

# Chapter 1

## The data augmentation algorithm: Theory and methodology

*James P. Hobert*

### 1.1 Basic Ideas and Examples

Assume that the function  $f_X : \mathbb{R}^p \rightarrow [0, \infty)$  is a probability density function (pdf). Suppose that  $g : \mathbb{R}^p \rightarrow \mathbb{R}$  is a function of interest and that we want to know the value of  $E_{f_X} g = \int_{\mathbb{R}^p} g(x) f_X(x) dx$ , but this integral cannot be computed analytically. There are many ways of approximating such intractable integrals and these include numerical integration, analytical approximations and Monte Carlo methods. In this chapter, we will describe a Markov chain Monte Carlo (MCMC) method called the *data augmentation (DA) algorithm*.

Here's the basic idea. In situations where classical Monte Carlo methods are not applicable because it is impossible to simulate from  $f_X$  directly, it is often possible to find a joint pdf

$f : \mathbb{R}^p \times \mathbb{R}^q \rightarrow [0, \infty)$  that satisfies two properties: (i) the  $x$ -marginal is  $f_X$ , that is,

$$\int_{\mathbb{R}^q} f(x, y) dy = f_X(x) ,$$

and (ii) simulating from the associated conditional pdfs,  $f_{X|Y}(x|y)$  and  $f_{Y|X}(y|x)$ , is straightforward. The DA algorithm is based on this joint pdf. The first property allows for the construction of a Markov chain that has  $f_X$  as an invariant pdf, and the second property provides a means of simulating this Markov chain. As long as the resulting chain is reasonably well behaved, simulations of it can be used to consistently estimate  $E_{f_X}g$ . We now begin to fill in the details, starting with the construction of the Markov chain.

As usual, let  $f_Y(y) = \int_{\mathbb{R}^p} f(x, y) dx$ . Also, define  $\mathbf{X} = \{x \in \mathbb{R}^p : f_X(x) > 0\}$  and  $\mathbf{Y} = \{y \in \mathbb{R}^q : f_Y(y) > 0\}$  and assume that  $f(x, y) = 0$  whenever  $(x, y) \notin \mathbf{X} \times \mathbf{Y}$ . Now define a function  $k : \mathbf{X} \times \mathbf{X} \rightarrow [0, \infty)$  as follows

$$k(x'|x) = \int_{\mathbf{Y}} f_{X|Y}(x'|y)f_{Y|X}(y|x) dy . \quad (1.1.1)$$

(We will not need to perform the integration in (1.1.1) - remember that we're still in the construction phase.) Since the integrand in (1.1.1) is a product of conditional densities,  $k$  is never negative. Furthermore,

$$\begin{aligned} \int_{\mathbf{X}} k(x'|x) dx' &= \int_{\mathbf{X}} \left[ \int_{\mathbf{Y}} f_{X|Y}(x'|y)f_{Y|X}(y|x) dy \right] dx' \\ &= \int_{\mathbf{Y}} f_{Y|X}(y|x) \left[ \int_{\mathbf{X}} f_{X|Y}(x'|y) dx' \right] dy \\ &= \int_{\mathbf{Y}} f_{Y|X}(y|x) dy \\ &= 1 . \end{aligned}$$

Hence, for each fixed  $x \in \mathbf{X}$ ,  $k(x'|x)$  is nonnegative and integrates to 1. The function  $k$  is therefore a Markov transition density (Mtd) that defines a Markov chain,  $X = \{X_n\}_{n=0}^{\infty}$ , with state space  $\mathbf{X}$ . The chain evolves as follows. If the current state of the chain is  $X_n = x$ , then the density of the next state,  $X_{n+1}$ , is  $k(\cdot|x)$ . This Markov chain is the basis of the DA algorithm and we now describe some of its properties.

The product  $k(x'|x)f_X(x)$  is symmetric in  $(x, x')$ . Indeed,

$$k(x'|x)f_X(x) = f_X(x) \int_{\mathcal{Y}} f_{X|Y}(x'|y)f_{Y|X}(y|x) dy = \int_{\mathcal{Y}} \frac{f(x', y)f(x, y)}{f_Y(y)} dy .$$

Thus, for all  $x, x' \in \mathcal{X}$ ,

$$k(x'|x)f_X(x) = k(x|x')f_X(x') , \tag{1.1.2}$$

which implies that the Markov chain  $X$  is *reversible* with respect to  $f_X$  (see, e.g., Ross, 1996, Section 4.7). Equation (1.1.2) is sometimes called the *detailed balance condition*. Integrating both sides of (1.1.2) with respect to  $x$  yields

$$\int_{\mathcal{X}} k(x'|x)f_X(x) dx = f_X(x') , \tag{1.1.3}$$

which shows that  $f_X$  is an *invariant density* for the Markov chain  $X$ . What does it mean for  $f_X$  to be invariant for  $X$ ? To answer this question, note that the integrand in (1.1.3) is the joint density of  $(X_0, X_1)$  when the starting value,  $X_0$ , is drawn from  $f_X$ . Thus, equation (1.1.3) implies that, when  $X_0 \sim f_X$ , the marginal density of  $X_1$  is also  $f_X$ . Actually, since  $X$  is a time homogeneous Markov chain, equation (1.1.3) also implies that, if  $X_n \sim f_X$ , then  $X_{n+1} \sim f_X$ . Hence, a simple induction argument leads to the conclusion that, if  $X_0 \sim f_X$ , then the marginal density of  $X_n$  is  $f_X$  for all  $n$ . In other words, when  $X_0 \sim f_X$ , the Markov chain  $X$  is a sequence of *dependent* random vectors with density  $f_X$ . Of course, in practice, it will not be possible to start the chain by drawing  $X_0$  from  $f_X$ . (If simulating directly from  $f_X$  is possible, then one should use classical Monte Carlo methods instead of the DA algorithm for the reasons laid out in Subsections 1.2.4 and 1.3.1.) Fortunately, as long as the Markov chain  $X$  is well-behaved (see Section 1.2.1), the marginal density of  $X_n$  will *converge* to the invariant density  $f_X$  no matter how the chain is started. And, more importantly, the estimator  $n^{-1} \sum_{i=0}^{n-1} g(X_i)$  will be strongly consistent for  $E_{f_X} g$ ; that is, this estimator will converge almost surely to  $E_{f_X} g$  as  $n \rightarrow \infty$ .

In order to keep things simple, we are considering only situations where  $f_X$  and  $f(x, y)$  are densities with respect to Lebesgue measure. However, all of the results and methodology that we discuss in this chapter can be easily extended to a much more general setting. See, for example, Section 2 of Hobert and Marchev (2008).

Now consider the practical issue of simulating the Markov chain  $X$ . Given that the current state of the chain is  $X_n = x$ , how do we draw  $X_{n+1}$  from the Mtd  $k(\cdot|x)$ ? The answer is based on a sequential simulation technique that we now describe. Suppose we would like to simulate a random vector from some pdf  $f_U(u)$ , but we cannot do this directly. Suppose further that  $f_U$  is the  $u$ -marginal of the joint pdf  $f_{U,V}(u, v)$  and that we have the ability to make draws from  $f_V(v)$  and from  $f_{U|V}(u|v)$  for fixed  $v$ . If we draw  $V \sim f_V(\cdot)$ , and then, conditional on  $V = v$ , we draw  $U \sim f_{U|V}(\cdot|v)$ , then the observed pair,  $(u, v)$ , is a draw from  $f_{U,V}$ , which means that  $u$  is a draw from  $f_U$ . This general technique will be employed many times throughout this chapter. We now explain how it is used to simulate from  $k(\cdot|x)$ .

Define

$$h(x', y|x) = f_{X|Y}(x'|y)f_{Y|X}(y|x) ,$$

and note that, for fixed  $x \in \mathbf{X}$ ,  $h(x', y|x)$  is a joint pdf in  $(x', y)$  with  $\int_{\mathbf{Y}} h(x', y|x) dy = k(x'|x)$ . We simply apply the technique described above with  $k(\cdot|x)$  and  $h(\cdot, \cdot|x)$  playing the roles of  $f_U(\cdot)$  and  $f_{U,V}(\cdot, \cdot)$ , respectively. All we need is the  $y$ -marginal of  $h(x', y|x)$ , which is  $f_{Y|X}(y|x)$ , and the conditional density of  $X'$  given  $Y = y$ , which is

$$\frac{h(x', y|x)}{f_{Y|X}(y|x)} = f_{X|Y}(x'|y) .$$

We now have a procedure for simulating one step of the DA algorithm. Indeed, if the current state is  $X_n = x$ , we simulate  $X_{n+1}$  as follows.

---

One iteration of the DA Algorithm:

1. Draw  $Y \sim f_{Y|X}(\cdot|x)$ , and call the observed value  $y$ .
  2. Draw  $X_{n+1} \sim f_{X|Y}(\cdot|y)$ .
- 

So, as long as we can simulate from the conditional densities,  $f_{X|Y}$  and  $f_{Y|X}$ , we can simulate

the Markov chain  $X$ . (Note that, as mentioned above, we do not need  $k(x'|x)$  in closed form.)

The genesis of the name *data augmentation algorithm* appears to be Tanner and Wong (1987) who used it to describe an iterative algorithm for approximating complex posterior distributions. On the last page of the paper, Tanner and Wong note that an “extreme” special case of their algorithm (in which their  $m$  is set equal to 1) yields a Markov chain whose transition density has the form (1.1.1). However, it does not appear to be the case that Tanner and Wong (1987) “invented” the DA algorithm (as we have defined it here), since other researchers, such as Swendsen and Wang (1987), were using it at about the same time. Here is our first example.

EXAMPLE 1. Suppose that  $f_X$  is the standard normal density, i.e.,  $f_X(x) = e^{-x^2/2}/\sqrt{2\pi}$ . Obviously, there is nothing intractable about this density. On the other hand, it is instructive to begin with a few simple examples in which the basic ideas of the algorithm are not overshadowed by the complexity of the target density. Take  $f(x, y) = (\sqrt{2\pi})^{-1} \exp\{-(x^2 - \sqrt{2}xy + y^2)\}$ , which is a bivariate normal density with means equal to zero, variances equal to one, and correlation equal to  $1/\sqrt{2}$ . The  $x$ -marginal is clearly standard normal and the two conditionals are also normal. Indeed,

$$Y|X = x \sim N\left(\frac{x}{\sqrt{2}}, \frac{1}{2}\right) \quad \text{and} \quad X|Y = y \sim N\left(\frac{y}{\sqrt{2}}, \frac{1}{2}\right)$$

Simulating from these conditionals is easy. For example, most statistically oriented programming languages, such as R (R Development Core Team, 2006), produce variates from the normal distribution and many other standard distributions. Hence, we have a viable DA algorithm that can be run by choosing an arbitrary starting value,  $X_0 = x_0$ , and then iterating the two-step procedure described above.

We now provide two more toy examples that will be put to good use. Two realistic examples are given later in this section.

EXAMPLE 2. Suppose that  $f_X(x) = 3x^2I_{(0,1)}(x)$ . If we take  $f(x, y) = 3xI(0 < y < x < 1)$ ,

then the  $x$ -marginal is  $f_X(x) = 3x^2 I_{(0,1)}(x)$  and the two conditional densities are given by

$$f_{Y|X}(y|x) = \frac{1}{x} I(0 < y < x) \quad \text{and} \quad f_{X|Y}(x|y) = \frac{2x}{1-y^2} I(y < x < 1).$$

Simulating from these conditionals is straightforward. Indeed, if  $U \sim \text{Uniform}(0,1)$ , then  $xU \sim f_{Y|X}(\cdot|x)$  and, using the probability integral transformation,  $\sqrt{U(1-y^2) + y^2} \sim f_{X|Y}(\cdot|y)$ .

EXAMPLE 3. Suppose that  $f_X(x)$  is a Student's  $t$  density with 4 degrees of freedom; that is,

$$f_X(x) = \frac{3}{8} \left(1 + \frac{x^2}{4}\right)^{-\frac{5}{2}}.$$

If we take

$$f(x, y) = \frac{4}{\sqrt{2\pi}} y^{\frac{3}{2}} \exp\left\{-y\left(\frac{x^2}{2} + 2\right)\right\} I_{(0,\infty)}(y),$$

then  $\int_{\mathbb{R}} f(x, y) dy = f_X(x)$ . Moreover, it's easy to show that  $X|Y = y \sim N(0, y^{-1})$  and that  $Y|X = x \sim \text{Gamma}\left(\frac{5}{2}, \frac{x^2}{2} + 2\right)$ . (We say  $W \sim \text{Gamma}(\alpha, \beta)$  if its density is proportional to  $w^{\alpha-1} e^{-w\beta} I(w > 0)$ .)

The popularity of the DA algorithm is due in part to the fact that, given an intractable  $f_X$ , there are general techniques available for constructing a potentially useful joint density  $f(x, y)$ . Here is one such technique. Suppose that  $f_X$  can be factorized as  $f_X(x) = q(x)l(x)$ . Now define

$$f(x, y) = q(x)I_{(0,l(x))}(y),$$

and note that

$$\int_{\mathbb{R}} f(x, y) dy = q(x) \int_{\mathbb{R}} I_{(0,l(x))}(y) dy = q(x) \int_0^{l(x)} dy = q(x)l(x) = f_X(x).$$

A simple calculation shows that  $Y|X = x \sim \text{Uniform}(0, l(x))$ , which is easy to sample. Thus, if it is also possible to draw from  $f_{X|Y}(x|y) \propto q(x)I_{(y,\infty)}(l(x))$ , then the DA algorithm can be applied. In this particular form, the DA algorithm is known as the *simple slice sampler* (Neal, 2003). The reader may verify that the DA algorithm developed in Example 2 is actually

a simple slice sampler based on the factorization  $f_X(x) = 3x^2I_{(0,1)}(x) = [3xI_{(0,1)}(x)][x] = [q(x)][l(x)]$ .

Another general technique for identifying an appropriate  $f(x, y)$  involves the concept of *missing data* that underlies the EM algorithm (Dempster et al., 1977). This technique is applicable when the target,  $f_X$ , is a posterior density. Let  $z$  denote some observed data, which is assumed to be a sample from a member of a family of pdfs  $\{p(z|\theta) : \theta \in \Theta\}$ , where  $\Theta \subset \mathbb{R}^p$ . If  $\pi(\theta)$  denotes the prior density, then the posterior density is given by  $\pi(\theta|z) = p(z|\theta)\pi(\theta)/c(z)$  where  $c(z) = \int_{\Theta} p(z|\theta)\pi(\theta) d\theta$  is the marginal density of the data. Assume that expectations with respect to  $\pi(\theta|z)$  are intractable; that is,  $\pi(\theta|z)$  is now playing the role of the problematic target  $f_X(x)$ .

Suppose we can identify missing data  $y \in \mathcal{Y} \subset \mathbb{R}^q$  such that the joint density of  $z$  and  $y$ , call it  $p(z, y|\theta)$ , satisfies

$$\int_{\mathcal{Y}} p(z, y|\theta) dy = p(z|\theta) . \quad (1.1.4)$$

Finding such missing data is often straightforward. Indeed, the joint density  $p(z, y|\theta)$  is precisely what is required to construct an EM algorithm for finding the maximum likelihood estimate of  $\theta$ ; that is, the maximizer of  $p(z|\theta)$  over  $\theta \in \Theta$  for fixed  $z$ . If such an EM algorithm already exists, we can simply use the corresponding missing data. Now define the *complete data posterior density* as

$$\pi(\theta, y|z) = \frac{p(z, y|\theta)\pi(\theta)}{\int_{\Theta} \int_{\mathcal{Y}} p(z, y|\theta)\pi(\theta) dy d\theta} = \frac{p(z, y|\theta)\pi(\theta)}{\int_{\Theta} p(z|\theta)\pi(\theta) d\theta} = \frac{p(z, y|\theta)\pi(\theta)}{c(z)} .$$

The key feature of the complete data posterior density is that its  $\theta$ -marginal is the target density,  $\pi(\theta|z)$ . Indeed,

$$\int_{\mathcal{Y}} \pi(\theta, y|z) dy = \frac{\pi(\theta)}{c(z)} \int_{\mathcal{Y}} p(z, y|\theta) dy = \frac{p(z|\theta)\pi(\theta)}{c(z)} = \pi(\theta|z) .$$

When an EM algorithm is constructed, the missing data is chosen to make likelihood calculations under  $p(z, y|\theta)$  much simpler than they are under the original density,  $p(z|\theta)$ . Such a choice will usually also result in conditional densities,  $\pi(\theta|y, z)$  and  $\pi(y|\theta, z)$ , that are easy to sample. Regardless of whether or not our missing data came from a preexisting EM al-

gorithm, as long as  $\pi(\theta|y, z)$  and  $\pi(y|\theta, z)$  can be straightforwardly sampled, we will have a viable DA algorithm with the complete data posterior density playing the role of  $f(x, y)$ . In particular,  $\theta$  plays the role of  $x$ , and everything is done conditionally on the observed data  $z$ .

EXAMPLE 4. Let  $Z_1, \dots, Z_m$  be a random sample from the location-scale Student's  $t$  density with known degrees of freedom,  $\nu > 0$ . The common density of the  $Z_i$ s is given by

$$\frac{\Gamma(\frac{\nu+1}{2})}{\sigma\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})} \left(1 + \frac{(z - \mu)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}.$$

Here  $(\mu, \sigma^2)$  is playing the role of  $\theta$ . The standard diffuse prior density for this location-scale problem is  $\pi(\mu, \sigma^2) \propto 1/\sigma^2$ . Of course, whenever an improper prior is used, it is important to check that the posterior is proper. In this case, the posterior is proper if and only if  $m \geq 2$  (Fernández and Steel, 1999) and we assume this throughout. The posterior density is an intractable bivariate density that is characterized by

$$\pi(\mu, \sigma^2|z) \propto (\sigma^2)^{-\frac{m+2}{2}} \prod_{i=1}^m \left(1 + \frac{(z_i - \mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}},$$

where  $z = (z_1, \dots, z_m)$ . Meng and van Dyk (1999) described a DA algorithm for this problem in which the missing data are based on the standard representation of a Student's  $t$  variate in terms of normal and  $\chi^2$  variates. Conditional on  $(\mu, \sigma^2)$ , let  $(Z_1, Y_1), \dots, (Z_m, Y_m)$  be independent and identically distributed (iid) pairs such that, for  $i = 1, \dots, m$ ,

$$\begin{aligned} Z_i|Y_i, \mu, \sigma^2 &\sim N(\mu, \sigma^2/Y_i) \\ Y_i|\mu, \sigma^2 &\sim \text{Gamma}(\nu/2, \nu/2). \end{aligned}$$

In this case,  $\mathbf{Y} = \mathbb{R}_+^m$  where  $\mathbb{R}_+ := (0, \infty)$ . Letting  $y = (y_1, \dots, y_m)$  we have

$$\begin{aligned} p(z, y|\mu, \sigma^2) &= \prod_{i=1}^m p(z_i|y_i, \mu, \sigma^2)p(y_i|\mu, \sigma^2) \\ &= \prod_{i=1}^m \frac{\sqrt{y_i}}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{y_i}{2\sigma^2}(z_i - \mu)^2\right\} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} y_i^{\frac{\nu}{2}-1} \exp\left\{-\frac{\nu y_i}{2}\right\}. \end{aligned}$$

Now,

$$\begin{aligned} \int_{\mathcal{Y}} p(z, y|\mu, \sigma^2) dy &= \prod_{i=1}^m \int_{\mathbb{R}_+} p(z_i|y_i, \mu, \sigma^2) p(y_i|\mu, \sigma^2) dy_i \\ &= \prod_{i=1}^m \frac{\Gamma(\frac{\nu+1}{2})}{\sigma\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})} \left(1 + \frac{(z_i - \mu)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}, \end{aligned}$$

so (1.1.4) is satisfied. The complete data posterior density is characterized by

$$\pi((\mu, \sigma^2), y|z) \propto \frac{1}{\sigma^2} \prod_{i=1}^m \frac{\sqrt{y_i}}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{y_i}{2\sigma^2}(z_i - \mu)^2\right\} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} y_i^{\frac{\nu}{2}-1} \exp\left\{-\frac{\nu y_i}{2}\right\}. \quad (1.1.5)$$

In order to implement the DA algorithm, we must be able to draw from  $\pi(y|\mu, \sigma^2, z)$  and from  $\pi(\mu, \sigma^2|y, z)$ . Since  $\pi(y|\mu, \sigma^2, z) \propto \pi(\mu, \sigma^2, y|z)$ , it is clear that the  $y_i$ s are conditionally independent given  $(\mu, \sigma^2, z)$  and, in fact,

$$Y_i|\mu, \sigma^2, z \sim \text{Gamma}\left(\frac{\nu+1}{2}, \frac{1}{2}\left(\frac{(z_i - \mu)^2}{\sigma^2} + \nu\right)\right). \quad (1.1.6)$$

We can simulate from  $\pi(\mu, \sigma^2|y, z)$  sequentially by first drawing from  $\pi(\sigma^2|y, z)$  and then from  $\pi(\mu|\sigma^2, y, z)$ . (Remember our sequential method of drawing from  $f_{U,V}$ ?) Let  $y_\cdot = \sum_{i=1}^m y_i$  and define

$$\hat{\mu} = \frac{1}{y_\cdot} \sum_{j=1}^m z_j y_j \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{y_\cdot} \sum_{j=1}^m y_j (z_j - \hat{\mu})^2.$$

Using the fact that  $\pi(\mu|\sigma^2, y, z) \propto \pi(\mu, \sigma^2, y|z)$ , it is straightforward to show that

$$\mu|\sigma^2, y, z \sim \text{N}\left(\hat{\mu}, \frac{\sigma^2}{y_\cdot}\right). \quad (1.1.7)$$

Finally,  $\pi(\sigma^2|y, z)$  is proportional to what remains when  $\mu$  is integrated out of (1.1.5). This integral can be computed in closed form and it follows that

$$\sigma^2|y, z \sim \text{IG}\left(\frac{m+1}{2}, \frac{y_\cdot \hat{\sigma}^2}{2}\right), \quad (1.1.8)$$

where  $\text{IG}(\alpha, \beta)$  is the distribution of  $1/W$  when  $W \sim \text{Gamma}(\alpha, \beta)$ . We now know how to run the DA algorithm for this problem. Given the current state,  $X_n = (\mu, \sigma^2)$ , we simulate

the next state,  $X_{n+1} = (\mu_{n+1}, \sigma_{n+1}^2)$ , by performing the following two steps:

1. Draw  $Y_1, \dots, Y_m$  independently according to (1.1.6), and call the result  $y = (y_1, \dots, y_m)$ .
2. Draw  $\sigma_{n+1}^2$  according to (1.1.8), and then draw  $\mu_{n+1}$  according to (1.1.7) with  $\sigma_{n+1}^2$  in place of  $\sigma^2$ .

The algorithm described above is actually a special case of a more general DA algorithm developed by Meng and van Dyk (1999) that can handle observations from the *multivariate* location-scale Student's  $t$  density.

We end this section by describing Albert and Chib's (1993) DA algorithm for Bayesian probit regression, which is one of the most widely used DA algorithms.

EXAMPLE 5. Let  $Z_1, \dots, Z_m$  be independent Bernoulli random variables such that  $\Pr(Z_i = 1) = \Phi(v_i^T \beta)$  where  $v_i$  is a  $p \times 1$  vector of known covariates associated with  $Z_i$ ,  $\beta$  is a  $p \times 1$  vector of unknown regression coefficients and  $\Phi(\cdot)$  denotes the standard normal distribution function. We have

$$\Pr(Z_1 = z_1, \dots, Z_m = z_m \mid \beta) = \prod_{i=1}^m [\Phi(v_i^T \beta)]^{z_i} [1 - \Phi(v_i^T \beta)]^{1-z_i},$$

where each  $z_i$  is binary; i.e., either 0 or 1. Consider a Bayesian analysis that employs a flat prior on  $\beta$ . Letting  $z = (z_1, \dots, z_m)$  denote the observed data, the marginal density is given by

$$c(z) = \int_{\mathbb{R}^p} \prod_{i=1}^m [\Phi(v_i^T \beta)]^{z_i} [1 - \Phi(v_i^T \beta)]^{1-z_i} d\beta.$$

Chen and Shao (2000) provide necessary and sufficient conditions on  $z$  and  $\{v_i\}_{i=1}^m$  for propriety of the posterior; that is, for  $c(z) < \infty$ . We assume throughout that these conditions are satisfied. The intractable posterior density of  $\beta$  is given by

$$\pi(\beta \mid z) = \frac{1}{c(z)} \prod_{i=1}^m [\Phi(v_i^T \beta)]^{z_i} [1 - \Phi(v_i^T \beta)]^{1-z_i}.$$

We now describe a DA algorithm for this problem that was developed by Albert and Chib (1993). Let  $\phi(u; \mu, \kappa^2)$  denote the  $N(\mu, \kappa^2)$  density function evaluated at the point  $u \in \mathbb{R}$ . Also, let  $\mathbb{R}_- = (-\infty, 0)$ , let  $y = (y_1, \dots, y_m)^T \in \mathbb{R}^m$  and consider the function

$$\pi(\beta, y | z) = \frac{1}{c(z)} \prod_{i=1}^m \left\{ I_{\mathbb{R}_+}(y_i) I_{\{1\}}(z_i) + I_{\mathbb{R}_-}(y_i) I_{\{0\}}(z_i) \right\} \phi(y_i; v_i^T \beta, 1), \quad (1.1.9)$$

where  $I_A(\cdot)$  is the usual indicator function of the set  $A$ . Integrating  $y$  out of  $\pi(\beta, y | z)$ , we have

$$\begin{aligned} & \frac{1}{c(z)} \int_{\mathbb{R}} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \prod_{i=1}^m \left\{ I_{\mathbb{R}_+}(y_i) I_{\{1\}}(z_i) + I_{\mathbb{R}_-}(y_i) I_{\{0\}}(z_i) \right\} \phi(y_i; v_i^T \beta, 1) dy_m \cdots dy_2 dy_1 \\ &= \frac{1}{c(z)} \prod_{i=1}^m \int_{\mathbb{R}} \left\{ I_{\mathbb{R}_+}(y_i) I_{\{1\}}(z_i) + I_{\mathbb{R}_-}(y_i) I_{\{0\}}(z_i) \right\} \phi(y_i; v_i^T \beta, 1) dy_i \\ &= \frac{1}{c(z)} \prod_{i=1}^m \left\{ I_{\{1\}}(z_i) \int_0^\infty \phi(y_i; v_i^T \beta, 1) dy_i + I_{\{0\}}(z_i) \int_{-\infty}^0 \phi(y_i; v_i^T \beta, 1) dy_i \right\} \\ &= \frac{1}{c(z)} \prod_{i=1}^m \left\{ I_{\{1\}}(z_i) \Phi(v_i^T \beta) + I_{\{0\}}(z_i) [1 - \Phi(v_i^T \beta)] \right\} \\ &= \frac{1}{c(z)} \prod_{i=1}^m [\Phi(v_i^T \beta)]^{z_i} [1 - \Phi(v_i^T \beta)]^{1-z_i} \\ &= \pi(\beta | z). \end{aligned}$$

Hence,  $\pi(\beta, y | z)$  is a joint density in  $(\beta, y)$  whose  $\beta$ -marginal is  $\pi(\beta | z)$ . Albert and Chib's (1993) DA algorithm is based on this joint density. We now derive the conditional densities,  $\pi(\beta | y, z)$  and  $\pi(y | \beta, z)$ . Let  $V$  denote the  $m \times p$  matrix whose  $i$ th row is  $v_i^T$ . (A necessary condition for propriety is that  $V$  have rank  $p$ .) Standard linear models-type calculations show that

$$\prod_{i=1}^m \phi(y_i; v_i^T \beta, 1) = (2\pi)^{-\frac{m}{2}} e^{-\frac{y^T(I-H)y}{2}} \exp \left\{ -\frac{1}{2} (\beta - \hat{\beta}(y))^T V^T V (\beta - \hat{\beta}(y)) \right\}, \quad (1.1.10)$$

where  $\hat{\beta}(y) = (V^T V)^{-1} V^T y$  and  $H = V(V^T V)^{-1} V^T$ . This implies that  $\pi(\beta | y, z)$  is a  $p$ -variate normal density with mean  $\hat{\beta}(y)$  and covariance matrix  $(V^T V)^{-1}$ .

Finally, let  $TN(\mu, \kappa^2, u)$  denote a normal distribution with mean  $\mu$  and variance  $\kappa^2$  that

is *truncated* to be positive if  $u = 1$  and negative if  $u = 0$ . It is clear from (1.1.9) that, given  $\beta$  and  $z$ , the  $Y_i$ s are independent with  $Y_i \sim \text{TN}(v_i^T \beta, 1, z_i)$ . We now know exactly how to implement the DA algorithm. Given the current state,  $X_n = \beta$ , we simulate the next state,  $X_{n+1}$ , by performing the following two steps:

1. Draw  $Y_1, \dots, Y_m$  independently such that  $Y_i \sim \text{TN}(v_i^T \beta, 1, z_i)$ , and call the result  $y = (y_1, \dots, y_m)^T$ .
2. Draw  $X_{n+1} \sim \text{N}(\hat{\beta}(y), (V^T V)^{-1})$ .

See Robert (1995) for an efficient method of simulating truncated normal random variables.

In the next section, we describe the theoretical properties of the Markov chain underlying the DA algorithm.

## 1.2 Properties of the DA Markov chain

### 1.2.1 Basic regularity conditions

In Section 1.1 we described how to construct and simulate a Markov chain,  $X$ , that has the intractable target,  $f_X$ , as an invariant density. Unfortunately, without additional assumptions, there is no guarantee that this chain will be useful for approximating expectations with respect to  $f_X$ . Here is a simple example from Roberts and Smith (1994) that illustrates one of the potential problems.

EXAMPLE 6. Suppose that  $f_X(x) = \frac{1}{2}I_{(0,2)}(x)$ . If we take

$$f(x, y) = \frac{1}{2} \left[ I_{(0,1)}(x)I_{(0,1)}(y) + I_{[1,2)}(x)I_{[1,2)}(y) \right],$$

then  $\int_{\mathbb{R}} f(x, y) dy = \frac{1}{2}I_{(0,2)}(x)$ , and

$$f_{X|Y}(x|y) = f_{Y|X}(y|x) = I_{(0,1)}(x)I_{(0,1)}(y) + I_{[1,2)}(x)I_{[1,2)}(y) .$$

Since the  $x$ -marginal of  $f(x, y)$  is  $f_X$  and simulation from the conditionals is easy, there is a DA algorithm based on  $f(x, y)$ . However, this algorithm is useless from a practical standpoint because the underlying Markov chain is not *irreducible*. For example, suppose we start the chain at  $x_0 = 1/2$ , and consider applying the two-step method to simulate  $X_1$ . First, we draw  $Y \sim \text{Uniform}(0, 1)$  and then, no matter what the result, we will draw  $X_1 \sim \text{Uniform}(0, 1)$ . Continuing along these lines shows that the chain will be stuck forever in the set  $(0, 1)$ . Hence, there is no sense in which the chain converges to  $f_X$ .

If the Markov chain  $X$  is  $\psi$ -irreducible, aperiodic and Harris recurrent, then the DA algorithm can be employed to effectively explore the intractable target density,  $f_X$ . When  $X$  satisfies these three properties, we call it *Harris ergodic*. Unfortunately, a good bit of technical Markov chain theory must be developed before these conditions can even be formally stated (Meyn and Tweedie, 1993; Roberts and Rosenthal, 2004). To avoid a lengthy technical discussion, we simply provide one sufficient condition for Harris ergodicity of  $X$  that is easy to check and holds for all of our examples and for many other DA algorithms that are used in practice.

Define a condition on the Mtd  $k$  as follows:

$$\text{Condition } \mathcal{K} : k(x'|x) > 0 \text{ for all } x', x \in \mathbf{X} .$$

Condition  $\mathcal{K}$  implies that the Markov chain  $X$  is Harris ergodic (see, e.g., Tan, 2008). In fact, condition  $\mathcal{K}$  implies that it is possible for the chain to move from any point  $x \in \mathbf{X}$  to any “big” set in a single step. To make this precise, let  $\lambda$  denote Lebesgue measure on  $\mathbf{X}$  and let  $P(\cdot, \cdot)$  denote the Markov transition function of the chain; that is, for  $x \in \mathbf{X}$  and a measurable set  $A$ ,

$$P(x, A) = \Pr(X_{n+1} \in A | X_n = x) = \int_A k(x'|x) dx' .$$

Under condition  $\mathcal{K}$ , if  $A$  is big in the sense that  $\lambda(A) > 0$ , then

$$P(x, A) = \int_A k(x'|x) dx' > 0,$$

which means that there is positive probability of moving from  $x$  to  $A$  in a single step. Recall that

$$k(x'|x) = \int_{\mathbf{Y}} f_{X|Y}(x'|y) f_{Y|X}(y|x) dy.$$

Clearly, if  $f(x, y)$  is strictly positive on  $\mathbf{X} \times \mathbf{Y}$ , then condition  $\mathcal{K}$  holds and the Markov chain  $X$  is Harris ergodic. We now check that the Markov chains developed in the examples of Section 1.1 are indeed Harris ergodic.

EXAMPLES 1 AND 3 CONT. In Example 1, we have  $\mathbf{X} = \mathbf{Y} = \mathbb{R}$ , while in Example 3,  $\mathbf{X} = \mathbb{R}$ ,  $\mathbf{Y} = \mathbb{R}_+$ . In both cases,  $f(x, y)$  is strictly positive on  $\mathbf{X} \times \mathbf{Y}$ . Hence, the Markov chains underlying the DA algorithms in Examples 1 and 3 are Harris ergodic.

EXAMPLE 4 CONT. The role of  $\mathbf{X}$  is played by  $\Theta = \mathbb{R} \times \mathbb{R}_+$  and  $\mathbf{Y} = \mathbb{R}_+^m$ . Note that the complete data posterior density (1.1.5) is strictly positive for all  $((\mu, \sigma^2), y) \in \Theta \times \mathbf{Y}$ . Hence, the chain  $X$  is Harris ergodic.

EXAMPLE 5 CONT. In this case,  $\mathbf{X} = \mathbb{R}^p$  and  $\mathbf{Y}$  is a Cartesian product of  $m$  half-lines ( $\mathbb{R}_+$  and  $\mathbb{R}_-$ ), where the  $i$ th component is  $\mathbb{R}_+$  if  $z_i = 1$ , and  $\mathbb{R}_-$  if  $z_i = 0$ . It is clear that the joint density (1.1.9) is strictly positive on  $\mathbf{X} \times \mathbf{Y}$ , and this implies that the Markov chain underlying Albert and Chib's (1993) algorithm is Harris ergodic.

Even when  $f(x, y)$  is not strictly positive on  $\mathbf{X} \times \mathbf{Y}$ , it is still often the case that condition  $\mathcal{K}$  holds.

EXAMPLE 2 CONT. The joint density is given by  $f(x, y) = 3xI(0 < y < x < 1)$ , which is not strictly positive on  $\mathbf{X} \times \mathbf{Y} = (0, 1) \times (0, 1)$ . However, we can show directly that condition  $\mathcal{K}$

holds. Indeed, for fixed  $x \in (0, 1)$  we have

$$\begin{aligned} k(x'|x) &= \int_{\mathbb{R}} \frac{2x'}{x(1-y^2)} I(0 < y < x) I(y < x' < 1) dy \\ &= \frac{2x'}{x} I(0 < x' < 1) \int_0^{\min\{x, x'\}} \frac{1}{1-y^2} dy \\ &= \frac{x'}{x} \log \left( \frac{1 + \min\{x, x'\}}{1 - \min\{x, x'\}} \right) I(0 < x' < 1). \end{aligned}$$

Hence,  $k(x'|x)$  is strictly positive for all  $x', x \in (0, 1)$  and Harris ergodicity follows. Actually, it is intuitively clear that the Markov chain has a positive probability of moving from any  $x \in (0, 1)$  to any set  $A \subset (0, 1)$  with  $\lambda(A) > 0$  in one step. Indeed, to get from  $x$  to the new state, we first draw  $Y \sim \text{Uniform}(0, x)$ , and then given  $Y = y$ , the new state is drawn from a density with support  $(y, 1)$ . Therefore, as long as the observed  $y$  is small enough, there will be a positive (conditional) probability of the new state being in any open set in  $(0, 1)$ .

It is not difficult to create examples of well-behaved DA algorithms for which condition  $\mathcal{K}$  fails to hold. Fortunately, there are many general results available for establishing that  $X$  is Harris ergodic in such situations. See, for example, Roberts and Smith (1994), Tierney (1994), Roberts and Rosenthal (2006) and Hobert et al. (2007). In the next subsection, we describe exactly what Harris ergodicity buys us.

### 1.2.2 Basic convergence properties

If  $X$  is Harris ergodic, then, no matter how the chain is started, the marginal distribution of  $X_n$  will converge to (the distribution associated with)  $f_X$ , and an analogue of the strong law of large numbers (SLLN) holds. To make this precise, some additional notation is required. Define the  $n$ -step Markov transition function as

$$P^n(x, A) = \Pr(X_n \in A | X_0 = x),$$

so  $P^1 \equiv P$ . Also, let  $\phi(\cdot)$  denote the probability measure corresponding to  $f_X$ ; that is, for measurable  $A$ ,  $\phi(A) = \int_A f_X(x) dx$ . If  $X$  is Harris ergodic, then the total variation distance

between the probability measures  $P^n(x, \cdot)$  and  $\phi(\cdot)$  decreases to 0 as  $n$  gets large. In symbols,

$$\|P^n(x, \cdot) - \phi(\cdot)\| \downarrow 0 \quad \text{as } n \rightarrow \infty, \quad (1.2.1)$$

where

$$\|P^n(x, \cdot) - \phi(\cdot)\| := \sup_A |P^n(x, A) - \phi(A)|.$$

Harris ergodicity is also sufficient for the ergodic theorem, which is the Markov chain version of the SLLN. Let  $L^1(f_X)$  denote the set of functions  $h : \mathbf{X} \rightarrow \mathbb{R}$  such that

$$\int_{\mathbf{X}} |h(x)| f_X(x) dx < \infty,$$

and, for  $h \in L^1(f_X)$ , define  $E_{f_X} h = \int_{\mathbf{X}} h(x) f_X(x) dx$ . The ergodic theorem implies that, if  $g \in L^1(f_X)$ , then, no matter what the distribution of  $X_0$ , we have

$$\bar{g}_n := \frac{1}{n} \sum_{i=0}^{n-1} g(X_i) \rightarrow E_{f_X} g$$

almost surely as  $n \rightarrow \infty$ ; i.e.,  $\bar{g}_n$  is a strongly consistent estimator of  $E_{f_X} g$ . The ergodic theorem justifies estimating  $E_{f_X} g$  with  $\bar{g}_n$  where  $X_0$  is any point (or has any distribution) from which it is convenient to start the simulation. An important practical question that this basic theory does not answer is “What is an appropriate value of  $n$ ?” Tools for answering this question will be presented in Section 1.3. For now, we simply point out that all rigorous methods of choosing an appropriate (Markov chain) Monte Carlo sample size are based on a central limit theorem (CLT) for  $\bar{g}_n$ . Assuming that  $\int_{\mathbf{X}} g^2(x) f_X(x) dx < \infty$ , a simple sufficient condition for the existence of such a CLT is that the Markov chain,  $X$ , converge to its stationary distribution at a geometric rate.

### 1.2.3 Geometric ergodicity

Assume that  $X$  is Harris ergodic. Note that (1.2.1) gives no information about the *rate* at which the total variation distance converges to 0. There are important practical benefits to using a DA algorithm for which this rate is (at least) geometrically fast. Formally, the chain

$X$  is called *geometrically ergodic* if there exist a function  $M : \mathsf{X} \rightarrow [0, \infty)$  and a constant  $\rho \in [0, 1)$  such that, for all  $x \in \mathsf{X}$  and all  $n = 1, 2, \dots$ ,

$$\|P^n(x, \cdot) - \phi(\cdot)\| \leq M(x) \rho^n . \quad (1.2.2)$$

Unfortunately, Harris ergodicity does not imply geometric ergodicity. The most straightforward method of proving that the Harris ergodic chain  $X$  is geometrically Harris ergodic is by establishing a certain type of *drift condition*, which we now introduce.

A function  $V : \mathsf{X} \rightarrow [0, \infty)$  is said to be *unbounded off compact sets* if, for each  $\beta \in \mathbb{R}$ , the sub-level set  $\{x \in \mathsf{X} : V(x) \leq \beta\}$  is compact. We say that a *geometric drift condition* holds if there exist a  $V : \mathsf{X} \rightarrow [0, \infty)$  that is unbounded off compact sets, and constants  $\lambda \in [0, 1)$  and  $L \in \mathbb{R}$  such that

$$\mathbb{E}[V(X_{n+1}) \mid X_n = x] \leq \lambda V(x) + L . \quad (1.2.3)$$

The function  $V$  is called the *drift function*. If  $f(x, y) > 0$  for all  $(x, y) \in \mathsf{X} \times \mathsf{Y}$ , then the existence of a geometric drift condition implies that  $X$  is geometrically ergodic (Tan, 2008). (See Meyn and Tweedie (1993, Chapter 15) for similar results that hold when  $f(x, y)$  is not strictly positive.) In practice, establishing a geometric drift condition is simply a matter of trial and error (and a lot of analysis). We now provide some pointers on calculating the expectation in (1.2.3).

Note that the left-hand side of (1.2.3) can be rewritten as

$$\begin{aligned} \mathbb{E}[V(X_{n+1}) \mid X_n = x] &= \int_{\mathsf{X}} V(x') k(x'|x) dx' \\ &= \int_{\mathsf{X}} V(x') \left[ \int_{\mathsf{Y}} f_{X|Y}(x'|y) f_{Y|X}(y|x) dy \right] dx' \\ &= \int_{\mathsf{Y}} \left[ \int_{\mathsf{X}} V(x') f_{X|Y}(x'|y) dx' \right] f_{Y|X}(y|x) dy . \end{aligned} \quad (1.2.4)$$

Thus, the expectation can be computed (or bounded) in two steps. The first step is to compute (or bound) the expectation of  $V(X')$  with respect to  $f_{X|Y}(\cdot|y)$ , call the result  $e(y)$ . The second step entails calculating (or bounding) the expectation of  $e(Y)$  with respect to  $f_{Y|X}(\cdot|x)$ . The fact that we are able to simulate straightforwardly from  $f_{X|Y}(x|y)$  and

$f_{Y|X}(y|x)$  often means that these conditional densities are easy to handle from an analytical standpoint. Hence, it is usually possible to calculate (or, at least get sharp upper bounds on) expectations with respect to  $f_{X|Y}(x|y)$  and  $f_{Y|X}(y|x)$ . We now give two simple examples illustrating how to prove that a DA algorithm is geometrically ergodic by establishing a geometric drift condition.

EXAMPLE 3 CONT. Recall that  $f(x, y)$  is strictly positive on  $\mathbf{X} \times \mathbf{Y}$ , so the drift technique can be used to establish geometric convergence in this example. Consider the drift function  $V(x) = x^2$ . For  $\beta < 0$ , the sub-level set  $\{x \in \mathbf{X} : V(x) \leq \beta\}$  is the empty set, for  $\beta = 0$ , it's the set  $\{0\}$ , and for  $\beta > 0$ , it's a closed interval. Thus,  $V$  is unbounded off compact sets. Recall that  $X|Y = y \sim \mathbf{N}(0, y^{-1})$ . Hence, the “inner expectation” in (1.2.4) can be evaluated as follows

$$\mathbf{E}[V(X') | y] = \mathbf{E}[(X')^2 | y] = \frac{1}{y}.$$

Now, using the fact that  $Y|X = x \sim \text{Gamma}(\frac{5}{2}, \frac{x^2}{2} + 2)$  yields

$$\mathbf{E}[V(X_{n+1}) | X_n = x] = \mathbf{E}[Y^{-1} | x] = \frac{1}{3}x^2 + \frac{4}{3} = \frac{1}{3}V(x) + \frac{4}{3}.$$

We have established that (1.2.3) holds with  $\lambda = \frac{1}{3}$  and  $L = \frac{4}{3}$ , and this shows that the Markov chain underlying this DA algorithm is geometrically ergodic.

In the toy example just considered, we were able to compute  $\mathbf{E}[V(X_{n+1}) | X_n = x]$  exactly and, luckily, the final expression involved the function  $V(x)$  in exactly the right way. Establishing geometric drift conditions in real examples is typically much more difficult, and often involves what Fill et al. (2000) describe as “difficult theoretical analysis.” Geometric drift conditions have been established for the Markov chains underlying the DA algorithms in Examples 4 and 5 (Marchev and Hobert, 2004; Roy and Hobert, 2007), but these calculations are too involved to present in this chapter. The next example is still a toy example, in the sense that the intractable target density is univariate, but it does provide a nice illustration of the type of bounding that is required in real examples.

EXAMPLE 7. Consider a simplification of the Student's  $t$  setup in Example 4 where the

variance is known and equal to 1. In this case, the posterior density is an intractable univariate density given by

$$\pi(\mu|z) \propto \prod_{i=1}^m \left( 1 + \frac{(z_i - \mu)^2}{\nu} \right)^{-\frac{\nu+1}{2}},$$

where  $z = (z_1, \dots, z_m)$ . Using the same missing data as before, the complete data posterior,  $\pi(\mu, y|z)$ , is proportional to the right-hand side of (1.1.5) with  $\sigma^2$  set equal to 1. Note that  $\pi(\mu, y|z)$  is strictly positive on  $\mathbf{X} \times \mathbf{Y} = \Theta \times \mathbf{Y} = \mathbb{R} \times \mathbb{R}_+^m$ . Of course, to run the DA algorithm, we need to be able to draw from  $\pi(y|\mu, z)$  and from  $\pi(\mu|y, z)$ . Recall that  $\hat{\mu} = \hat{\mu}(y) = \frac{1}{y} \sum_{j=1}^m z_j y_j$ . (Since the data,  $z$ , is fixed, we suppress this dependence in the notation.) It's easy to show that  $\mu|y, z \sim \text{N}(\hat{\mu}(y), \frac{1}{y})$  and that the  $y_i$ s are conditionally independent given  $(\mu, z)$  with

$$Y_i|\mu, z \sim \text{Gamma}\left(\frac{\nu+1}{2}, \frac{(z_i - \mu)^2 + \nu}{2}\right).$$

For notational convenience, we will denote the DA Markov chain as  $\{\mu_n\}_{n=0}^\infty$  (instead of the usual  $\{X_n\}_{n=0}^\infty$ ). The Mtd of the DA algorithm is then given by

$$k(\mu'|\mu) = \int_{\mathbf{Y}} \pi(\mu'|y, z) \pi(y|\mu, z) dy.$$

We now show that this Markov chain is geometrically ergodic as long as  $\nu > 1$  and  $m > 1/(\nu - 1)$ . The drift function we use is  $V(\mu) = \sum_{i=1}^m (z_i - \mu)^2$ . It's easy to see that  $V$  is unbounded off compact sets. Indeed, fix  $\beta \in \mathbb{R}$  and consider the sub-level set  $\{\mu \in \mathbb{R} : V(\mu) \leq \beta\}$ . Let  $\bar{z} = m^{-1} \sum_{i=1}^m z_i$ . If  $\beta < \sum_{i=1}^m (z_i - \bar{z})^2$ , then the sub-level set is the empty set, and if  $\beta \geq \sum_{i=1}^m (z_i - \bar{z})^2$ , the sub-level set is a closed interval.

Let  $z_*$  and  $z^*$  denote the minimum and the maximum of the  $z_i$ s, respectively. Since  $\hat{\mu}(y)$  is a convex combination of  $z_1, \dots, z_m$ , it follows that  $\hat{\mu}(y) \in [z_*, z^*]$  for all  $y \in \mathbf{Y}$ . The inner

expectation in (1.2.4) can now be bounded as follows

$$\begin{aligned}
\mathbb{E}[V(\mu') \mid y, z] &= \mathbb{E}\left[\sum_{i=1}^m (z_i - \mu')^2 \mid y, z\right] \\
&= \sum_{i=1}^m \mathbb{E}\left[(z_i - \mu')^2 \mid y, z\right] \\
&= \sum_{i=1}^m \text{Var}[(z_i - \mu') \mid y, z] + \sum_{i=1}^m \left\{\mathbb{E}[(z_i - \mu') \mid y, z]\right\}^2 \\
&= \sum_{i=1}^m \text{Var}[\mu' \mid y, z] + \sum_{i=1}^m \left\{z_i - \mathbb{E}[\mu' \mid y, z]\right\}^2 \\
&= \frac{m}{y} + \sum_{i=1}^m (z_i - \hat{\mu}(y))^2 \\
&\leq \frac{m}{y} + m(z^* - z_*)^2.
\end{aligned}$$

Now, since the harmonic mean is less than or equal to the arithmetic mean, we have

$$\frac{m}{y} = \frac{1}{\frac{1}{m} \sum_{i=1}^m \frac{1}{y_i}} \leq \frac{1}{m} \sum_{i=1}^m y_i^{-1}.$$

We conclude that

$$\mathbb{E}[V(\mu') \mid y, z] \leq \frac{1}{m} \sum_{i=1}^m y_i^{-1} + m(z^* - z_*)^2.$$

Therefore, as long as  $\nu > 1$ , we have

$$\begin{aligned}
\mathbb{E}[V(\mu_{n+1}) \mid \mu_n = \mu] &\leq \mathbb{E}\left[\left(\frac{1}{m} \sum_{i=1}^m Y_i^{-1} + m(z^* - z_*)^2\right) \mid \mu, z\right] \\
&= \frac{1}{m} \sum_{i=1}^m \mathbb{E}[Y_i^{-1} \mid \mu, z] + m(z^* - z_*)^2 \\
&= \frac{1}{m(\nu - 1)} \sum_{i=1}^m [(z_i - \mu)^2 + \nu] + m(z^* - z_*)^2 \\
&= \frac{1}{m(\nu - 1)} \sum_{i=1}^m (z_i - \mu)^2 + \frac{\nu}{(\nu - 1)} + m(z^* - z_*)^2 \\
&= \frac{1}{m(\nu - 1)} V(\mu) + \frac{\nu}{(\nu - 1)} + m(z^* - z_*)^2.
\end{aligned}$$

We have established that, when  $\nu > 1$ , (1.2.3) holds with  $\lambda = \frac{1}{m(\nu-1)}$ . Thus, the Markov chain is geometrically ergodic whenever  $\nu > 1$  and  $m(\nu - 1) > 1$ .

Of course, the fact that our analysis did not lead to a geometric drift condition for the (extreme) situations where  $\nu \leq 1$  and/or  $m(\nu - 1) \leq 1$  *does not imply* that the DA chain converges at a sub-geometric rate in those cases. Indeed, it may be the case that a more delicate analysis of  $E[V(\mu_{n+1}) \mid \mu_n = \mu]$  would show that these chains are geometric as well. Or, we might have to resort to changing the drift function. Unfortunately, there are currently no simple methods of proving that a DA chain is *not* geometrically ergodic.

The drift method that we have described and illustrated in this subsection provides only *qualitative* information about the rate of convergence in the sense that, once (1.2.3) has been established, all we can say is that *there exist*  $M$  and  $\rho$  satisfying (1.2.2), but we cannot say what they are. There are other (more complicated) versions of this method that, in addition to establishing the existence of  $M$  and  $\rho$ , provide an upper bound on  $M(x)\rho^n$  that decreases to zero geometrically in  $n$ . These methods were developed and refined in a series of papers beginning with Meyn and Tweedie (1994) and Rosenthal (1995). For an overview, see Jones and Hobert (2001). The final subsection of this chapter concerns CLTs for the estimator  $\bar{g}_n$ .

#### 1.2.4 Central limit theorems

Harris ergodicity alone does not imply the existence of CLTs. However, as we now explain, if the DA Markov chain,  $X$ , is geometrically Harris ergodic, then there will be CLTs for square integrable functions. Let  $L^2(f_X)$  denote the set of functions  $h : \mathbf{X} \rightarrow \mathbb{R}$  such that

$$\int_{\mathbf{X}} h^2(x) f_X(x) dx < \infty .$$

Assume that  $g \in L^2(f_X)$  and define  $c_k = \text{Cov}[g(X_0), g(X_k)]$  for  $k \in \{1, 2, 3, \dots\}$ , where the covariances are calculated under the assumption that  $X_0 \sim f_X$ . For example,

$$c_1 = \int_{\mathbf{X}} \int_{\mathbf{X}} (g(x') - E_{f_X} g)(g(x) - E_{f_X} g) k(x'|x) f_X(x) dx dx' ,$$

where we have used the fact that  $X_0 \sim f_X$  implies that  $X_1 \sim f_X$ , so the expected value of  $g(X_1)$  is  $E_{f_X}g$ . Liu et al. (1994, Lemma 3.2) noted that this expression can be rearranged as follows

$$\begin{aligned}
c_1 &= \int_{\mathcal{X}} \int_{\mathcal{X}} (g(x') - E_{f_X}g)(g(x) - E_{f_X}g)k(x'|x)f_X(x) dx dx' \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} (g(x') - E_{f_X}g)(g(x) - E_{f_X}g) \left[ \int_{\mathcal{Y}} f_{X|Y}(x'|y)f_{Y|X}(y|x) dy \right] f_X(x) dx dx' \\
&= \int_{\mathcal{Y}} \int_{\mathcal{X}} \int_{\mathcal{X}} (g(x') - E_{f_X}g)(g(x) - E_{f_X}g)f_{X|Y}(x'|y)f_{X|Y}(x|y)f_Y(y) dx dx' dy \quad (1.2.5) \\
&= \int_{\mathcal{Y}} \left[ \int_{\mathcal{X}} (g(x) - E_{f_X}g)f_{X|Y}(x|y) dx \right]^2 f_Y(y) dy \\
&= \text{Var} \left\{ E[(g(X') - E_{f_X}g) | Y'] \right\},
\end{aligned}$$

where  $(X', Y') \sim f(x, y)$ . This shows that  $c_1 > 0$ . In fact, this result can be used in conjunction with the reversibility of  $X$  to show that  $c_k > 0$  for all  $k \in \{1, 2, 3, \dots\}$ .

Assume that  $X$  is geometrically Harris ergodic and that  $g \in L^2(f_X)$ . As before, put  $\bar{g}_n = \frac{1}{n} \sum_{i=0}^{n-1} g(X_i)$ . Define  $\sigma^2 = E_{f_X}g^2 - (E_{f_X}g)^2$  and  $\kappa^2 = \sigma^2 + 2 \sum_{k=1}^{\infty} c_k$ . Results in Roberts and Rosenthal (1997) and Chan and Geyer (1994) imply that  $\kappa^2 < \infty$  and that, as  $n \rightarrow \infty$ ,

$$\sqrt{n}(\bar{g}_n - E_{f_X}g) \xrightarrow{d} N(0, \kappa^2). \quad (1.2.6)$$

This CLT *does not require* that  $X_0 \sim f_X$  - it holds for all starting distributions, including degenerate ones. We note that the reversibility of  $X$  plays a major role in the existence of the CLT (1.2.6). In the next section, we explain how to consistently estimate the asymptotic variance,  $\kappa^2$ . But first, we briefly compare the estimators of  $E_{f_X}g$  based on DA and classical Monte Carlo.

Let  $X_1^*, X_2^*, \dots$  be an iid sequence from  $f_X$ . The classical Monte Carlo estimator of  $E_{f_X}g$  is  $\bar{g}_n^* := \frac{1}{n} \sum_{i=1}^n g(X_i^*)$ . If  $g \in L^1(f_X)$ , then, by the SLLN,  $\bar{g}_n^*$  is a strongly consistent estimator of  $E_{f_X}g$ . If, in addition,  $g \in L^2(f_X)$ , then standard results from iid theory tell us that, as  $n \rightarrow \infty$ ,

$$\sqrt{n}(\bar{g}_n^* - E_{f_X}g) \xrightarrow{d} N(0, \sigma^2). \quad (1.2.7)$$

If  $c_1 \neq 0$  (as will typically be the case), then  $\kappa^2/\sigma^2 > 1$ , so the asymptotic relative efficiency (ARE) of  $\bar{g}_n^*$  with respect to  $\bar{g}_n$  is larger than one. Therefore, if it is possible to make an iid draw from  $f_X$ , and the computational effort of doing so is similar to the effort of simulating a single iteration of the DA algorithm, then the classical Monte Carlo estimator is to be preferred over the estimator based on the DA algorithm. In the next section, we explain how these CLTs can be used in practice to choose an appropriate Monte Carlo sample size.

### 1.3 Choosing the Monte Carlo Sample Size

#### 1.3.1 Classical Monte Carlo

We begin by describing how the Monte Carlo sample size is chosen in the classical Monte Carlo context. Assume that  $g \in L^2(f_X)$  and recall that the classical Monte Carlo estimator of  $E_{f_X} g$  is  $\bar{g}_n^* = \frac{1}{n} \sum_{i=1}^n g(X_i^*)$ , where  $X_1^*, X_2^*, \dots$  are iid from  $f_X$ . The main motivation for using  $\bar{g}_n^*$  as an estimator of  $E_{f_X} g$  is that  $\bar{g}_n^*$  converges to  $E_{f_X} g$  almost surely as  $n \rightarrow \infty$ . Obviously, in practice we cannot use an infinite sample size, so we must find a finite value of  $n$  such that the error in  $\bar{g}_n^*$  is (likely to be) acceptably small. To make this more precise, suppose we are willing to live with an error of size  $\Delta$ . In other words, we would like to be able to assert that the interval given by  $\bar{g}_n^* \pm \Delta$  is highly likely to contain the true, unknown value of  $E_{f_X} g$ . As we now explain, this can be accomplished through routine use of the CLT given in (1.2.7).

Let  $\hat{\sigma}_n^2$  denote the usual sample variance of the  $g(X_i^*)$ s; that is,

$$\hat{\sigma}_n^2 = \frac{1}{(n-1)} \sum_{i=1}^n (g(X_i^*) - \bar{g}_n^*)^2.$$

Basic asymptotic theory tell us that, since  $\hat{\sigma}_n^2$  is a consistent estimator of  $\sigma^2$ ,

$$\frac{\sqrt{n}(\bar{g}_n^* - E_{f_X} g)}{\sqrt{\hat{\sigma}_n^2}} \xrightarrow{d} N(0, 1).$$

Thus, for large  $n$ , the interval  $\bar{g}_n^* \pm 2\hat{\sigma}_n/\sqrt{n}$  will contain the unknown value of  $E_{f_X}g$  with probability (approximately) equal to 0.95. With this in mind, we can proceed as follows. Choose an initial sample size, say  $n'$ , and make  $n'$  iid draws from  $f_X$ . (Hopefully,  $n'$  is large enough that  $\hat{\sigma}_{n'}$  is a reasonable estimate of  $\sigma^2$ .) If the observed value of  $2\hat{\sigma}_{n'}/\sqrt{n'}$  is less than  $\Delta$ , then the current estimate of  $E_{f_X}g$  is good enough and we stop. Otherwise, if  $2\hat{\sigma}_{n'}/\sqrt{n'} > \Delta$ , then additional simulation is required. Moreover, the current estimate of  $\sigma^2$  can be used to calculate approximately how much more simulation will be necessary to achieve the stated precision. Indeed, we require an  $n$  such that  $2\hat{\sigma}_n/\sqrt{n} < \Delta$ , so assuming that our estimate of  $\sigma^2$  has stabilized,  $n > 4\hat{\sigma}_{n'}^2/\Delta^2$  should suffice.

There are two major obstacles blocking the use of a similar program for choosing  $n$  in the DA context. First, as we have already seen, even when the Markov chain  $X$  is Harris ergodic, the second moment condition,  $g \in L^2(f_X)$ , is not enough to guarantee that the estimator  $\bar{g}_n$  satisfies a CLT. To be sure that CLTs hold for  $L^2(f_X)$  functions, the practitioner must either (i) employ a DA algorithm that is known to be geometrically ergodic, or (ii) establish geometric ergodicity of the DA algorithm in question. The second problem is that, even when the CLT in (1.2.6) is known to hold, consistent estimation of the asymptotic variance,  $\kappa^2$ , is a challenging problem because this variance has a fairly complex form and because the dependence among the variables in the Markov chain complicates asymptotic analysis. Consistent estimators of  $\kappa^2$  have been developed using techniques from time series analysis and using the method of batch means, but these estimators are much more complicated than  $\hat{\sigma}_n^2$  both practically and theoretically. Good entry points into the statistical literature on methods of estimating  $\kappa^2$  are Geyer (1992), Jones et al. (2006) and Flegal et al. (2008).

There is no getting around the fact that establishing the existence of CLTs is harder for Markov chains than it is for iid sequences. However, it is possible to circumvent the difficulties associated with consistent estimation of  $\kappa^2$ . Indeed, there is an alternative form of the CLT in (1.2.6) that is developed by introducing *regenerations* into the Markov chain. The advantage of this new CLT is that consistent estimation of its asymptotic variance is very simple. The price we have to pay for this added simplicity is that the user must develop a *minorization condition* for the Mtd  $k(\cdot|\cdot)$ . Fortunately, the form of  $k$  lends itself to constructing a minorization condition. Before we can fully explain regeneration and

minorization, we have to introduce three new Markov chains that are all closely related to  $X$ .

### 1.3.2 Three Markov chains that are closely related to $X$

Recall from Section 1.1 that, for fixed  $x \in \mathbf{X}$ , the function  $h(x', y'|x) = f_{X|Y}(x'|y')f_{Y|X}(y'|x)$  is a joint pdf in  $(x', y')$ . Now, define  $\tilde{k} : (\mathbf{X} \times \mathbf{Y}) \times (\mathbf{X} \times \mathbf{Y}) \rightarrow [0, \infty)$  as

$$\tilde{k}(x', y'|x, y) = h(x', y'|x) = f_{X|Y}(x'|y')f_{Y|X}(y'|x).$$

For each fixed  $(x, y) \in \mathbf{X} \times \mathbf{Y}$ ,  $\tilde{k}(x', y'|x, y)$  is nonnegative and integrates to 1. Hence, the function  $\tilde{k}$  is an Mtd that defines a Markov chain,  $(X, Y) = \{(X_n, Y_n)\}_{n=0}^{\infty}$ , with state space  $\mathbf{X} \times \mathbf{Y}$ . If the current state of the chain is  $(X_n, Y_n) = (x, y)$ , then the density of the next state,  $(X_{n+1}, Y_{n+1})$ , is  $\tilde{k}(\cdot, \cdot|x, y)$ . Furthermore, the chain  $(X, Y)$  has invariant density  $f(x, y)$ ; indeed,

$$\begin{aligned} \int_{\mathbf{X}} \int_{\mathbf{Y}} \tilde{k}(x', y'|x, y) f(x, y) dy dx &= f_{X|Y}(x'|y') \int_{\mathbf{X}} f_{Y|X}(y'|x) \left[ \int_{\mathbf{Y}} f(x, y) dy \right] dx \\ &= f_{X|Y}(x'|y') \int_{\mathbf{X}} f(x, y') dx \\ &= f_{X|Y}(x'|y') f_Y(y') \\ &= f(x', y'). \end{aligned}$$

We refer to  $(X, Y)$  as the ‘‘Gibbs chain’’ because it is, in fact, just the Markov chain that is induced by the two-variable Gibbs sampler based on the joint density  $f(x, y)$ . The analogue of condition  $\mathcal{K}$  for the Gibbs chain is

$$\text{Condition } \tilde{\mathcal{K}} : \tilde{k}(x', y'|x, y) > 0 \text{ for all } (x, y), (x', y') \in \mathbf{X} \times \mathbf{Y}.$$

Condition  $\tilde{\mathcal{K}}$  implies that the Gibbs chain is Harris ergodic. A sufficient condition for condition  $\tilde{\mathcal{K}}$  is that  $f(x, y) > 0$  for all  $(x, y) \in \mathbf{X} \times \mathbf{Y}$ .

The reader has probably already noticed that  $\tilde{k}(x', y'|x, y)$  does not actually depend on  $y$ .

In terms of the Markov chain, this means that the future state,  $(X_{n+1}, Y_{n+1})$ , depends on the current state,  $(X_n, Y_n)$  only through  $X_n$ . This fact can be used to show that the conditional distribution of  $X_{n+1}$  given  $(X_0, X_1, \dots, X_n)$  does not depend on  $(X_0, X_1, \dots, X_{n-1})$ . In other words, the sequence  $X = \{X_n\}_{n=0}^\infty$  is itself a Markov chain on  $\mathbf{X}$ . Moreover, its Mtd is

$$\int_{\mathbf{Y}} \tilde{k}(x', y' | x, y) dy' = \int_{\mathbf{Y}} f_{X|Y}(x' | y') f_{Y|X}(y' | x) dy' = k(x' | x) ;$$

that is, the marginal sequence  $X = \{X_n\}_{n=0}^\infty$  from the Gibbs chain is the original DA Markov chain. (This is why it made sense to use the symbol  $X_n$  to denote the  $x$ -coordinate of Gibbs chain.) It follows that we can view our estimator,  $\bar{g}_n = n^{-1} \sum_{i=0}^{n-1} g(X_i)$ , as being an estimator based on the Gibbs chain. Formally,  $\bar{g}_n = n^{-1} \sum_{i=0}^{n-1} \tilde{g}(X_i, Y_i)$ , where  $\tilde{g}(x, y) = g(x)$ . This correspondence allows us work with the Gibbs chain instead of  $X$ , which turns out to be easier because, unlike  $k$ ,  $\tilde{k}$  is a known closed form function.

Concerning simulation of the Gibbs chain, recall that our two-step procedure for simulating one iteration of the DA algorithm involves drawing from the joint pdf  $h(x', y' | x)$  and throwing away the  $y$ -coordinate. In other words, the two-step procedure given in Section 1.1 actually simulates the Gibbs chain and just ignores the  $y$ -coordinates.

Not surprisingly, the marginal sequence  $Y = \{Y_n\}_{n=0}^\infty$  from the Gibbs chain is also a Markov chain. This chain lives on  $\mathbf{Y}$  and its Mtd is

$$k_y(y' | y) = \int_{\mathbf{X}} f_{Y|X}(y' | x) f_{X|Y}(x | y) dx .$$

It follows that  $Y$  can be viewed as the Markov chain underlying a DA algorithm for the target density  $f_Y(y)$ , and, as such,  $Y$  is reversible with respect to  $f_Y$ . There is actually an alternative estimator of  $E_{f_X} g$  based on  $Y$  that we now describe. Define

$$\hat{g}(y) = \int_{\mathbf{X}} g(x) f_{X|Y}(x | y) dx ,$$

and note that  $\int_{\mathbf{Y}} \hat{g}(y) f_Y(y) dy = \int_{\mathbf{X}} g(x) f_X(x) dx = E_{f_X} g$ . Thus, if we can write  $\hat{g}$  in closed form, which is often the case in practice, then we can compute the alternative estimator of

$E_{f_X} g$  given by

$$\frac{1}{n} \sum_{i=0}^{n-1} \hat{g}(Y_i) . \quad (1.3.1)$$

If  $Y$  is Harris ergodic, then, like  $\bar{g}_n$ , the estimator (1.3.1) is strongly consistent for  $E_{f_X} g$ . In fact, Liu et al. (1994) proved that, if  $X_0 \sim f_X$  and  $Y_0 \sim f_Y$ , then the alternative estimator has a smaller (small sample) variance than  $\bar{g}_n$ . (Comparing variances is appropriate here since, if  $X_0 \sim f_X$  and  $Y_0 \sim f_Y$ , then both estimators are unbiased.) We note that the methods described below for computing a valid asymptotic standard error for  $\bar{g}_n$  can just as easily be applied to the estimator (1.3.1).

Finally, consider the Mtd given by

$$\tilde{k}(y', x'|y, x) = f_{Y|X}(y'|x') f_{X|Y}(x'|y) ,$$

and denote the corresponding Markov chain by  $(Y', X') = \{(Y'_n, X'_n)\}_{n=0}^{\infty}$ . Of course,  $(Y', X')$  is just the Markov chain induced by the two-variable Gibbs sampler for  $f(x, y)$  with the variables in the opposite order. The chain  $(Y', X')$  behaves just like  $(X, Y)$ . Indeed,  $f(x, y)$  remains invariant and, by symmetry, the marginal sequences  $\{X'_n\}_{n=0}^{\infty}$  and  $\{Y'_n\}_{n=0}^{\infty}$  are equivalent (distributionally) to  $X = \{X_n\}_{n=0}^{\infty}$  and  $Y = \{Y_n\}_{n=0}^{\infty}$ . Consequently, we can also view our estimator,  $\bar{g}_n = n^{-1} \sum_{i=0}^{n-1} g(X_i)$ , as being an estimator based on the chain  $(Y', X')$ ; i.e.,  $\bar{g}_n = n^{-1} \sum_{i=0}^{n-1} \tilde{g}(Y'_i, X'_i)$ , where  $\tilde{g}(y, x) = g(x)$ . In some cases, it is more convenient to work with  $\tilde{k}$  than with  $\tilde{k}$ . An important fact that will be used later is that all four of the Markov chains discussed in this section  $(X, Y, (X, Y)$  and  $(Y', X')$ ) converge at exactly the same rate (Diaconis et al., 2008; Roberts and Rosenthal, 2001). Therefore, either all four chains are geometrically ergodic, or none of them is. We now describe the minorization condition and how it is used to induce regenerations, which can in turn be used to derive the alternative CLT.

### 1.3.3 Minorization, regeneration and an alternative CLT

We assume throughout this subsection that the Gibbs chain is Harris ergodic. Suppose we can find a function  $s : \mathsf{X} \rightarrow [0, 1)$  with  $E_{f_X} s > 0$  and a joint pdf  $d : \mathsf{X} \times \mathsf{Y} \rightarrow [0, \infty)$  such that

$$\tilde{k}(x', y'|x, y) \geq s(x) d(x', y') \text{ for all } (x, y), (x', y') \in \mathsf{X} \times \mathsf{Y}. \quad (1.3.2)$$

Equation (1.3.2) is called a *minorization condition* (Jones and Hobert, 2001; Meyn and Tweedie, 1993; Roberts and Rosenthal, 2004). Here is a simple example.

EXAMPLE 2 CONT. Here we have  $\mathsf{X} = \mathsf{Y} = (0, 1)$ , and we can develop a minorization condition as follows

$$\begin{aligned} \tilde{k}(x', y'|x, y) &= f_{X|Y}(x'|y') f_{Y|X}(y'|x) \\ &= \frac{2x'}{(1-y'^2)} I(y' < x' < 1) \frac{1}{x} I(0 < y' < x < 1) \\ &\geq \frac{2x'}{(1-y'^2)} I(y' < x' < 1) \frac{1}{x} I(0 < y' < 0.5) I(0.5 < x < 1) \\ &= \frac{1}{x} I(0.5 < x < 1) \frac{2x'}{(1-y'^2)} I(y' < x' < 1) I(0 < y' < 0.5) \\ &= \left[ \frac{1}{2x} I(0.5 < x < 1) \right] \left[ \frac{4x'}{(1-y'^2)} I(y' < x' < 1) I(0 < y' < 0.5) \right] \\ &= s(x) d(x', y'), \end{aligned}$$

where we have used the fact that

$$\int_0^1 \int_0^1 \frac{2x}{(1-y^2)} I(y < x < 1) I(0 < y < 0.5) dx dy = \frac{1}{2}.$$

Note that the density  $d$  is not strictly positive on  $\mathsf{X} \times \mathsf{Y}$ .

The minorization condition (1.3.2) can be used to represent the Mtd  $\tilde{k}$  as a mixture of two other Mtds. First, define

$$r(x', y'|x, y) = \frac{\tilde{k}(x', y'|x, y) - s(x)d(x', y')}{1 - s(x)},$$

and note that  $r(x', y'|x, y)$  is an Mtd. Indeed, (1.3.2) implies that  $r$  is nonnegative and it's also clear that  $\int_{\mathbf{X}} \int_{\mathbf{Y}} r(x', y'|x, y) dy' dx' = 1$ . We can now express  $\tilde{k}$  as follows

$$\tilde{k}(x', y'|x, y) = s(x)d(x', y') + (1 - s(x))r(x', y'|x, y) . \quad (1.3.3)$$

If we think of  $s(x)$  and  $1 - s(x)$  as two fixed numbers in  $[0, 1]$  whose sum is 1, then the right-hand side of (1.3.3) can be viewed as a mixture of two Mtds,  $d(x', y')$  and  $r(x', y'|x, y)$ . Since  $d(x', y')$  does not depend on  $(x, y)$ , the Markov chain defined by  $d$  is actually an iid sequence and this is the key to introducing regenerations. Technically speaking, the regenerations do not occur in the Gibbs chain itself, but in an augmented Markov chain that we now describe.

For  $(x, y) \in \mathbf{X} \times \mathbf{Y}$ , let  $f_1(\delta|(x, y))$  denote a Bernoulli( $s(x)$ ) probability mass function; that is,  $f_1(1|(x, y)) = s(x)$  and  $f_1(0|(x, y)) = 1 - s(x)$ . Also, for  $(x', y'), (x, y) \in \mathbf{X} \times \mathbf{Y}$  and  $\delta \in \{0, 1\}$  define

$$f_2((x', y')|\delta, (x, y)) = d(x', y') I(\delta = 1) + r(x', y'|x, y) I(\delta = 0) . \quad (1.3.4)$$

Note that  $f_2$  is a pdf in  $(x', y')$ . Finally, define

$$k_s((x', y'), \delta'|(x, y), \delta) = f_1(\delta'|(x, y)) f_2((x', y')|\delta', (x, y)) . \quad (1.3.5)$$

Now,  $k_s$  is non-negative and

$$\sum_{\delta' \in \{0,1\}} \int_{\mathbf{Y}} \int_{\mathbf{X}} k_s((x', y'), \delta'|(x, y), \delta) dx' dy' = \sum_{\delta' \in \{0,1\}} f_1(\delta'|(x, y)) = 1 .$$

Therefore,  $k_s$  is an Mtd and the corresponding Markov chain, which we denote by  $((X, Y), \delta) = \{(X_n, Y_n), \delta_n\}_{n=0}^{\infty}$ , lives on  $(\mathbf{X} \times \mathbf{Y}) \times \{0, 1\}$ . This is called the *split chain* (Nummelin, 1984, Section 4.4).

Before we elucidate the regeneration properties of the split chain, we describe the relationship between the split chain and the Gibbs chain. Note that  $k_s$  does not actually depend on  $\delta$ . Thus, arguments similar to those used in Subsection 1.3.2 show that the marginal

sequence  $\{(X_n, Y_n)\}_{n=0}^\infty$  from the split chain is itself a Markov chain with Mtd given by

$$\begin{aligned} & k_s((x', y'), 1|(x, y), \delta) + k_s((x', y'), 0|(x, y), \delta) \\ &= f_1(1|(x, y))f_2((x', y')|1, (x, y)) + f_1(0|(x, y))f_2((x', y')|0, (x, y)) \\ &= s(x)d(x', y') + (1 - s(x))r(x', y'|x, y) \\ &= \tilde{k}(x', y'|x, y) . \end{aligned}$$

We conclude that the marginal sequence  $\{(X_n, Y_n)\}_{n=0}^\infty$  from the split chain is (distributionally) equivalent to the Gibbs chain. Moreover, the split chain inherits Harris ergodicity from the Gibbs chain (Nummelin, 1984, Section 4.4). As before, we can view the estimator  $\bar{g}_n$  as being based on the split chain.

The split chain experiences a regeneration every time the binary component visits the set  $\{1\}$ . To see this, suppose we start the split chain with  $\delta_0 = 1$  and  $(X_0, Y_0) \sim d(\cdot, \cdot)$ . It is clear from (1.3.5) and (1.3.4) that, no matter what the value of the current state,  $((X_n, Y_n), \delta_n)$ , if  $\delta_{n+1} = 1$  then  $(X_{n+1}, Y_{n+1}) \sim d(\cdot, \cdot)$  and the process stochastically restarts itself; that is, the Markov chain regenerates. We now use the regenerative structure of the split chain to recast our estimator of  $E_{f_X}g$  in such a way that iid theory can be used to analyze it. This leads to an alternative CLT whose asymptotic variance is very easy to estimate.

Let  $\tau_0, \tau_1, \tau_2, \dots$  denote the *regeneration times*; that is, the random times at which the split chain regenerates. Then  $\tau_0 = 0$  and, for  $t = 1, 2, 3, \dots$ , we have

$$\tau_t = \min \{i > \tau_{t-1} : \delta_i = 1\} .$$

This notation allows us to identify the “tours” that the split chain takes in between regenerations:

$$\left\{ \left( (X_{\tau_{t-1}}, Y_{\tau_{t-1}}), \delta_{\tau_{t-1}} \right), \dots, \left( (X_{\tau_t}, Y_{\tau_t}), \delta_{\tau_t} \right) : t = 1, 2, 3, \dots \right\} .$$

These tours are independent stochastic copies of each other, and hence standard techniques from iid theory (such as the SLLN and the CLT) can be used in the asymptotic analysis of the resulting ergodic averages. In other words, the regenerative structure that we have introduced

allows us to circumvent (to some extent) the complications caused by the dependence among the random vectors in the Markov chain.

Consider running the split chain for  $R$  tours; that is, the chain is started with  $\delta_0 = 1$  and  $(X_0, Y_0) \sim d(\cdot, \cdot)$  and is run until the  $R$ th time that a  $\delta_n = 1$ . (Some practical advice concerning simulation of the split chain will be given later.) For  $t = 1, 2, \dots, R$ , define  $N_t = \tau_t - \tau_{t-1}$ , which is the length of the  $t$ th tour, and  $S_t = \sum_{i=\tau_{t-1}}^{\tau_t-1} g(X_i)$ . Because the tours are independent stochastic copies of each other, the pairs  $(N_1, S_1), \dots, (N_R, S_R)$  are iid. The total length of the simulation is  $\sum_{t=1}^R N_t = \tau_R$ , which is, of course, random. Our estimator of  $E_{f_X} g$  will be

$$\bar{g}_R = \frac{1}{\tau_R} \sum_{i=0}^{\tau_R-1} g(X_i) = \frac{\sum_{t=1}^R S_t}{\sum_{t=1}^R N_t}.$$

Clearly, the only difference between  $\bar{g}_R$  and the usual ergodic average,  $\bar{g}_n$ , is that here, the sample size is random. However,  $\tau_R \rightarrow \infty$  almost surely as  $R \rightarrow \infty$  and it follows that  $\bar{g}_R$  is also strongly consistent for  $E_{f_X} g$  as  $R \rightarrow \infty$ . The advantage of  $\bar{g}_R$  over the usual ergodic average is that it can be expressed in terms of the iid pairs  $\{(N_t, S_t)\}_{t=1}^R$ . Results in Hobert et al. (2002) show that, if the Gibbs chain (or, equivalently, the DA Markov chain) is geometrically ergodic and  $E_{f_X} |g|^{2+\alpha} < \infty$  for some  $\alpha > 0$ , then as  $R \rightarrow \infty$

$$\sqrt{R}(\bar{g}_R - E_{f_X} g) \xrightarrow{d} N(0, \gamma^2), \quad (1.3.6)$$

where

$$\gamma^2 = \frac{E[(S_1 - N_1 E_{f_X} g)^2]}{[E N_1]^2}.$$

Note that this asymptotic variance is written in terms of a single tour,  $(N_1, S_1)$ . Results in Hobert et al. (2002) show that the geometric ergodicity of the Gibbs chain together with the “ $2 + \alpha$ ” moment condition on  $g$  imply that  $E N_1^2$  and  $E S_1^2$  are both finite. Once these moments are known to be finite, routine asymptotics can be used to show that

$$\hat{\gamma}_R^2 = \frac{R \sum_{t=1}^R (S_t - \bar{g}_R N_t)^2}{\tau_R^2}$$

is a strongly consistent estimator of  $\gamma^2$  as  $R \rightarrow \infty$ . Note the simple form of this estimator.

A couple of comments are in order concerning the two different CLTs, (1.3.6) and (1.2.6). First, both CLTs are based on the assumption that the DA Markov chain,  $X$ , is geometrically ergodic. However, while (1.2.6) requires only the usual second moment condition,  $E_{f_X} g^2 < \infty$ , (1.3.6) requires the slightly stronger condition that  $E_{f_X} |g|^{2+\alpha} < \infty$  for some  $\alpha > 0$ . Second, the two asymptotic variances are related via the formula  $\kappa^2 = \gamma^2 / E_{f_X} s$  (Hobert et al., 2002). This makes sense intuitively because  $E_{f_X} s$  is the average probability of regeneration (under stationarity) and hence  $1/E_{f_X} s$  seems like a reasonable guess at the average tour length.

We conclude that, if  $X$  is geometrically ergodic and the “ $2 + \alpha$ ” moment condition on  $g$  is satisfied, then we can employ the DA algorithm in the same way that classical Monte Carlo is used. Indeed, we can simulate  $R'$  tours of the split chain, where  $R'$  is some initial sample size. (Hopefully,  $R'$  is large enough that  $\hat{\gamma}_{R'}^2$  is a reasonable estimate of  $\gamma^2$ .) If  $2\hat{\gamma}_{R'}^2 / \sqrt{R'} \leq \Delta$ , then the current estimate of  $E_{f_X} g$  is good enough and we stop. Otherwise, if  $2\hat{\gamma}_{R'}^2 / \sqrt{R'} > \Delta$ , then additional tours must be simulated.

### 1.3.4 Simulating the split chain

Exploiting the techniques described in the previous subsection in practice requires the ability to simulate the split chain. The form of  $k_s$  actually lends itself to the sequential simulation technique described in Section 1.1. If the current state is  $((X_n, Y_n), \delta_n) = ((x, y), \delta)$ , then the future state,  $((X_{n+1}, Y_{n+1}), \delta_{n+1})$ , can be simulated as follows. First, draw  $\delta_{n+1} \sim \text{Bernoulli}(s(x))$  and then, conditional on  $\delta_{n+1} = \delta'$ , draw  $(X_{n+1}, Y_{n+1})$  from

$$f_2((\cdot, \cdot) | \delta', (x, y)) ;$$

i.e., if  $\delta' = 1$ , draw  $(X_{n+1}, Y_{n+1}) \sim d(\cdot, \cdot)$ , and if  $\delta' = 0$ , draw  $(X_{n+1}, Y_{n+1}) \sim r(\cdot, \cdot | x, y)$ . Here is an example where this method is viable.

EXAMPLE 2 CONT. Recall that we developed a minorization condition of the form (1.3.2) for this example earlier in this subsection. We now verify that it is straightforward to

simulate from  $d(\cdot, \cdot)$  and from  $r(\cdot, \cdot | x, y)$ . First, it is easy to show that if  $(U, V) \sim d(\cdot, \cdot)$ , then marginally,  $V$  is  $\text{Uniform}(0, 0.5)$ , and the conditional density of  $U$  given  $V = v$  is  $f_{X|Y}(u|v) = \frac{2u}{1-v^2} I(v < u < 1)$ . Hence, simulating from  $d$  is easy. Now consider  $r$ . Since,  $s(x) = 0$  when  $x \in (0, 0.5)$ , we must have  $r(x', y' | x, y) = \tilde{k}(x', y' | x, y)$  when  $x \in (0, 0.5)$ . On the other hand, when  $x \in (0.5, 1)$ , then routine calculations show that

$$r(x', y' | x, y) = \frac{2x'}{(1-y'^2)(x-0.5)} I(y' < x' < 1) I(0.5 < y' < x),$$

and, in this case, it's easy to show that if  $(U, V) \sim r(\cdot, \cdot | x, y)$ , then marginally,  $V$  is  $\text{Uniform}(0.5, x)$ , and the conditional density of  $U$  given  $V = v$  is  $f_{X|Y}(u|v)$ , so it is also easy to draw from  $r$ . Note that the supports of  $d(\cdot, \cdot)$  and  $r(\cdot, \cdot | x, y)$  are mutually exclusive. We conclude that the sequential method outlined above can be used to simulate the split chain in this example.

In the toy example just considered, it is straightforward to simulate from  $r(\cdot, \cdot | x, y)$ . However, this will typically not be the case in real examples where  $\tilde{k}(x', y' | x, y)$  is a high dimensional, complex Mtd. Fortunately, Mykland et al. (1995) noticed a clever way of circumventing the need to draw from  $r$ . Their idea amounts to using the sequential simulation technique, but in the opposite order. Indeed, one way to draw from  $(X_{n+1}, Y_{n+1}) | \delta_{n+1} | (X_n, Y_n)$  is to first draw from  $(X_{n+1}, Y_{n+1}) | (X_n, Y_n)$  and then from  $\delta_{n+1} | (X_{n+1}, Y_{n+1}), (X_n, Y_n)$ . A little thought reveals that these two steps are simple and do not involve drawing from  $r$ . First, we established above that  $(X_{n+1}, Y_{n+1}) | (X_n, Y_n) = (x, y) \sim \tilde{k}(\cdot, \cdot | x, y)$ , so this step can be accomplished by simulating a single iteration of the Gibbs chain (by drawing from  $f_{Y|X}$  and then from  $f_{X|Y}$ ). Furthermore, given  $(X_n, Y_n)$  and  $(X_{n+1}, Y_{n+1})$ ,  $\delta_{n+1}$  has a Bernoulli distribution with success probability given by

$$\Pr(\delta_{n+1} = 1 | X_n = x, Y_n = y, X_{n+1} = x', Y_{n+1} = y') = \frac{s(x) d(x', y')}{\tilde{k}(x', y' | x, y)}. \quad (1.3.7)$$

Here is a summary of how Mykland et al.'s (1995) method is used to simulate the split chain. If the current state is  $(X_n, Y_n) = (x, y)$ , then we simply draw  $(X_{n+1}, Y_{n+1})$  in the usual way, and then we go back and “fill in” the value of  $\delta_{n+1}$  by simulating a Bernoulli with

success probability (1.3.7). Even though we only draw from  $d$  once (at the start) and we never actually draw from  $r$  at all, there is a regeneration in the chain each time  $\delta_n = 1$ . In fact, we can even avoid the single draw from  $d$  (although, even in real problems, it is usually pretty easy to draw from  $d$ ). Starting the chain from an arbitrary point, but then throwing away everything from the beginning up to and including the first Bernoulli that equals 1, is equivalent to drawing  $(X_0, Y_0) \sim d(\cdot, \cdot)$ . Finally, note the rather striking fact that the only difference between simulating the split chain and the Gibbs chain is a single Bernoulli draw per iteration! In fact, if computer code is available that runs the DA algorithm, then a few minor modifications will yield a program that runs the split chain instead. Here is an example illustrating the use of (1.3.7).

EXAMPLE 2 CONT. If the  $n$ th and  $(n+1)$ st states of the Gibbs chain are  $(X_n, Y_n) = (x, y)$  and  $(X_{n+1}, Y_{n+1}) = (x', y')$ , then it must be the case that  $x, x' \in (0, 1)$  and  $y' \in (0, \min\{x, x'\})$ . Now, applying (1.3.7), the probability that a regeneration occurred is

$$\Pr(\delta_{n+1} = 1 \mid X_n = x, Y_n = y, X_{n+1} = x', Y_{n+1} = y') = I(0.5 < x < 1) I(0 < y' < 0.5).$$

In hindsight, this formula is actually “obvious.” First, if  $x \notin (0.5, 1)$ , then  $s(x) = 0$ , and regeneration could not have occurred. Likewise, if  $y' \notin (0, 0.5)$ , then  $d$  could not have been used to draw  $(X_{n+1}, Y_{n+1})$  so, again, regeneration could not have occurred. On the other hand, if  $x \in (0.5, 1)$  and  $y' \in (0, 0.5)$ , then there must have been a regeneration because  $r(\cdot, \cdot \mid x, y)$  could not have been used to draw  $(X_{n+1}, Y_{n+1})$ .

In the next subsection, we give a general method of developing the minorization condition (1.3.2).

### 1.3.5 A general method for constructing the minorization condition

The minorization condition for Example 1 was derived in a somewhat *ad-hoc* manner. We now describe a general recipe, due to Mykland et al. (1995), for constructing a minorization condition. This technique is most effective when  $f(x, y)$  is strictly positive on  $\mathsf{X} \times \mathsf{Y}$ . Fix a

“distinguished point”  $x_* \in \mathsf{X}$  and a set  $D \subset \mathsf{Y}$ . Then we can write

$$\begin{aligned} \tilde{k}(x', y'|x, y) &= f_{X|Y}(x'|y')f_{Y|X}(y'|x) \\ &= \frac{f_{Y|X}(y'|x)}{f_{Y|X}(y'|x_*)} f_{X|Y}(x'|y')f_{Y|X}(y'|x_*) \\ &\geq \left[ \inf_{y \in D} \frac{f_{Y|X}(y|x)}{f_{Y|X}(y|x_*)} \right] f_{X|Y}(x'|y')f_{Y|X}(y'|x_*)I_D(y') \\ &= c \left[ \inf_{y \in D} \frac{f_{Y|X}(y|x)}{f_{Y|X}(y|x_*)} \right] \frac{1}{c} f_{X|Y}(x'|y')f_{Y|X}(y'|x_*)I_D(y') , \end{aligned}$$

where

$$c = \int_{\mathsf{Y}} \int_{\mathsf{X}} f_{X|Y}(x|y)f_{Y|X}(y|x_*)I_D(y) dx dy = \int_D f_{Y|X}(y|x_*) dy .$$

Thus, we have a minorization condition  $\tilde{k}(x', y'|x, y) \geq s(x)d(x', y')$  with

$$s(x) = c \inf_{y \in D} \frac{f_{Y|X}(y|x)}{f_{Y|X}(y|x_*)} \quad \text{and} \quad d(x', y') = \frac{1}{c} f_{X|Y}(x'|y')f_{Y|X}(y'|x_*)I_D(y') .$$

Fortunately, the value of  $c$  is not required in practice. The success probability in (1.3.7) involves  $s(x)$  and  $d(x', y')$  only through their product, so  $c$  cancels out. Furthermore, it is possible to make draws from  $d(x', y')$  without knowing the value of  $c$ . We first draw  $Y'$  from its marginal density,  $c^{-1}f_{Y|X}(y'|x_*)I_D(y')$ , by repeatedly drawing from  $f_{Y|X}(\cdot|x_*)$  until the result is in the set  $D$ . Then, given  $Y' = y'$ , we draw  $X'$  from  $f_{X|Y}(\cdot|y')$ .

Since the asymptotics described in Subsection 1.3.3 are for large  $R$ , the more frequently the split chain regenerates, the better. Thus, in practice, one should choose the point  $x_*$  and the set  $D$  so that regenerations occur frequently. This can be done by trial and error. In applications, we have found it useful fix  $x_*$  (at a preliminary estimate of the mean of  $f_X$ ) and then vary the set  $D$ . Note that, according to (1.3.7), a regeneration could only have occurred if  $y' \in D$ , so it is tempting to make  $D$  large. However, as  $D$  gets larger,  $s(x)$  becomes smaller, which means that the probability of regeneration becomes smaller. Hence, a balance must be struck. For examples, see Mykland et al. (1995), Jones and Hobert (2001), Roy and Hobert (2007) and Tan and Hobert (2008). We now provide two examples illustrating Mykland et al.’s (1995) method.

EXAMPLE 3 CONT. Recall that  $X|Y = y \sim N(0, y^{-1})$  and  $Y|X = x \sim \text{Gamma}(\frac{5}{2}, \frac{x^2}{2} + 2)$ . Thus,

$$\begin{aligned} \frac{f_{Y|X}(y|x)}{f_{Y|X}(y|x_*)} &= \frac{[\Gamma(\frac{5}{2})]^{-1} (\frac{x^2}{2} + 2)^{\frac{5}{2}} y^{\frac{3}{2}} \exp\left\{-y\left(\frac{x^2}{2} + 2\right)\right\}}{[\Gamma(\frac{5}{2})]^{-1} (\frac{x_*^2}{2} + 2)^{\frac{5}{2}} y^{\frac{3}{2}} \exp\left\{-y\left(\frac{x_*^2}{2} + 2\right)\right\}} \\ &= \left(\frac{x^2 + 4}{x_*^2 + 4}\right)^{\frac{5}{2}} \exp\left\{-\frac{y}{2}(x^2 - x_*^2)\right\}. \end{aligned}$$

So if we take  $D = [d_1, d_2]$  where  $0 < d_1 < d_2 < \infty$ , we have

$$\inf_{y \in D} \frac{f_{Y|X}(y|x)}{f_{Y|X}(y|x_*)} = \left(\frac{x^2 + 4}{x_*^2 + 4}\right)^{\frac{5}{2}} \exp\left\{-\frac{d_2}{2}(x^2 - x_*^2)I(x^2 > x_*^2) - \frac{d_1}{2}(x^2 - x_*^2)I(x^2 \leq x_*^2)\right\}.$$

Thus,

$$\begin{aligned} \Pr(\delta_{n+1} = 1 | X_n = x, Y_n = y, X_{n+1} = x', Y_{n+1} = y') &= \frac{s(x) d(x', y')}{\tilde{k}(x', y'|x, y)} \\ &= \left[ \inf_{y \in [d_1, d_2]} \frac{f_{Y|X}(y|x)}{f_{Y|X}(y|x_*)} \right] \frac{f_{Y|X}(y'|x_*)}{f_{Y|X}(y'|x)} I_{[d_1, d_2]}(y') \\ &= \exp\left\{(x^2 - x_*^2) \left[ \frac{y'}{2} - \frac{d_2}{2} I(x^2 > x_*^2) - \frac{d_1}{2} I(x^2 \leq x_*^2) \right]\right\} I_{[d_1, d_2]}(y'). \end{aligned}$$

A draw from  $d(x', y')$  can be made by drawing a truncated gamma and then a normal.

Here is a more realistic example.

EXAMPLE 4 CONT. The variable of interest is  $(\mu, \sigma^2)$ , which lives in  $\mathbf{X} = \mathbb{R} \times \mathbb{R}_+$ , and the augmented data,  $y$ , live in  $\mathbf{Y} = \mathbb{R}_+^m$ . In order to keep the notation under control, we use the symbol  $\eta$  in place of  $\sigma^2$ . In this example, it turns out to be more convenient to use  $\tilde{k}$ , which is given by

$$\tilde{k}(y', (\mu', \eta') | y, (\mu, \eta)) = \pi(y' | \mu', \eta', z) \pi(\mu', \eta' | y, z),$$

where the conditional densities on the right-hand side are defined in (1.1.6)-(1.1.8). Fix a

distinguished point  $y_* \in \mathsf{Y}$  and let  $D = [d_1, d_2] \times [d_3, d_4]$  where  $-\infty < d_1 < d_2 < \infty$  and  $0 < d_3 < d_4 < \infty$ . Now, letting  $y_s$  denote the sum of the components of  $y_*$ , we have

$$\begin{aligned} & \frac{\pi(\mu, \eta | y, z)}{\pi(\mu, \eta | y_*, z)} \\ &= \frac{\frac{\sqrt{y}}{\sqrt{\eta 2\pi}} \exp \left\{ -\frac{y}{2\eta} (\mu - \hat{\mu}(y))^2 \right\} \left( \frac{y \hat{\sigma}^2(y)}{2} \right)^{\frac{m+1}{2}} \Gamma^{-1} \left( \frac{m+1}{2} \right) \eta^{-\frac{m+1}{2}-1} \exp \left\{ -\frac{y \hat{\sigma}^2(y)}{2\eta} \right\}}{\frac{\sqrt{y_s}}{\sqrt{\eta 2\pi}} \exp \left\{ -\frac{y_s}{2\eta} (\mu - \hat{\mu}(y_*))^2 \right\} \left( \frac{y_s \hat{\sigma}^2(y_*)}{2} \right)^{\frac{m+1}{2}} \Gamma^{-1} \left( \frac{m+1}{2} \right) \eta^{-\frac{m+1}{2}-1} \exp \left\{ -\frac{y_s \hat{\sigma}^2(y_*)}{2\eta} \right\}} \\ &= \frac{\sqrt{y}}{\sqrt{y_s}} \left( \frac{y \hat{\sigma}^2(y)}{y_s \hat{\sigma}^2(y_*)} \right)^{\frac{m+1}{2}} \exp \left\{ -\frac{1}{2\eta} \left[ y (\mu - \hat{\mu}(y))^2 + y \hat{\sigma}^2(y) - y_s (\mu - \hat{\mu}(y_*))^2 - y_s \hat{\sigma}^2(y_*) \right] \right\} \\ &= \frac{\sqrt{y}}{\sqrt{y_s}} \left( \frac{y \hat{\sigma}^2(y)}{y_s \hat{\sigma}^2(y_*)} \right)^{\frac{m+1}{2}} \exp \left\{ -\frac{1}{2\eta} Q(\mu; y, y_*) \right\}, \end{aligned}$$

where  $Q(\mu; y, y_*)$  is a quadratic function of  $\mu$  whose coefficients are determined by  $y$  and  $y_*$ . Now consider minimizing the exponential over  $(\mu, \eta) \in [d_1, d_2] \times [d_3, d_4]$ . Let  $\tilde{\mu}$  denote the maximizer of  $Q(\mu; y, y_*)$  over  $\mu \in [d_1, d_2]$ , which is easy to compute once  $y$  and  $y_*$  are specified. Clearly, if  $Q(\tilde{\mu}; y, y_*) \geq 0$ , then the exponential is minimized at  $(\mu, \eta) = (\tilde{\mu}, d_3)$ . On the other hand, if  $Q(\tilde{\mu}; y, y_*) < 0$ , then the minimizer is  $(\mu, \eta) = (\tilde{\mu}, d_4)$ . Let  $\underline{\eta} = d_3$  if  $Q(\tilde{\mu}; y, y_*) \geq 0$  and  $d_4$  if  $Q(\tilde{\mu}; y, y_*) < 0$ . Then we can write

$$s(y) = c \inf_{(\mu, \eta) \in [d_1, d_2] \times [d_3, d_4]} \frac{\pi(\mu, \eta | y, z)}{\pi(\mu, \eta | y_*, z)} = c \frac{\sqrt{y}}{\sqrt{y_s}} \left( \frac{y \hat{\sigma}^2(y)}{y_s \hat{\sigma}^2(y_*)} \right)^{\frac{m+1}{2}} \exp \left\{ -\frac{1}{2\underline{\eta}} Q(\tilde{\mu}; y, y_*) \right\},$$

and

$$d(y', (\mu', \eta')) = \frac{1}{c} \pi(y' | \mu', \eta', z) \pi(\mu', \eta' | y_*, z) I_D(\mu', \eta').$$

Putting all of this together, if the  $n$ th and  $(n+1)$ st states of the Gibbs chain are  $(X_n, Y_n) = ((\mu, \eta), y)$  and  $(X_{n+1}, Y_{n+1}) = ((\mu', \eta'), y')$ , then the probability that a regeneration occurred;

i.e., that  $\delta_{n+1} = 1$ , is given by

$$\begin{aligned} \frac{s(y) d(y', (\mu', \eta'))}{\tilde{k}(y', (\mu', \eta')|y, (\mu, \eta))} &= \left[ \inf_{(\mu, \eta) \in [d_1, d_2] \times [d_3, d_4]} \frac{\pi(\mu, \eta|y, z)}{\pi(\mu, \eta|y_*, z)} \right] \frac{\pi(\mu', \eta'|y_*, z)}{\pi(\mu', \eta'|y, z)} I_D(\mu', \eta') \\ &= \exp \left\{ -\frac{1}{2\underline{\eta}} Q(\tilde{\mu}; y, y_*) + \frac{1}{2\eta'} Q(\mu'; y_*, y) \right\} I_D(\mu', \eta'). \end{aligned}$$

In the final section of this chapter, we describe a simple method of improving the DA algorithm.

## 1.4 Improving the DA Algorithm

Suppose the current state of the DA algorithm is  $X_n = x$ . As we know, the move to  $X_{n+1}$  involves two steps: draw  $Y \sim f_{Y|X}(\cdot|x)$  and then, conditional on  $Y = y$ , draw  $X_{n+1} \sim f_{X|Y}(\cdot|y)$ . Consider adding an extra step in between these two. Suppose that, *after* having drawn  $Y = y$ , but *before* drawing  $X_{n+1}$ , a random move is made from  $y$  to a new point in  $\mathsf{Y}$ , call it  $Y'$ . Then, conditional on  $Y' = y'$ , draw  $X_{n+1} \sim f_{X|Y}(\cdot|y')$ . Graphically, we are changing the algorithm from  $X \rightarrow Y \rightarrow X'$  to  $X \rightarrow Y \rightarrow Y' \rightarrow X'$ . Of course, this must all be done subject to the restriction that  $f_X$  remains invariant. Intuitively, this extra random move within  $\mathsf{Y}$  should reduce the correlation between  $X_n$  and  $X_{n+1}$ , thereby improving the mixing properties of the DA Markov chain. On the other hand, the new algorithm requires more computational effort per iteration and this must be weighed against any improvement in mixing. In this section, we describe techniques for constructing relatively inexpensive extra moves that often result in dramatic improvements in mixing. Here is a brief description of one of these techniques.

Suppose that  $\mathsf{G} \subset \mathbb{R}^d$  and that we have a class of functions  $t_g : \mathsf{Y} \rightarrow \mathsf{Y}$  indexed by  $g \in \mathsf{G}$ . In Section 1.4.4 we show that, if this class possesses a certain group structure, then there exists a parametric family of densities on  $\mathsf{G}$ , indexed by  $y \in \mathsf{Y}$ , call it  $\xi(g; y)$ , that can be used to make the extra move  $Y \rightarrow Y'$ . It proceeds as follows. Given  $Y = y$ , draw  $G \sim \xi(\cdot; y)$ ,

call the result  $g$ , and set  $Y' = t_g(y)$ . In other words, the extra move takes  $y$  to the random point  $Y' = t_G(y)$  where  $G$  is drawn from a distribution that is constructed to ensure that  $f_X$  remains invariant. Typically,  $d$  is small, say 1 or 2, so drawing from  $\xi(\cdot; y)$  is inexpensive. A potential downside of small  $d$  is that, for fixed  $y$ , the set  $\{t_g(y) : g \in G\}$  is a low dimensional subset of  $Y$  (that includes the point  $y$ ). Thus, the potential “shake-up” resulting from the move to  $Y' = t_G(y)$  may not be significant. However, it turns out that, even when  $d = 1$ , this shake-up often results in huge improvements. We now begin a careful development of these ideas.

#### 1.4.1 The PX-DA and marginal augmentation algorithms

Recall that our DA algorithm is based on the pdf  $f(x, y)$  whose  $x$ -marginal is  $f_X$ . As above, let  $G \subset \mathbb{R}^d$  and suppose that we have a class of functions  $t_g : Y \rightarrow Y$  indexed by  $g \in G$ . Assume that for each fixed  $g$ ,  $t_g(y)$  is one-to-one and differentiable in  $y$ . Let  $J_g(z)$  denote the Jacobian of the transformation  $z = t_g^{-1}(y)$ , so, e.g., in the univariate case,  $J_g(z) = \frac{\partial}{\partial z} t_g(z)$ . Note that

$$\int_Y f(x, t_g(y)) |J_g(y)| dy = \int_Y f(x, z) dz = f_X(x). \quad (1.4.1)$$

Now suppose that  $w : G \rightarrow [0, \infty)$  is a pdf and define  $f^{(w)} : X \times Y \times G \rightarrow [0, \infty)$  as follows

$$f^{(w)}(x, y, g) = f(x, t_g(y)) |J_g(y)| w(g). \quad (1.4.2)$$

It is clear from (1.4.1) that  $f^{(w)}(x, y, g)$  is a pdf whose  $x$ -marginal is  $f_X(x)$ , and hence the pdf defined by

$$f^{(w)}(x, y) = \int_G f^{(w)}(x, y, g) dg$$

also has  $f_X$  as its  $x$ -marginal. Thus, if straightforward sampling from  $f_{X|Y}^{(w)}(x|y)$  and  $f_{Y|X}^{(w)}(y|x)$  is possible, then we have a new DA algorithm that can be compared with the one based on  $f(x, y)$ . (For the remainder of this chapter, we assume that all Markov chains on  $X$  are Harris ergodic.) As we will see, it is often possible to choose  $t_g$  and  $w$  in such a way that there is little difference between these two DA algorithms in terms of computational effort per iteration. However, under mild regularity conditions that are described below, the new algorithm

beats the original in terms of both convergence rate and ARE. The idea of introducing the extra parameter,  $g$ , to form a new DA algorithm was developed independently by Meng and van Dyk (1999), who called it *marginal augmentation*, and Liu and Wu (1999), who called it *parameter expanded-data augmentation* (or PX-DA). We find Liu and Wu's (1999) terminology a bit more convenient, so we call the new DA algorithm based on  $f^{(w)}(x, y)$  a *PX-DA algorithm*. Here's a simple example.

EXAMPLE 3 CONT. Set  $G = \mathbb{R}_+$  and let  $t_g(y) = gy$ . If we take  $w(g)$  to be a Gamma( $\alpha, \beta$ ) pdf, then we have

$$\begin{aligned} f^{(w)}(x, y, g) &= f(x, t_g(y)) |J_g(y)| w(g) \\ &= \left[ \frac{4}{\sqrt{2\pi}} (gy)^{\frac{3}{2}} \exp \left\{ -gy \left( \frac{x^2}{2} + 2 \right) \right\} I_{\mathbb{R}_+}(y) \right] (g) \left[ \frac{\beta^\alpha}{\Gamma(\alpha)} g^{\alpha-1} \exp\{-g\beta\} I_{\mathbb{R}_+}(g) \right]. \end{aligned}$$

It follows that

$$f^{(w)}(x, y) = \int_{\mathbb{R}_+} f^{(w)}(x, y, g) dg = \frac{4\beta^\alpha \Gamma\left(\frac{5}{2} + \alpha\right)}{\Gamma(\alpha)\sqrt{2\pi}} y^{\frac{3}{2}} \left[ y \left( \frac{x^2}{4} + 1 \right) + \beta \right]^{-\left(\frac{5}{2} + \alpha\right)}.$$

According to the theory above, every  $(\alpha, \beta) \in \mathbb{R}_+ \times \mathbb{R}_+$  yields a different version of  $f^{(w)}(x, y)$  and every one of them has the Student's  $t$  density with 4 degrees of freedom as its  $x$ -marginal.

Now consider the conditional densities under  $f^{(w)}(x, y)$ . It's easy to show that  $f_{X|Y}^{(w)}(\cdot|y)$  is a scaled Student's  $t$  density and that  $f_{Y|X}^{(w)}(\cdot|x)$  is a scaled  $F$  density. In fact, if the current state of the PX-DA Markov chain is  $X_n = x$ , then the next state,  $X_{n+1}$ , can be simulated by performing the following two steps:

1. Draw  $U$  from the  $F$ -distribution with 5 numerator degrees of freedom and  $2\alpha$  denominator degrees of freedom, and call the realized value  $u$ . Then set  $y = \frac{10\beta}{\alpha(x^2+4)} u$ .
2. Draw  $V$  from the Student's  $t$  distribution with  $2(\alpha + 2)$  degrees of freedom, and set  $X_{n+1} = \sqrt{\frac{2(y+\beta)}{y(\alpha+2)}} V$ .

(Note that Step 2 is as difficult as drawing directly from the target pdf,  $f_X$ , which is a

Student's  $t$  density, but keep in mind that this is just a toy example that we are using for illustration.) We now have infinitely many viable PX-DA algorithms - one for each  $(\alpha, \beta)$  pair. This brings up an obvious question. Are any of these PX-DA algorithms better than the original DA algorithm, and if so, is there a best one? These questions are answered below.

In the toy example just considered, the conditional densities  $f_{X|Y}^{(w)}$  and  $f_{Y|X}^{(w)}$  have standard forms. Unfortunately, in real examples, it will typically be impossible to sample directly from (or even compute) these conditionals. However, by exploiting the relationship between  $f^{(w)}(x, y)$  and  $f^{(w)}(x, y, g)$ , it is possible to develop *indirect* methods of drawing from  $f_{X|Y}^{(w)}$  and  $f_{Y|X}^{(w)}$  that use only draws from  $f_{X|Y}$ ,  $f_{Y|X}$ ,  $w(g)$  and one other density. (Recall that we have been operating under the assumption that it is easy to sample from  $f_{X|Y}$  and  $f_{Y|X}$  since Section 1.1.) We begin with  $f_{Y|X}^{(w)}(y|x)$ . Note that

$$\begin{aligned} f_{Y|X}^{(w)}(y|x) &= \frac{\int_{\mathbf{G}} f^{(w)}(x, y, g) dg}{f_X(x)} \\ &= \int_{\mathbf{G}} \frac{f(x, t_g(y))}{f_X(x)} |J_g(y)| w(g) dg \\ &= \int_{\mathbf{G}} f_{Y|X}(t_g(y)|x) |J_g(y)| w(g) dg . \end{aligned} \tag{1.4.3}$$

Now suppose that  $Y' \sim f_{Y|X}(\cdot|x)$ ,  $G \sim w(\cdot)$ , and  $Y'$  and  $G$  are independent. Then the integrand in (1.4.3) is the joint density of  $(G, Y)$  where  $Y = t_G^{-1}(Y')$ . Consequently,  $Y = t_G^{-1}(Y')$  has density  $f_{Y|X}^{(w)}(\cdot|x)$ . This provides a simple method of drawing from  $f_{Y|X}^{(w)}(\cdot|x)$ . Indeed, we draw  $Y'$  and  $G$  independently from  $f_{Y|X}(\cdot|x)$  and  $w(\cdot)$ , respectively, and then take  $Y = t_G^{-1}(Y')$ .

Sampling from  $f_{X|Y}^{(w)}$  is a little trickier. Clearly,

$$f_{X|Y}^{(w)}(x|y) = \int_{\mathbf{G}} f_{X,G|Y}^{(w)}(x, g|y) dg = \int_{\mathbf{G}} f_{X|Y,G}^{(w)}(x|y, g) f_{G|Y}^{(w)}(g|y) dg .$$

Thus, we can use the sequential simulation technique from Section 1.1 to draw from  $f_{X|Y}^{(w)}(x|y)$  as follows. First, draw  $G \sim f_{G|Y}^{(w)}(\cdot|y)$  and then, conditional on  $G = g$ , draw  $X \sim f_{X|Y,G}^{(w)}(\cdot|y, g)$ .

But now the question is: Can we draw from  $f_{G|Y}^{(w)}$  and  $f_{X|Y,G}^{(w)}$ ? It's actually simple to draw from  $f_{X|Y,G}^{(w)}$  because

$$f_{X|Y,G}^{(w)}(x|y, g) = \frac{f^{(w)}(x, y, g)}{\int_{\mathbf{X}} f^{(w)}(x, y, g) dx} = \frac{f(x, t_g(y)) |J_g(y)| w(g)}{f_Y(t_g(y)) |J_g(y)| w(g)} = f_{X|Y}(x|t_g(y)).$$

In other words, drawing from  $f_{X|Y,G}^{(w)}(\cdot|y, g)$  is equivalent to drawing from  $f_{X|Y}(\cdot|t_g(y))$ . Now,

$$f_{G|Y}^{(w)}(g|y) = \frac{\int_{\mathbf{X}} f^{(w)}(x, y, g) dx}{\int_{\mathbf{G}} \int_{\mathbf{X}} f^{(w)}(x, y, g) dx dg} \propto \int_{\mathbf{X}} f^{(w)}(x, y, g) dx = f_Y(t_g(y)) |J_g(y)| w(g).$$

There is no simple trick for drawing from  $f_{G|Y}^{(w)}$ . Moreover, at first glance, sampling from the normalized version of  $f_Y(t_g(y)) |J_g(y)| w(g)$  appears challenging because this function involves  $f_Y$ , from which it is impossible to sample. (Indeed, if we could draw directly from  $f_Y$ , then we could use the sequential simulation technique to get exact draws from the target,  $f_X$ , and we wouldn't need MCMC!) Fortunately,  $g$  typically has much lower dimension than  $y$  and in such cases it is often possible to draw from  $f_{G|Y}^{(w)}(g|y)$  despite the intractability of  $f_Y$ . Hence, our method of drawing from  $f_{X|Y}^{(w)}(\cdot|y)$  is as follows. Draw  $G \sim f_{G|Y}^{(w)}(\cdot|y)$ , and then conditional on  $G = g$ , draw  $X \sim f_{X|Y}(\cdot|t_g(y))$ .

As we know from previous sections, performing one iteration of the PX-DA algorithm entails drawing from  $f_{Y|X}^{(w)}(\cdot|x)$  and then from  $f_{X|Y}^{(w)}(\cdot|y)$ . Liu and Wu (1999) noticed that making these two draws using the indirect techniques described above can be represented as a *three-step* procedure in which the first and third steps *are the same* as the original DA algorithm. Indeed, if the current state of the PX-DA Markov chain is  $X_n = x$ , then we can simulate  $X_{n+1}$  as follows.

---

One iteration of the PX-DA Algorithm:

1. Draw  $Y \sim f_{Y|X}(\cdot|x)$ , and call the observed value  $y$ .
  2. Draw  $G \sim w(\cdot)$ , call the result  $g$ , then draw  $G' \sim f_{G|Y}^{(w)}(\cdot|t_g^{-1}(y))$ , call the result  $g'$ , and finally set  $y' = t_{g'}(t_g^{-1}(y))$ .
  3. Draw  $X_{n+1} \sim f_{X|Y}(\cdot|y')$ .
- 

Here is a recapitulation of what has been done so far in this subsection. We started with a DA algorithm for  $f_X$  based on a joint density  $f(x, y)$ . The density  $f(x, y)$  was used to create an entire family of joint densities,  $f^{(w)}(x, y)$ , one for each density  $w(\cdot)$ . Each member of this family has  $f_X$  as its  $x$ -marginal and can therefore be used to create a new DA algorithm. We call these PX-DA algorithms. Running a PX-DA algorithm requires drawing from  $f_{X|Y}^{(w)}$  and  $f_{Y|X}^{(w)}$ , and simple, indirect methods of making these draws were developed. Finally, we provided a representation of the PX-DA algorithm as a *three-step* algorithm in which the first and third steps are the same as the two steps of the original DA algorithm.

From a computational standpoint, the only difference between the original DA algorithm and the PX-DA algorithm is that one extra step (Step 2) must be performed at each iteration of the PX-DA algorithm. However, when  $g$  has relatively low dimension, as is usually the case in practice, the computational cost of the extra step is inconsequential compared to the cost of Steps 1 and 3. In such cases, the DA and PX-DA algorithms are (essentially) equivalent in terms of cost per iteration. What is amazing is the extent to which the mixing properties of the DA algorithm can be improved without really altering the computational complexity of the algorithm (see; e.g., Liu and Wu, 1999; Meng and van Dyk, 1999; van Dyk and Meng, 2001). Moreover, there is empirical evidence that the relative improvement of PX-DA over DA actually increases as the dimension of  $\mathbf{X}$  increases (Meng and van Dyk, 1999). Subsection 1.4.3 contains a rigorous theoretical comparison of the DA and PX-DA algorithms. We end this subsection with a real example that was developed and studied in

Liu and Wu (1999), van Dyk and Meng (2001) and Roy and Hobert (2007).

EXAMPLE 5 CONT. In this example,  $\pi(\beta, y|z)$  plays the role of  $f(x, y)$ . Take  $\mathbf{G} = \mathbb{R}_+$  and  $t_g(y) = gy = (gy_1, \dots, gy_m)$  and take  $w$  as follows

$$w(g; \alpha, \delta) = \frac{2\delta^\alpha}{\Gamma(\alpha)} g^{2\alpha-1} e^{-g^2\delta} I_{\mathbb{R}_+}(g), \quad (1.4.4)$$

where  $\alpha, \delta \in \mathbb{R}_+$ . This is just the density of the square root of a gamma variate; that is, if  $U \sim \text{Gamma}(\alpha, \delta)$ , then  $G = \sqrt{U}$  has density (1.4.4). Substituting (1.1.10) into (1.1.9) and integrating with respect to  $\beta$  shows that

$$\pi(y|z) = \frac{\exp\left\{-\frac{y^T(I-H)y}{2}\right\}}{|V^T V|^{\frac{1}{2}} c(z) (2\pi)^{\frac{m-p}{2}}} \prod_{i=1}^m \left\{ I_{\mathbb{R}_+}(y_i) I_{\{1\}}(z_i) + I_{\mathbb{R}_-}(y_i) I_{\{0\}}(z_i) \right\}.$$

Thus,

$$\begin{aligned} f_{G|Y}^{(w)}(g|y) &\propto \pi(t_g(y)|z) |J_g(y)| w(g) \\ &\propto \left[ \exp\left\{-\frac{1}{2}(gy)^T(I-H)(gy)\right\} \right] (g^m) \left[ g^{2\alpha-1} \exp\{-g^2\delta\} I_{\mathbb{R}_+}(g) \right] \\ &= \exp\left\{-g^2 \left[ \frac{y^T(I-H)y}{2} + \delta \right]\right\} g^{m+2\alpha-1} I_{\mathbb{R}_+}(g). \end{aligned}$$

Note that  $f_{G|Y}^{(w)}(g|y)$  has the same form as  $w(g; \alpha, \delta)$ , which means that a draw from  $f_{G|Y}^{(w)}(g|y)$  can be made by simulating a gamma and taking its square root. Putting all of this together, if the current state of the PX-DA algorithm is  $X_n = \beta$ , then we simulate the next state,  $X_{n+1}$ , by performing the following three steps:

1. Draw  $Y_1, \dots, Y_m$  independently such that  $Y_i \sim \text{TN}(v_i^T \beta, 1, z_i)$ , and call the result  $y = (y_1, \dots, y_m)^T$ .
2. Draw  $U \sim \text{Gamma}(\alpha, \delta)$ , call the result  $u$ , and set  $\tilde{y} = y/\sqrt{u}$ . Draw

$$V \sim \text{Gamma}\left(\frac{m}{2} + \alpha, \frac{\tilde{y}^T(I-H)\tilde{y}}{2} + \delta\right),$$

call the result  $v$ , and set  $y' = \sqrt{v}\tilde{y}$ .

3. Draw  $X_{n+1} \sim N(\hat{\beta}(y'), (V^T V)^{-1})$ .

Sampling from the truncated normal distribution is typically done using an accept-reject algorithm, and Step 1 of the above procedure involves the simulation of  $m$  truncated normals. Obviously, the computational burden of Step 2, which requires only two univariate draws from the gamma distribution, is relatively minor. On the other hand, as the examples in Liu and Wu (1999) and van Dyk and Meng (2001) demonstrate, the PX-DA algorithm mixes much faster than the DA algorithm.

As a prelude to our theoretical comparison of DA and PX-DA, we introduce a bit of operator theory.

#### 1.4.2 The operator associated with a reversible Markov chain

It is well known that techniques from spectral theory (see, e.g., Rudin, 1991, Part III) can be used to analyze reversible Markov chains. The reason for this is that every reversible Markov chain defines a self-adjoint operator on the space of functions that are square integrable with respect to the invariant density. Examples of the use of spectral theory in the analysis of reversible Markov chains can be found in Diaconis and Stroock (1991), Chan and Geyer (1994), Liu et al. (1994, 1995), Roberts and Rosenthal (1997) and Mira and Geyer (1999). Our theoretical comparison of PX-DA and DA involves ideas from this theory.

Define

$$L_0^2(f_X) = \left\{ h \in L^2(f_X) : \int_{\mathbf{X}} h(x) f_X(x) dx = 0 \right\},$$

and, for  $g, h \in L_0^2(f_X)$ , define the inner product as  $\langle g, h \rangle = \int_{\mathbf{X}} g(x) h(x) f_X(x) dx$ . The corresponding norm is given by  $\|g\| = \sqrt{\langle g, g \rangle}$ . Let  $a : \mathbf{X} \times \mathbf{X} \rightarrow [0, \infty)$  denote a generic Mtd that is reversible with respect to  $f_X$ ; that is,  $a(x'|x)f_X(x) = a(x|x')f_X(x')$  for all  $x, x' \in \mathbf{X}$ . Let  $\Psi = \{\Psi_n\}_{n=0}^{\infty}$  denote the corresponding Markov chain and assume that  $\Psi$  is Harris ergodic. The Mtd  $a$  defines an operator,  $A$ , that maps  $g \in L_0^2(f_X)$  to a new function in

$L_0^2(f_X)$  given by

$$(Ag)(x) = \int_{\mathbf{X}} g(x') a(x'|x) dx' .$$

Note that  $(Ag)(x) = \mathbb{E}[g(\Psi_{n+1})|\Psi_n = x]$ . To verify that  $Ag$  is square integrable with respect to  $f_X$ , use Jensen's inequality, Fubini, the invariance of  $f_X$ , and the fact that  $g \in L_0^2(f_X)$  as follows

$$\begin{aligned} \int_{\mathbf{X}} [(Ag)(x)]^2 f_X(x) dx &= \int_{\mathbf{X}} \left[ \int_{\mathbf{X}} g(x') a(x'|x) dx' \right]^2 f_X(x) dx \\ &\leq \int_{\mathbf{X}} \left[ \int_{\mathbf{X}} g^2(x') a(x'|x) dx' \right] f_X(x) dx \\ &= \int_{\mathbf{X}} g^2(x') \left[ \int_{\mathbf{X}} a(x'|x) f_X(x) dx \right] dx' \\ &= \int_{\mathbf{X}} g^2(x') f_X(x') dx' < \infty . \end{aligned}$$

That  $Ag$  has mean zero follows from Fubini, the invariance of  $f_X$ , and the fact that  $g$  has mean zero:

$$\begin{aligned} \int_{\mathbf{X}} (Ag)(x) f_X(x) dx &= \int_{\mathbf{X}} \left[ \int_{\mathbf{X}} g(x') a(x'|x) dx' \right] f_X(x) dx \\ &= \int_{\mathbf{X}} g(x') \left[ \int_{\mathbf{X}} a(x'|x) f_X(x) dx \right] dx' \\ &= \int_{\mathbf{X}} g(x') f_X(x') dx' = 0 . \end{aligned}$$

We now demonstrate that the operator  $A$  is indeed self-adjoint (Rudin, 1991, Section 12).

Using Fubini and the fact that  $a(x'|x)f_X(x)$  is symmetric in  $(x, x')$ , we have for  $g, h \in L_0^2(f_X)$

$$\begin{aligned}
\langle Ag, h \rangle &= \int_{\mathbf{X}} (Ag)(x) h(x) f_X(x) dx \\
&= \int_{\mathbf{X}} \left[ \int_{\mathbf{X}} g(x') a(x'|x) dx' \right] h(x) f_X(x) dx \\
&= \int_{\mathbf{X}} \int_{\mathbf{X}} g(x') h(x) a(x'|x) f_X(x) dx dx' \\
&= \int_{\mathbf{X}} g(x') \left[ \int_{\mathbf{X}} h(x) a(x|x') dx \right] f_X(x') dx' \\
&= \int_{\mathbf{X}} g(x') (Ah)(x') f_X(x') dx' \\
&= \langle g, Ah \rangle .
\end{aligned}$$

The norm of the operator  $A$  is defined as

$$\|A\| = \sup_{g \in L_0^2(f_X), \|g\|=1} \|Ag\| .$$

Obviously,  $\|A\| \geq 0$ . In fact,  $\|A\| \in [0, 1]$ . Indeed,  $\|Ag\|^2 = \int_{\mathbf{X}} [(Ag)(x)]^2 f_X(x) dx$  and the calculations above imply that  $\|Ag\|^2 \leq \|g\|^2$ . The quantity  $\|A\|$  is closely related to the convergence properties of the Markov chain  $\Psi$ . For example,  $\Psi$  is geometrically ergodic if and only if  $\|A\| < 1$  (Roberts and Rosenthal, 1997; Roberts and Tweedie, 2001). Loosely speaking, the closer  $\|A\|$  is to 0, the faster  $\Psi$  converges to its stationary distribution (see, e.g., Liu et al., 1995). Because of this, Monte Carlo Markov chains are sometimes ordered according to their operator norms. In particular, if there are two different chains available that are both reversible with respect to  $f_X$ , we prefer the one with the smaller operator norm (see, e.g., Liu et al., 1994; Liu and Wu, 1999; Meng and van Dyk, 1999). In the next subsection, we compare DA and PX-DA in terms of operator norms as well as performance in the CLT.

### 1.4.3 A theoretical comparison of the DA and PX-DA algorithms

The Mtd of the PX-DA algorithm is given by

$$k_w(x'|x) = \int_{\mathbf{Y}} f_{X|Y}^{(w)}(x'|y) f_{Y|X}^{(w)}(y|x) dy .$$

However, there is an alternative representation of  $k_w$  that is based on the general three-step procedure for simulating the PX-DA algorithm that was given in Subsection 1.4.1. This representation turns out to be much more useful for comparing DA and PX-DA. Recall that Step 2 of the three-step procedure entails making the transition  $y \rightarrow y'$  by drawing  $Y'$  from a distribution that depends on  $y$ . Hence, this step can be viewed as performing a single iteration of a Markov chain whose state space is  $\mathbf{Y}$ . If we denote the corresponding Mtd as  $l_w(y'|y)$ , then we can re-express the Mtd of the PX-DA algorithm as

$$k_w(x'|x) = \int_{\mathbf{Y}} \int_{\mathbf{Y}} f_{X|Y}(x'|y') l_w(y'|y) f_{Y|X}(y|x) dy dy' . \quad (1.4.5)$$

Liu and Wu's (1999) Theorem 1 implies that  $f_Y$  is an invariant density for  $l_w$ ; that is,

$$\int_{\mathbf{Y}} l_w(y'|y) f_Y(y) dy = f_Y(y') .$$

This invariance implies that  $f_X$  is an invariant density for  $k_w(x'|x)$ :

$$\begin{aligned} \int_{\mathbf{X}} k_w(x'|x) f_X(x) dx &= \int_{\mathbf{X}} \left[ \int_{\mathbf{Y}} \int_{\mathbf{Y}} f_{X|Y}(x'|y') l_w(y'|y) f_{Y|X}(y|x) dy dy' \right] f_X(x) dx \\ &= \int_{\mathbf{Y}} f_{X|Y}(x'|y') \left[ \int_{\mathbf{Y}} l_w(y'|y) f_Y(y) dy \right] dy' \\ &= \int_{\mathbf{Y}} f_{X|Y}(x'|y') f_Y(y') dy' \\ &= f_X(x') . \end{aligned}$$

Of course, we did not need (1.4.5) to conclude that  $f_X$  is invariant for  $k_w(x'|x)$ . Indeed, the fact that  $k_w(x'|x)$  is the Mtd of a DA algorithm implies that  $k_w$  is reversible with respect to  $f_X$ , and hence that  $f_X$  is invariant for  $k_w$ . Note, however, that the previous calculation still goes through if  $l_w$  is replaced by *any* Mtd having  $f_Y$  as an invariant density. This suggests

a generalization of (1.4.5).

Let  $l : \mathsf{Y} \times \mathsf{Y} \rightarrow [0, \infty)$  be any Mtd that has  $f_Y(y)$  as an invariant density. Define the function  $k_l : \mathsf{X} \times \mathsf{X} \rightarrow [0, \infty)$  as follows

$$k_l(x'|x) = \int_{\mathsf{Y}} \int_{\mathsf{Y}} f_{X|Y}(x'|y') l(y'|y) f_{Y|X}(y|x) dy dy' . \quad (1.4.6)$$

The reader can easily verify that, for each fixed  $x \in \mathsf{X}$ ,  $\int_{\mathsf{X}} k_l(x'|x) dx' = 1$ . Hence,  $k_l$  is an Mtd that defines a Markov chain on  $\mathsf{X}$ , and the arguments above show that  $f_X$  is an invariant density for  $k_l$ . This is a generalization of (1.4.5) in the sense that the set of Mtds having  $f_Y$  as an invariant density is much larger than the set of Mtds of the form  $l_w$ . Hobert and Marchev (2008) studied  $k_l$  and established that (under weak regularity conditions) the MCMC algorithm based on  $k_l$  is better (in terms of convergence rate and ARE) than the DA algorithm. This leads to the conclusion that every PX-DA algorithm is better than the DA algorithm upon which it is based. In order to state the results precisely, we need a couple of definitions.

If there exists a joint pdf  $f^*(x, y)$  with  $\int_{\mathsf{Y}} f^*(x, y) dy = f_X(x)$  such that

$$k_l(x'|x) = \int_{\mathsf{Y}} f_{X|Y}^*(x'|y) f_{Y|X}^*(y|x) dy ,$$

then we say that  $k_l$  is *representable*. Clearly, if  $k_l$  is representable, then it is also reversible with respect to  $f_X(x)$ . (Note that, by definition,  $k_w$  is representable with  $f^{(w)}(x, y)$  playing the role of  $f^*(x, y)$ .) The second definition involves the CLT discussed in Subsection 1.2.4. Let  $X = \{X_n\}_{n=0}^{\infty}$  denote the Markov chain underlying the original DA algorithm based on  $f(x, y)$ . Suppose  $g \in L^2(f_X)$  and, as before, let  $\bar{g}_n = \frac{1}{n} \sum_{i=0}^{n-1} g(X_i)$ . If  $\bar{g}_n$  satisfies a CLT, then let  $\kappa_g^2$  denote the corresponding asymptotic variance. If there is no CLT for  $\bar{g}_n$ , then set  $\kappa_g^2$  equal to  $\infty$ . (Since we have not assumed that  $X$  is geometrically ergodic, a CLT for  $\bar{g}_n$  may or may not exist.) Now let  $X^* = \{X_n^*\}_{n=0}^{\infty}$  denote the Markov chain associated with  $k_l(x'|x)$ , and define  $\kappa_g^{*2}$  analogously using  $\bar{g}_n^* = \frac{1}{n} \sum_{i=0}^{n-1} g(X_i^*)$  in place of  $\bar{g}_n$ . If  $\kappa_g^{*2} \leq \kappa_g^2$  for every  $g \in L^2(f_X)$ , then we say that  $k_l$  is *more efficient than*  $k$ .

Hobert and Marchev (2008) established two general results that facilitate comparison

of the DA algorithm and the MCMC algorithm based on  $k_l$ : (i) if  $k_l$  is reversible with respect to  $f_X$ , then  $k_l$  is more efficient than  $k$ , and (ii) if  $k_l$  is representable, then  $\|K_l\| \leq \|K\|$ , where  $K_l$  and  $K$  are the operators on  $L_0^2(f_X)$  associated with  $k_l$  and  $k$ , respectively. (Hobert and Rosenthal (2007) show that, in (ii), representability can be replaced by a weaker condition at no expense.) Now, consider the implications of these results with regard to the PX-DA algorithm. Since  $k_w$  is representable, both of Hobert and Marchev's (2008) results are applicable and we may conclude that every PX-DA algorithm is better than the corresponding DA algorithm in terms of both convergence rate and ARE. (The norm comparison result was actually established in Liu and Wu (1999) and Meng and van Dyk (1999) using different techniques.)

In addition to providing information about the relative convergence rates of  $X$  and  $X^*$ , the inequality  $\|K_l\| \leq \|K\|$  also has a nice practical application. We know from Subsection 1.4.2 that a reversible Markov chain is geometrically ergodic if and only if the norm of the corresponding operator is strictly less than 1. Therefore, if we can prove that the DA Markov chain,  $X$ , is geometrically ergodic (by, say, establishing a geometric drift condition), then it follows that  $\|K_l\| \leq \|K\| < 1$ , which implies that  $X^*$  is also geometrically ergodic. This allows one to prove that  $X^*$  is geometric without having to work directly with  $k_l$ , which, from an analytical standpoint, is much more cumbersome than  $k$ .

It is important to keep in mind that the comparison results described above are really only useful in situations where at least one of the two chains being compared is known to be geometrically ergodic. For example, if all we know is that  $\|K_l\| \leq \|K\|$ , then it may be the case that  $X$  and  $X^*$  are both bad chains with norm 1 and neither should be used in practice. Similarly, if there are no CLTs, then the fact that  $k_l$  is more efficient than  $k$  isn't very useful.

Finally, there is one very simple sufficient condition for  $k_l$  to be reversible with respect to  $f_X$  and that is the reversibility of  $l(y'|y)$  with respect to  $f_Y(y)$ . Indeed, suppose that

$l(y'|y) f_Y(y)$  is symmetric in  $(y, y')$ . Then

$$\begin{aligned}
k_l(x'|x) f_X(x) &= f_X(x) \int_{\mathcal{Y}} \int_{\mathcal{Y}} f_{X|Y}(x'|y') l(y'|y) f_{Y|X}(y|x) dy dy' \\
&= \int_{\mathcal{Y}} \int_{\mathcal{Y}} f_{X|Y}(x'|y') l(y'|y) f(x, y) dy dy' \\
&= \int_{\mathcal{Y}} \int_{\mathcal{Y}} f_{X|Y}(x'|y') l(y'|y) f_Y(y) f_{X|Y}(x|y) dy dy' \\
&= \int_{\mathcal{Y}} \int_{\mathcal{Y}} f_{X|Y}(x'|y') l(y|y') f_Y(y') f_{X|Y}(x|y) dy dy' \\
&= \int_{\mathcal{Y}} \int_{\mathcal{Y}} f(x', y') l(y|y') f_{X|Y}(x|y) dy dy' \\
&= f_X(x') \int_{\mathcal{Y}} \int_{\mathcal{Y}} f_{X|Y}(x|y) l(y|y') f_{Y|X}(y'|x') dy dy' \\
&= k_l(x|x') f_X(x').
\end{aligned}$$

There is also a simple sufficient condition on  $l(y'|y)$  for representability of  $k_l$  (see Hobert and Marchev, 2008).

We know that each pdf  $w(g)$  yields its own PX-DA algorithm. In the next subsection, we show that, under certain conditions, there is a limiting version of the PX-DA algorithm that beats all the others.

#### 1.4.4 Is there a best PX-DA algorithm?

The results in the previous subsection show that every PX-DA algorithm is better than the original DA algorithm based on  $f(x, y)$ . The question then becomes, is there a particular PX-DA algorithm that beats all the others? There are actually theoretical arguments as well as empirical evidence suggesting that the PX-DA algorithm will perform better as the pdf  $w(\cdot)$  becomes more “diffuse” (Liu and Wu, 1999; Meng and van Dyk, 1999; van Dyk and Meng, 2001). On the other hand, it is clear that our development of the PX-DA algorithm breaks down if  $w$  is improper. In particular, if  $w$  is improper, then (1.4.2) is no longer a pdf. Moreover, Step 2 of the PX-DA algorithm requires a draw from  $w$ , which is obviously not possible when  $w$  is improper. However, Liu and Wu (1999) showed that, if there is a certain

group structure present in the problem, then it is possible to construct a valid PX-DA-like algorithm using an improper *Haar density* in place of  $w$ . Moreover, the results from the previous subsection can be used to show that this *Haar PX-DA algorithm* is better than any PX-DA algorithm based on a proper  $w$ .

Suppose that the set  $\mathbf{G}$  is a topological group; that is, a group such that the functions  $(g_1, g_2) \mapsto g_1 g_2$  and  $g \mapsto g^{-1}$  are continuous. (An example of such a group is the *multiplicative group*,  $\mathbb{R}_+$ , where the binary operation defining the group is multiplication, the identity element is 1, and  $g^{-1} = 1/g$ .) Let  $e$  denote the group's identity element and assume that  $t_e(y) = y$  for all  $y \in \mathbf{Y}$  and that  $t_{g_1 g_2}(y) = t_{g_1}(t_{g_2}(y))$  for all  $g_1, g_2 \in \mathbf{G}$  and all  $y \in \mathbf{Y}$ . In other words, we are assuming that  $t_g(y)$  represents  $\mathbf{G}$  acting topologically on the left of  $\mathbf{Y}$  (Eaton, 1989, Chapter 2).

A function  $\chi : \mathbf{G} \rightarrow \mathbb{R}_+$  is called a *multiplier* if  $\chi$  is continuous and  $\chi(g_1 g_2) = \chi(g_1) \chi(g_2)$  for all  $g_1, g_2 \in \mathbf{G}$ . Assume that Lebesgue measure on  $\mathbf{Y}$  is *relatively (left) invariant* with multiplier  $\chi$ ; that is, assume that for any  $g \in \mathbf{G}$  and any integrable function  $h : \mathbf{Y} \rightarrow \mathbb{R}$ , we have

$$\chi(g) \int_{\mathbf{Y}} h(t_g(y)) dy = \int_{\mathbf{Y}} h(y) dy .$$

Here is a simple example.

EXAMPLE 5 CONT. Again, take  $\mathbf{G} = \mathbb{R}_+$  and  $t_g(y) = gy = (gy_1, \dots, gy_m)$ . Now think of  $\mathbf{G} = \mathbb{R}_+$  as the multiplicative group and note that, for any  $y \in \mathbb{R}_+^m$  and any  $g_1, g_2 \in \mathbf{G}$ , we have  $t_e(y) = y$  and

$$t_{g_1 g_2}(y) = g_1 g_2 y = g_1(g_2 y) = t_{g_1}(t_{g_2}(y)) .$$

Hence, the compatibility conditions are satisfied. Now, for any  $g \in \mathbf{G}$ , we have

$$\int_{\mathbf{Y}} h(t_g(y)) dy = \int_{\mathbb{R}_+^m} h(gy) dy = g^{-m} \int_{\mathbb{R}_+^m} h(y) dy ,$$

which shows that Lebesgue measure on  $\mathbf{Y} = \mathbb{R}_+^m$  is relatively invariant with multiplier  $\chi(g) = g^m$ .

Suppose the group  $\mathbf{G}$  has a *left-Haar measure* of the form  $\nu_l(g) dg$  where  $dg$  denotes Lebesgue measure on  $\mathbf{G}$ . Left-Haar measure satisfies

$$\int_{\mathbf{G}} h(\tilde{g}g) \nu_l(g) dg = \int_{\mathbf{G}} h(g) \nu_l(g) dg \quad (1.4.7)$$

for all  $\tilde{g} \in \mathbf{G}$  and all integrable functions  $h : \mathbf{G} \rightarrow \mathbb{R}$ . In most applications, this measure will be improper; that is,  $\int_{\mathbf{G}} \nu_l(g) dg = \infty$ . (When the left-Haar measure is the same as the right-Haar measure, which satisfies the obvious analogue of (1.4.7), the group is called *unimodular*.) Finally, assume that

$$q(y) := \int_{\mathbf{G}} f_Y(t_g(y)) \chi(g) \nu_l(g) dg$$

is strictly positive for all  $y \in \mathbf{Y}$  and finite for (almost) all  $y \in \mathbf{Y}$ .

We now state (a generalized version of) Liu and Wu's (1999) Haar PX-DA algorithm. If the current state is  $X_n^* = x$ , we simulate  $X_{n+1}^*$  as follows.

One iteration of the Haar PX-DA Algorithm:

1. Draw  $Y \sim f_{Y|X}(\cdot|x)$ , and call the observed value  $y$ .
2. Draw  $G$  from the density proportional to  $f_Y(t_g(y)) \chi(g) \nu_l(g)$ , call the result  $g$ , and set  $y' = t_g(y)$ .
3. Draw  $X_{n+1}^* \sim f_{X|Y}(\cdot|y')$ .

This algorithm *is not a PX-DA algorithm*, but its Mtd does take the form (1.4.6). Indeed, if we let  $l_H(y'|y)$  denote the Mtd of the Markov chain on  $\mathbf{Y}$  that is simulated at Step 2, then we can write the Mtd of the Haar PX-DA algorithm as follows

$$k_H(x'|x) = \int_{\mathbf{Y}} \int_{\mathbf{Y}} f_{X|Y}(x'|y') l_H(y'|y) f_{Y|X}(y|x) dy dy' .$$

Hobert and Marchev (2008) show that  $l_H(y'|y)$  is reversible with respect to  $f_Y$ , which, of course, implies that  $f_Y$  is an invariant density for  $l_H(y'|y)$ . Moreover, these authors also prove that  $k_H$  is representable. Hence, the comparison results from the previous subsection are applicable and imply that the Haar PX-DA algorithm is better than the DA algorithm in terms of both convergence rate and ARE. However, what we really want to compare is Haar PX-DA and PX-DA, and this is the subject of the remainder of this section.

Hobert and Marchev (2008) show that, for any fixed proper pdf  $w(\cdot)$ ,  $k_H$  can be re-expressed as

$$k_H(x'|x) = \int_{\mathbf{Y}} \int_{\mathbf{Y}} f_{X|Y}^{(w)}(x'|y') l^{(w)}(y'|y) f_{Y|X}^{(w)}(y|x) dy dy' , \quad (1.4.8)$$

where  $f_{X|Y}^{(w)}$  and  $f_{Y|X}^{(w)}$  are as defined in Subsection 1.4.1, and  $l^{(w)}(y'|y)$  is an Mtd on  $\mathbf{Y}$  that is reversible with respect to  $f_Y^{(w)}(y) := \int_{\mathbf{Y}} f^{(w)}(x, y) dx$ . Now consider the significance of equation (1.4.8) in the context of the results of Subsection 1.4.3. In particular, we know that the PX-DA algorithm driven by  $f^{(w)}(x, y)$  is itself a DA algorithm, and (1.4.8) shows that  $k_H$  is related to  $k_w$  in exactly the same way that  $k_l$  is related to  $k$ . Therefore, since  $k_H$  is representable, we may appeal to the comparison results once more to conclude that Haar PX-DA is better than every PX-DA algorithm in terms of both convergence rate and ARE.

Finally, note that Step 2 of the Haar PX-DA algorithm involves only one draw from a density on  $\mathbf{G}$ , whereas the regular PX-DA algorithm calls for two such draws in its Step 2. Thus, from a computational standpoint, the Haar PX-DA algorithm is actually simpler than the PX-DA algorithm. We conclude with an application to the probit example.

EXAMPLE 5 CONT. Recall that  $\mathbf{G}$  is the multiplicative group,  $\mathbb{R}_+$ , and  $t_g(y) = gy = (gy_1, \dots, gy_m)$ . Note that, for any  $\tilde{g} \in \mathbf{G}$ , we have

$$\int_0^\infty h(\tilde{g}g) \frac{1}{g} dg = \int_0^\infty h(g) \frac{1}{g} dg ,$$

which shows that  $\frac{dg}{g}$  is a left-Haar measure for the multiplicative group. (This group is

actually abelian and hence unimodular.) Thus,

$$\pi(t_g(y)|z) \chi(g) \nu_l(g) \propto g^{m-1} \exp \left\{ -g^2 \left[ \frac{y^T(I-H)y}{2} \right] \right\} I_{\mathbb{R}_+}(g),$$

and it follows that

$$q(y) \propto \int_0^\infty g^{m-1} \exp \left\{ -g^2 \left[ \frac{y^T(I-H)y}{2} \right] \right\} dg = \frac{2^{\frac{m}{2}} \Gamma(\frac{m}{2})}{[y^T(I-H)y]^{\frac{m}{2}}},$$

which is clearly positive for all  $y \in \mathbf{Y}$  and finite for (almost) all  $y \in \mathbf{Y}$ . We can now write down the Haar PX-DA algorithm. Given the current state,  $X_n^* = \beta$ , we simulate the next state,  $X_{n+1}^*$ , by performing the following three steps:

1. Draw  $Y_1, \dots, Y_m$  independently such that  $Y_i \sim \text{TN}(v_i^T \beta, 1, z_i)$ , and call the result  $y = (y_1, \dots, y_m)^T$ .

2. Draw

$$V \sim \text{Gamma} \left( \frac{m}{2}, \frac{y^T(I-H)y}{2} \right),$$

call the result  $v$ , and set  $y' = \sqrt{v}y$ .

3. Draw  $X_{n+1}^* \sim \text{N}(\hat{\beta}(y'), (V^T V)^{-1})$ .

In Subsection 1.4.1, we developed a family of PX-DA algorithms for this problem, one for each  $(\alpha, \delta) \in \mathbb{R}_+ \times \mathbb{R}_+$ . The results in Subsection 1.4.3 imply that every member of that family is better than the original DA algorithm based on  $f(x, y)$ . Moreover, the results described in this subsection show that the Haar PX-DA algorithm above is better than every member of that family of PX-DA algorithms.

Roy and Hobert (2007) proved that this Haar PX-DA algorithm is geometrically ergodic by establishing that the much simpler DA algorithm of Albert and Chib (1993) is geometrically ergodic, and then appealing to the fact that  $\|K_H\| \leq \|K\|$ . These authors also provided substantial empirical evidence suggesting that the ARE of the Haar PX-DA estimator with respect to the DA estimator is often much larger than 1.

## **Acknowledgments**

The author is grateful to Trung Ha, Galin Jones, Aixin Tan and an anonymous referee for helpful comments and suggestions. This work was supported by NSF Grants DMS-05-03648 and DMS-08-05860.

# Bibliography

- Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679.
- Chan, K. S. and Geyer, C. J. (1994). Comment on “Markov chains for exploring posterior distributions” by L. Tierney. *The Annals of Statistics*, 22:1747–1758.
- Chen, M.-H. and Shao, Q.-M. (2000). Propriety of posterior distribution for dichotomous quantal response models. *Proceedings of the American Mathematical Society*, 129:293–302.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38.
- Diaconis, P., Khare, K., and Saloff-Coste, L. (2008). Gibbs sampling, exponential families and orthogonal polynomials (with discussion). *Statistical Science*, 23:151–200.
- Diaconis, P. and Stroock, D. (1991). Geometric bounds for eigenvalues of Markov chains. *The Annals of Applied Probability*, 1:36–61.
- Eaton, M. L. (1989). *Group Invariance Applications in Statistics*. Institute of Mathematical Statistics and the American Statistical Association, Hayward, California and Alexandria, Virginia.
- Fernández, C. and Steel, M. F. J. (1999). Multivariate Student- $t$  regression models: Pitfalls and inference. *Biometrika*, 86:153–167.
- Fill, J. A., Machida, M., Murdoch, D. J., and Rosenthal, J. S. (2000). Extension of Fill’s

- perfect rejection sampling algorithm to general chains. *Random Structures and Algorithms*, 17:290–316.
- Flegal, J. M., Haran, M., and Jones, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science*, 23:250–260.
- Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science*, 7:473–511.
- Hobert, J. P., Jones, G. L., Presnell, B., and Rosenthal, J. S. (2002). On the applicability of regenerative simulation in Markov chain Monte Carlo. *Biometrika*, 89:731–743.
- Hobert, J. P. and Marchev, D. (2008). A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms. *The Annals of Statistics*, 36:532–554.
- Hobert, J. P. and Rosenthal, J. S. (2007). Norm comparisons for data augmentation. *Advances and Applications in Statistics*, 7:291–302.
- Hobert, J. P., Tan, A., and Liu, R. (2007). When is Eaton’s Markov chain irreducible? *Bernoulli*, 13:641–652.
- Jones, G. L., Haran, M., Caffo, B. S., and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 101:1537–1547.
- Jones, G. L. and Hobert, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, 16:312–34.
- Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to comparisons of estimators and augmentation schemes. *Biometrika*, 81:27–40.
- Liu, J. S., Wong, W. H., and Kong, A. (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *Journal of the Royal Statistical Society, Series B*, 57:157–169.
- Liu, J. S. and Wu, Y. N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94:1264–1274.

- Marchev, D. and Hobert, J. P. (2004). Geometric ergodicity of van Dyk and Meng's algorithm for the multivariate student's  $t$  model. *Journal of the American Statistical Association*, 99:228–238.
- Meng, X.-L. and van Dyk, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, 86:301–320.
- Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, London.
- Meyn, S. P. and Tweedie, R. L. (1994). Computable bounds for geometric convergence rates of Markov chains. *The Annals of Applied Probability*, 4:981–1011.
- Mira, A. and Geyer, C. J. (1999). Ordering Monte Carlo Markov chains. Technical Report No. 632, School of Statistics, University of Minnesota.
- Mykland, P., Tierney, L., and Yu, B. (1995). Regeneration in Markov chain samplers. *Journal of the American Statistical Association*, 90:233–41.
- Neal, R. M. (2003). Slice sampling (with discussion). *The Annals of Statistics*, pages 705–767.
- Nummelin, E. (1984). *General Irreducible Markov Chains and Non-negative Operators*. Cambridge, London.
- R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing*, 5:121–125.
- Roberts, G. and Smith, A. F. M. (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and their Applications*, 49:207–216.
- Roberts, G. O. and Rosenthal, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, 2:13–25.
- Roberts, G. O. and Rosenthal, J. S. (2001). Markov chains and de-initializing processes. *Scandinavian Journal of Statistics*, 28:489–504.

- Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71.
- Roberts, G. O. and Rosenthal, J. S. (2006). Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains. *The Annals of Applied Probability*, 16:2123–2139.
- Roberts, G. O. and Tweedie, R. L. (2001). Geometric  $L^2$  and  $L^1$  convergence are equivalent for reversible Markov chains. *Journal of Applied Probability*, 38A:37–41.
- Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90:558–566.
- Ross, S. M. (1996). *Stochastic Processes*. John Wiley and Sons, New York, 2nd edition edition.
- Roy, V. and Hobert, J. P. (2007). Convergence rates and asymptotic standard errors for Markov chain Monte Carlo algorithms for Bayesian probit regression. *Journal of the Royal Statistical Society, Series B*, 69:607–623.
- Rudin, W. (1991). *Functional Analysis*. McGraw-Hill, New York, 2nd edition.
- Swendsen, R. H. and Wang, J.-S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58:86–88.
- Tan, A. (2008). *Analysis of Markov chain Monte Carlo algorithms for random effects models*. PhD thesis, Department of Statistics, University of Florida.
- Tan, A. and Hobert, J. P. (2008). Block Gibbs sampling for Bayesian random effects models with improper priors: Convergence and regeneration. Technical report, University of Florida.
- Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82:528–550.
- Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics*, 22:1701–1762.

- van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation (with discussion).  
*Journal of Computational and Graphical Statistics*, 10:1–50.