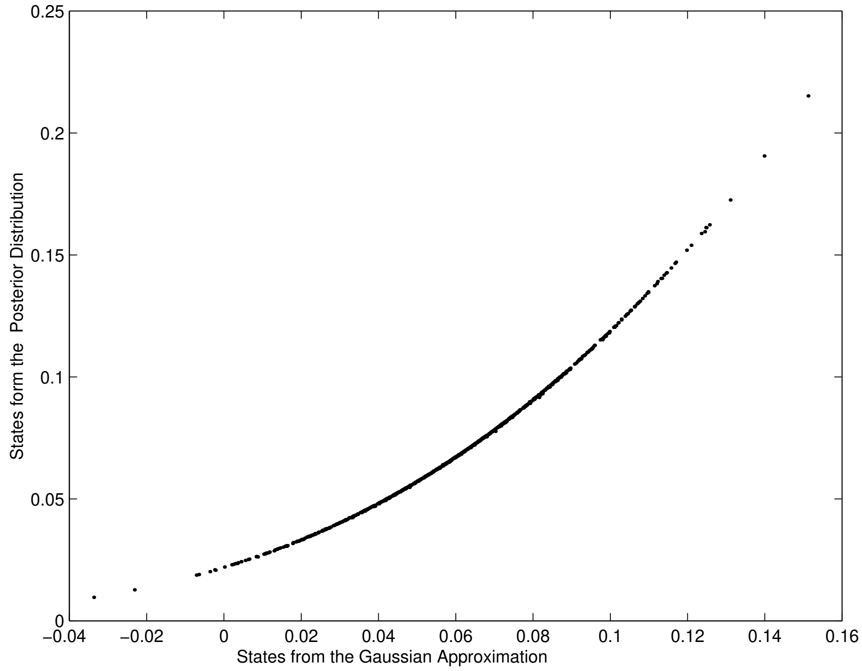

Figure 4: Plot showing the relationship between λ_1 values for the two coupled chains

		Estimate Based on			Relative Efficiency		
Parameter	Precise Estimate	One Chain ($\overline{y}$)	First-order Model ($\widehat{\mu}^{(1)}$)	Third-Order Model ($\widehat{\mu}^{(3)}$)	$\overline{y}$ vs. $\widehat{\mu}^{(1)}$	$\overline{y}$ vs. $\widehat{\mu}^{(3)}$	$\widehat{\mu}^{(1)}$ vs. $\widehat{\mu}^{(3)}$
θ	2.4895321	2.4674625	2.4874835	2.4884441			
	0.0002776	0.0303696	0.0064195	0.0042179	22	52	2.3
λ_1	0.0702695	0.0708782	0.0701613	0.0702695			
	0.0000003	0.0008800	0.0001644	0.0000056	29	24000	850
λ_2	0.1541290	0.1573713	0.1545877	0.1542453			
	0.0000057	0.0028258	0.0007910	0.0000655	13	1900	150
λ_3	0.1040727	0.1052354	0.1039408	0.1040741			
	0.0000007	0.0013133	0.0002370	0.0000120	31	12000	390
λ_4	0.1232198	0.1233812	0.1232257	0.1232186			
	0.0000005	0.0010237	0.0001227	0.0000071	70	21000	300
λ_5	0.6264700	0.6309189	0.6243196	0.6260541			
	0.0000333	0.0098097	0.0021989	0.0004975	20	390	20
λ_6	0.6133804	0.6089641	0.6134456	0.6133659			
	0.0000084	0.0042671	0.0004532	0.0001238	89	1200	13
λ_7	0.8240495	0.8280188	0.8229350	0.8264169			
	0.0001540	0.0191607	0.0054209	0.0019331	12	98	7.9
λ_8	0.8242431	0.8505378	0.8292412	0.8247778			
	0.0001599	0.0193552	0.0071622	0.0021677	7.3	80	11
λ_9	1.2951942	1.3148071	1.2966271	1.2939680			
	0.0000974	0.0224054	0.0046176	0.0014573	24	240	10
λ_{10}	1.8407347	1.8321169	1.8416890	1.8401319			
	0.0000594	0.0143102	0.0017195	0.0008827	69	260	3.8

Table 1: The estimates based on 900 states of the coupled chains and their standard errors,along with the relative efficiencies rounded to two significant digits.

	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_0	1.0000	-0.0205	-0.0596	-0.0302	-0.0247	-0.1952	-0.1065	-0.2737	-0.2737	-0.3437	-0.2836
x_1	-0.0205	1.0000	0.0012	0.0006	0.0005	0.0040	0.0022	0.0056	0.0056	0.0070	0.0058

Table 2: Correlations between x_0 (θ) and x_1 (λ_1) and all the other the components of the Gaussian approximation.

	Coverage probabilities			
	Simple Linear Regression		Third Order Regression	
Parameter	95%	90%	95%	90%
θ	0.95	0.88	0.95	0.89
λ_1	0.92	0.86	0.95	0.91
λ_2	0.92	0.84	0.93	0.84
λ_3	0.95	0.93	0.95	0.91
λ_4	0.95	0.91	0.96	0.92
λ_5	0.93	0.89	0.96	0.92
λ_6	0.97	0.93	0.94	0.89
λ_7	0.96	0.93	0.96	0.89
λ_8	0.91	0.86	0.96	0.91
λ_9	0.93	0.89	0.95	0.89
λ_{10}	0.93	0.88	0.96	0.89

Table 3: The fraction of 95% and 90% confidence intervals that contain the precise estimate. Each confidence interval was determined from an estimate based on a pair of chains 900 iterations long. The standard errors for the coverage probabilities for the 95% confidence intervals are 0.015; for the 90% confidence intervals, the standard errors are 0.021.

6 Discussion

We have shown that estimators based on coupling to a chain that samples from a Gaussian approximation can be much more precise than the traditional Markov chain Monte Carlo estimators based on one chain. This method is applicable to those Bayesian problems for which the posterior distribution can be approximated reasonably well with a Gaussian distribution. The success of this method is related to the coupling technique used. The two sampling techniques used in this paper, Gibbs sampling and the Metropolis algorithm, both have computational drawbacks. Gibbs sampling seems to produce samples that are highly correlated, but at the expense of having to compute inverse cumulative distribution functions, which for some conditional distributions might be expensive. Moreover, for more difficult problems, the conditional distributions will not be available, and therefore Gibbs sampling will not be applicable. For the Metropolis algorithm, the inefficiency is introduced by the small step size needed to keep the rejection rate small and therefore the correlation between chains high. These inefficiencies can be avoided by using other sampling techniques, such as higher-order Langevin methods, which can produce very low rejection rates using reasonable step sizes.

Finding the mean and the variance-covariance matrix for the Gaussian approximation requires time of order m^3, where m is the number of parameters, and one step Gibbs sampling requires time of order m^2. The methods of this paper are therefore probably not useful when m is more than a few hundred.

Methods similar to those we have presented can be applied to problems where samples from several similar distributions are needed. These problems occur when assessing the effect on the posterior distribution of deleting observations (Peruggia 1997) or changing the prior or likelihood (Gelman, *et al.* 1995, Chapter 12). These authors use importance sampling to obtain estimates for expected values with respect to all distributions by drawing a sample from one of the distributions, and then reweighting these sample points to reflect the other distributions. Unfortunately, the importance weights can vary wildly when the distributions are too different. We propose the following strategy for sampling from many different, but similar, distributions. Simulate a long Markov chain converging to one of the distributions, from which a precise estimate for the expected values of the parameters with respect to this distribution can be found. For the other distributions, run short chains coupled with the first part of the long chain, and then use the methods presented in Sections 3 and 4 to find accurate estimates of the expected values of the parameters with respect to these other distributions, taking advantage of the precise estimates from the long chain.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada and by the Institute for Robots and Intelligent Systems. The first author was supported by an Ontario Graduate Scholarship.

References

Cheng, R. C. H. (1978) "Analysis of simulation experiments under normality assumptions", *Journal of Operational Research Society*, vol. 29, pp. 493-497.

Frigessi, A., Gåsemyr, J. and Rue, H. (2000) "Antithetic coupling of two Gibbs sampler chains", *Annals of Statistics*, vol. 28.

Gelfand, A. E and Smith, A. F. M. (1990) "Sampling-based approaches to calculating marginal densities", *Journal of American Statistical Association*, vol. 85, pp. 398-409.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995) *Bayesian Data Analysis*, London: Chapman & Hall.

Kelijnen, J. P. C. (1974) *Statistical Techniques in Simulation*, New York: Marcel Dekker.

Lavenberg, S. S., Moeller, T. L. and Welch, P. D. (1982) "Statistical Results on Control Variables With Application to Queueing Network Simulation", *Operations Research*, vol. 30, pp. 182-202.

Peruggia, M. (1997) "On the variability of case-deletion Importance sampling Weights in the Bayesian linear model", *Journal of the American Statistical Association*, vol. 92, pp. 199-207.

Propp, J. C. and Wilson, D. B. (1996) "Exact sampling with coupled Markov chains and applications to statistical mechanics", *Random Structures and Algorithms*, vol. 9, pp. 223-252.

Ripley, B. D. (1987) *Stochastic Simulation*, New York: Wiley.

Tanner, M. A. (1993) *Tools for Statistical Inference*, Second Edition, New York: Springer Verlag.

Thisted, R. A. (1988) *Elements of Statistical Computing*, Chapman and Hall/CRC.