

# Discussion of Yu & Meng's Paper

James P. Hobert and Jorge Carlos Román

Department of Statistics

University of Florida

December 2010

We begin by congratulating Professors Yu and Meng on an outstanding paper, and thanking Professor Levine for giving us the opportunity to discuss their work. Our discussion focuses mainly on the GIS & ASIS algorithms. Section 1 concerns the relationship between the GIS and sandwich algorithms. In Section 2, we consider a family of toy GIS algorithms based on the bivariate normal distribution, and show how this family is related to the toy example in Section 2 of Yu & Meng (2011) (hereafter Y&M). Finally, in Section 3, we provide a simple example of a non-reversible GIS algorithm.

## 1 $\{\mathbf{DA\ algorithms}\} \subset \{\mathbf{sandwich\ algorithms}\} \subset \{\mathbf{GIS\ algorithms}\}$

Let  $f_X : \mathsf{X} \rightarrow [0, \infty)$  be an intractable target density, and suppose that  $f : \mathsf{X} \times \mathsf{Y} \rightarrow [0, \infty)$  is a joint density whose  $x$ -marginal is the target; i.e.,  $\int_{\mathsf{Y}} f(x, y) dy = f_X(x)$ . If straightforward sampling from the associated conditional densities is possible, then we can use the data augmentation (DA) algorithm to explore  $f_X$ . Of course, running the algorithm entails alternating between draws from  $f_{Y|X}$  and  $f_{X|Y}$ , which simulates the Markov chain whose Markov transition density (Mtd) is

$$k_{\text{DA}}(x'|x) = \int_{\mathsf{Y}} f_{X|Y}(x'|y) f_{Y|X}(y|x) dy .$$

If we denote the DA Markov chain by  $\{X_n\}_{n=0}^{\infty}$ , then  $k_{\text{DA}}(\cdot|x)$  is simply the conditional density of  $X_{n+1}$  given that  $X_n = x$ . It's easy to see that  $k_{\text{DA}}(x'|x) f_X(x)$  is symmetric in  $(x, x')$ , so the DA Markov chain is reversible. We assume throughout that all Markov chains on the target space,  $\mathsf{X}$ , satisfy the usual regularity conditions: Harris recurrence, irreducibility and aperiodicity.

Following Liu and Wu (1999), Meng and van Dyk (1999) and van Dyk and Meng (2001), Hobert and Marchev (2008) introduced an alternative to DA that employs an extra move on the  $Y$  space that is “sandwiched” between the two conditional draws. Define  $f_Y(y) = \int_X f(x, y) dx$  and suppose that  $R(y, dy')$  is *any* Markov transition function (Mtf) on  $Y$  that is reversible with respect to  $f_Y$ ; i.e.,  $R(y, dy')f_Y(y)dy = R(y', dy)f_Y(y')dy'$ . The *sandwich algorithm* simulates the Markov chain whose Mtd is

$$k_S(x'|x) = \int_Y \int_Y f_{X|Y}(x'|y') R(y, dy') f_{Y|X}(y|x) dy .$$

Again, it’s easy to see that  $k_S(x'|x) f_X(x)$  is symmetric in  $(x, x')$ . To run the sandwich algorithm, we simply run the DA algorithm as usual, except that after each  $y$  is drawn, we perform the extra step  $y' \sim R(y, \cdot)$  before drawing the new  $x$ . In practice,  $R$  is usually chosen so that, for fixed  $y$ , the chain driven by  $R(y, \cdot)$  lives in a one-dimensional subspace of  $Y$ . Consequently, drawing from  $R$  is much less expensive (computationally) than drawing from the conditional densities. Note that the sandwich algorithm reduces to DA if we take  $R$  to be the trivial Mtf whose chain is absorbed at the starting point.

Here is our interpretation of Y&M’s GIS algorithm. Suppose that, in addition to  $f(x, y)$ , we have another joint density  $\tilde{f} : X \times Y \rightarrow [0, \infty)$  for which  $\int_Y \tilde{f}(x, y) dy = f_X(x)$ . Now let  $Q(y, dy')$  be *any* Mtf on the space  $Y$  that satisfies

$$\int_Y Q(y, dy') f_Y(y) dy = \tilde{f}_Y(y') dy' , \quad (1)$$

where  $\tilde{f}_Y(y) = \int_X \tilde{f}(x, y) dx$ . Note that  $\tilde{f}_Y(y)$  need not be the same as  $f_Y(y)$ . Y&M’s GIS algorithm simulates the Markov chain whose Mtd is

$$k_{YM}(x'|x) = \int_Y \int_Y \tilde{f}_{X|Y}(x'|y') Q(y, dy') f_{Y|X}(y|x) dy . \quad (2)$$

This chain is not necessarily reversible (see Section 3), but  $f_X(x)$  is the invariant density. As with the sandwich algorithm, we allow the Markov chain defined by  $Q$  to be reducible, as long as the GIS chain itself is well behaved. We can see that the sandwich chain is a special case of the GIS chain by taking  $\tilde{f}(x, y) = f(x, y)$  and  $Q(y, dy') = R(y, dy')$ .

Here is a simple example of a GIS algorithm. Given  $f(x, y)$  and  $\tilde{f}(x, y)$ , let  $Q_A(y, dy') = q_A(y'|y) dy'$ , where the Mtd  $q_A$  is defined as

$$q_A(y'|y) = \int_X \tilde{f}_{Y|X}(y'|x) f_{X|Y}(x|y) dx . \quad (3)$$

Clearly,  $\int_Y q_A(y'|y) f_Y(y) dy = \tilde{f}_Y(y')$ , so condition (1) is satisfied. Now, plugging into (2), we see that the resulting GIS algorithm has Mtd given by

$$k_A(x'|x) = \int_Y \int_X \int_Y \tilde{f}_{X|Y}(x'|y') \tilde{f}_{Y|X}(y'|x'') f_{X|Y}(x''|y) f_{Y|X}(y|x) dy dx'' dy' ,$$

which is exactly the Mtd associated with the *alternating scheme* defined by Y&M's equation (2.12). Hence, the alternating scheme is a special case of GIS.

It seems that our definition of the GIS algorithm is slightly more general than that of Y&M. Suppose that  $f(x, y)$  and  $\tilde{f}(x, y)$  are fixed. In contrast with our version of GIS, in which any  $Q$  satisfying (1) will do, Y&M restrict attention to those  $Q$ s that stem from joint distributions on  $X \times Y \times Y$  that are consistent with  $f(x, y)$  and  $\tilde{f}(x, y)$ . For example, suppose that  $g : X \times Y \times Y \rightarrow [0, \infty)$  is a joint density such that

$$\int_Y g(x, y, y') dy' = f(x, y) \quad \text{and} \quad \int_Y g(x, y', y) dy' = \tilde{f}(x, y) . \quad (4)$$

Then take  $Q(y, dy') = q(y'|y) dy'$  where

$$q(y'|y) = \frac{\int_X g(x, y, y') dx}{f_Y(y)} . \quad (5)$$

It's easy to see that this  $Q$  satisfies condition (1). An example of a joint density that satisfies (4) is  $g(x, y, y') = f(x, y)\tilde{f}(x, y')/f_X(x)$ . Interestingly, applying (5) with this particular  $g$  yields  $q_A$ .

On the other hand, given  $f(x, y)$ ,  $\tilde{f}(x, y)$  and an arbitrary  $Q$  that satisfies (1), there does not necessarily exist a joint distribution that is consistent with all three of these. Hence, the class of  $Q$ s considered by Y&M is a strict subset of the  $Q$ s satisfying (1). However, some of Y&M's theoretical results concerning the GIS algorithm still hold when the set of allowable  $Q$ s is expanded to include all those satisfying (1). An example is given below.

We now discuss relationships among the convergence rates of the Markov chains that we have defined. As in Y&M's Section 5.1, let  $L_0^2(f_X)$  denote the set of real-valued functions with domain  $X$  that are square integrable and have mean zero with respect to  $f_X$ . Suppose that  $k(x'|x)$  is a (generic) Mtd satisfying  $\int_X k(x'|x) f_X(x) dx = f_X(x')$ . This Mtd defines an operator,  $K : L_0^2(f_X) \rightarrow L_0^2(f_X)$ , that maps  $g \in L_0^2(f_X)$  to  $Kg \in L_0^2(f_X)$  where

$$(Kg)(x) = \int_X g(x') k(x'|x) dx' .$$

Let  $\|K\|$  and  $r(K)$  denote the norm and spectral radius of  $K$ , respectively, which both lie in  $[0, 1]$ . For definitions, see Retherford (1993). As explained in Rosenthal (2003),  $r(K)$  is equal to the

(asymptotic) convergence rate of the Markov chain defined by  $k$ . Since  $r(K) \leq \|K\|$  and small values of  $r(K)$  are associated with fast convergence, the norm provides an upper bound on the “slowness” of the chain. In the special case where the chain is reversible,  $r(K) = \|K\|$  and the chain is geometrically ergodic if and only if  $\|K\| < 1$  (Roberts and Rosenthal, 1997; Roberts and Tweedie, 2001).

Let  $K_{\text{DA}}$  and  $K_{\text{YM}}$  denote the Markov operators defined by  $k_{\text{DA}}$  and  $k_{\text{YM}}$ , respectively. Also, let  $\tilde{K}_{\text{DA}}$  denote the Markov operator associated with the DA chain based on  $\tilde{f}(x, y)$ . Y&M’s Theorem 1 states that

$$r(K_{\text{YM}}) \leq \|K_{\text{YM}}\| \leq \mathcal{R}_{12} \sqrt{\|K_{\text{DA}}\| \|\tilde{K}_{\text{DA}}\|}, \quad (6)$$

where  $\mathcal{R}_{12}$  is the maximal correlation of the pair  $(Y_0, Y_1)$ , whose joint probability distribution is  $Q(y, dy') f_Y(y) dy$ . We note that, in the formal statement of Y&M’s Theorem 1, there is an explicit assumption concerning the existence of a joint distribution that is consistent with  $f(x, y)$  and  $\tilde{f}(x, y)$ , but the proof (in Appendix B) does not seem to rely on this assumption.

The inequality (6) can be used to prove a new result concerning the sandwich algorithm. As mentioned above, we can recover the sandwich chain from the GIS chain by taking  $\tilde{f}(x, y) = f(x, y)$  and  $Q(y, dy') = R(y, dy')$ . In this case, (6) becomes

$$r(K_{\text{S}}) = \|K_{\text{S}}\| \leq \mathcal{R}_{12} \sqrt{\|K_{\text{DA}}\| \|K_{\text{DA}}\|} = \mathcal{R}_{12} \|K_{\text{DA}}\|, \quad (7)$$

which was noted by Y&M (at the top of page 12). Now,  $R(y, dy') f_Y(y) dy$  is actually the joint distribution of the first two steps of the stationary version of the Markov chain driven by  $R$ . Therefore, by Liu, Wong and Kong’s (1994) Lemma 2.3,  $\mathcal{R}_{12} = \|K_{\text{R}}\|$ , where  $K_{\text{R}} : L_0^2(f_Y) \rightarrow L_0^2(f_Y)$  is the Markov operator defined by  $R(y, dy')$ . Hence, (7) becomes

$$r(K_{\text{S}}) = \|K_{\text{S}}\| \leq \|K_{\text{R}}\| \|K_{\text{DA}}\|. \quad (8)$$

The inequality (8) strengthens and generalizes a result of Hobert and Rosenthal (2007) who showed that, if  $K_{\text{S}}$  is a positive operator, then  $\|K_{\text{S}}\| \leq \|K_{\text{DA}}\|$ . The additional factor of  $\|K_{\text{R}}\|$  on the right-hand side of (8) is important because it shows that a sub-geometric DA chain can be transformed into a geometrically ergodic sandwich chain by adding an extra move according to a geometrically ergodic chain on the  $Y$  space. In other words, the left-hand side of (8) will be less than 1 as long as  $\|K_{\text{R}}\| < 1$ , even if  $\|K_{\text{DA}}\| = 1$ .

It should be noted that, in most practical applications of the sandwich algorithm,  $R$  is *idempotent*; that is,

$$\int_Y R(y, dy'') R(y'', dy') = R(y, dy'). \quad (9)$$

Loosely speaking, if  $R$  is idempotent, then  $K_R$  is a projection and  $\|K_R\| = 1$ , so (8) becomes  $\|K_S\| \leq \|K_{DA}\|$ , which is exactly Hobert and Rosenthal's (2007) result (which is applicable here because  $K_S$  is a positive operator when  $R$  is idempotent). See Khare and Hobert (2010) for more on this.

## 2 A Family of GIS Algorithms for a Normal Target

Suppose that the target density,  $f_X(x)$ , is  $N(\mu, \sigma^2)$ , and that  $f(x, y)$  is bivariate normal, denoted by  $BN(\mu, \sigma^2, \theta, \tau^2, \rho)$ . Obviously, the  $x$ -marginal of  $f(x, y)$  is the target. We now specify an  $\tilde{f}$  so that we can construct a family of GIS algorithms. Fix  $c \in \mathbb{R}$ . If  $(U_1, U_2) \sim BN(\mu, \sigma^2, \theta, \tau^2, \rho)$ , then  $(W_1, W_2) = (U_1, U_2 + cU_1)$  is  $BN(\mu, \sigma^2, \tilde{\theta}, \tilde{\tau}^2, \tilde{\rho})$ , where  $\tilde{\theta} = c\mu + \theta$ ,  $\tilde{\tau}^2 = c^2\sigma^2 + \tau^2 + 2c\rho\sigma\tau$ , and

$$\tilde{\rho} = \frac{c\sigma + \rho\tau}{\sqrt{(c\sigma + \rho\tau)^2 + (1 - \rho^2)\tau^2}}.$$

Let  $\tilde{f}(x, y)$  denote this second bivariate normal density, and note that the  $x$ -marginal of  $\tilde{f}$  is still the target. Obviously,  $f_Y(y)$  is  $N(\theta, \tau^2)$  and  $\tilde{f}_Y(y)$  is  $N(\tilde{\theta}, \tilde{\tau}^2)$ . Now, let  $Q(y, dy') = q(y'|y) dy'$  where  $q(\cdot|y)$  is a univariate normal density given by

$$N\left(y + c\mu + c\rho(\sigma/\tau)(y - \theta), c^2\sigma^2(1 - \rho^2)\right).$$

A simple calculation shows that  $\int_{\mathbb{R}} q(y'|y) f_Y(y) dy = \tilde{f}_Y(y')$ , so this  $Q$  satisfies condition (1). We now have all the ingredients that we need to construct a family of GIS algorithms, indexed by  $(\theta, \tau^2, \rho)$  and  $c$ . A direct calculation shows that the resulting Mtd,  $k_{YM}(\cdot|x)$ , is normal with mean

$$\mu + \rho \left[ \frac{(c\sigma + \rho\tau)(\rho c\sigma + \tau)}{(c\sigma + \rho\tau)^2 + (1 - \rho^2)\tau^2} \right] (x - \mu)$$

and variance

$$\sigma^2 - \rho^2\sigma^2 \left[ \frac{(c\sigma + \rho\tau)(\rho c\sigma + \tau)}{(c\sigma + \rho\tau)^2 + (1 - \rho^2)\tau^2} \right]^2.$$

Interestingly,  $k_{\text{YM}}(x'|x) f_X(x)$  is symmetric in  $(x, x')$ , so every member of this family of GIS chains is reversible. Consequently, the convergence rate (spectral radius) is exactly equal to the norm,  $\|K_{\text{YM}}\|$ .

By Liu et al.'s (1994) Lemma 2.3, we know that  $\|K_{\text{YM}}\|$  is equal to the maximal correlation of a random pair with joint density  $k_{\text{YM}}(x'|x) f_X(x)$ . It's easy to show that this joint density is bivariate normal, so the maximal correlation is equal to the absolute value of the correlation in the bivariate normal, and we have

$$\|K_{\text{YM}}\| = \frac{|\rho| |c\sigma + \rho\tau| |\rho c\sigma + \tau|}{(c\sigma + \rho\tau)^2 + (1 - \rho^2)\tau^2}, \quad (10)$$

which is less than one if and only if  $|\rho| < 1$ . It is interesting to compare the exact value in (10) with the upper bound from Y&M's Theorem 1; that is, with the right-hand side of (6). Liu et al.'s (1994) Theorem 3.2 implies that  $\|K_{\text{DA}}\|$  is equal to the square of the maximal correlation of a random pair with joint density  $f(x, y)$ , which is bivariate normal. Thus,  $\|K_{\text{DA}}\| = \rho^2$ . Similarly,  $\|\tilde{K}_{\text{DA}}\| = \tilde{\rho}^2$ . Moreover, the joint distribution  $Q(y, dy') f_Y(y) dy$  is also bivariate normal, and it follows that

$$\mathcal{R}_{12} = \frac{|\rho c\sigma + \tau|}{\sqrt{(c\sigma + \rho\tau)^2 + (1 - \rho^2)\tau^2}}.$$

Therefore,

$$\begin{aligned} \mathcal{R}_{12} \sqrt{\|K_{\text{DA}}\|} \sqrt{\|\tilde{K}_{\text{DA}}\|} &= \frac{|\rho c\sigma + \tau|}{\sqrt{(c\sigma + \rho\tau)^2 + (1 - \rho^2)\tau^2}} \sqrt{\rho^2} \sqrt{\frac{(c\sigma + \rho\tau)^2}{(c\sigma + \rho\tau)^2 + (1 - \rho^2)\tau^2}} \\ &= \frac{|\rho| |c\sigma + \rho\tau| |\rho c\sigma + \tau|}{(c\sigma + \rho\tau)^2 + (1 - \rho^2)\tau^2} \\ &= \|K_{\text{YM}}\|. \end{aligned}$$

So, for every member of this family of GIS algorithms, the upper bound on  $\|K_{\text{YM}}\|$  from Y&M's Theorem 1 is exactly equal to  $\|K_{\text{YM}}\|$ .

When the norm of a Markov operator is zero, the corresponding algorithm is “perfect” in the sense that it produces an iid sample from the target. For example, it's clear that the DA algorithm based on  $f(x, y)$  is perfect when  $\rho = 0$ , since in that case,  $f(x, y) = f_X(x) f_Y(y)$ . Equation (6) shows that the GIS algorithm must be perfect whenever one of the underlying DA algorithms is perfect.

Now consider a specific example in which  $\mu = \theta = 0$ ,  $\sigma^2 = \tau^2 = 1$ , and  $c = -1$ . In this case,  $\|K_{\text{DA}}\| = \rho^2$ ,  $\|\tilde{K}_{\text{DA}}\| = (1 - \rho)/2$ , and  $\|K_{\text{YM}}\| = |\rho|(1 - \rho)/2$ . Figure 1 shows a plot of the convergence rates of the three different algorithms as  $\rho$  ranges between -1 and 1. It is interesting

to note that the first DA algorithm actually converges strictly faster than the GIS algorithm for all  $\rho$  in  $(-1, 1/3) \setminus \{0\}$ . This example serves as a warning that it is possible for a GIS algorithm to converge more slowly than the faster of the two underlying DA algorithms.

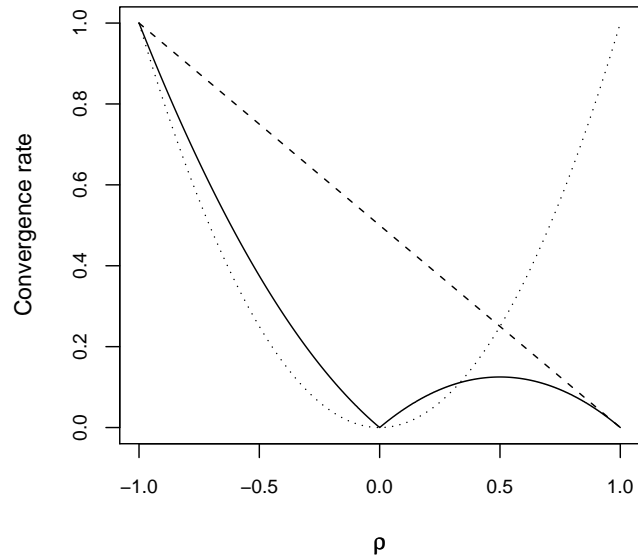


Figure 1: This plot shows how convergence rate varies with  $\rho$  for the DA chain based on  $f$  (dotted line), the DA chain based on  $\tilde{f}$  (dashed line), and the GIS chain (solid line). In this example  $\mu = \theta = 0$ ,  $\sigma^2 = \tau^2 = 1$ , and  $c = -1$ .

## 2.1 Analysis of Y&M's normal hierarchy

Consider the following simple hierarchy

$$\begin{aligned}
 Z|X, Y &\sim N(Y, 1) \\
 Y|X &\sim N(X, V) \\
 X &\sim N(0, A) ,
 \end{aligned}
 \tag{11}$$

where  $V$  and  $A$  are fixed positive numbers. This is the same as Y&M's normal hierarchy (from their Section 2.1), except that we've replaced the flat (Haar) prior by a proper normal prior. (Also,

in order to keep our notation consistent, we're using  $(X, Y, Z)$  in place of Y&M's  $(\theta, Y_{mis}, Y_{obs})$ . As in Y&M, we take the target to be the posterior density of  $X$  given the data,  $z$ . This posterior, which is denoted by  $f_X(x)$ , is

$$\mathbf{N}\left(\frac{Az}{V + A + 1}, \frac{A(V + 1)}{V + A + 1}\right).$$

The joint density of  $(X, Y)$  given  $z$ , which we denote by  $f(x, y)$ , is bivariate normal with  $\mu = Az/(V + A + 1)$ ,  $\sigma^2 = A(V + 1)/(V + A + 1)$ ,

$$\theta = \frac{(A + V)z}{A + V + 1},$$

$$\tau^2 = \frac{A + V}{A + V + 1},$$

and

$$\rho = \frac{\sqrt{A}}{\sqrt{(A + V)(V + 1)}}.$$

Hence, we can employ the GIS algorithm developed earlier in this section. For general  $c$ , the convergence rate is given by

$$\|K_{\text{YM}}\| = \frac{A^2 |1 + V/A + c| |1 + c(V + 1)|}{(V + A) \left[ A(1 + c(V + 1))^2 + V(V + A + 1) \right]}. \quad (12)$$

Despite the fact that we are not using a Haar prior, we can still get perfect GIS algorithms by setting  $c = -(V + 1)^{-1}$  or  $c = -V/A - 1$ . Figure 2 shows a plot of the convergence rate of the GIS algorithm as  $c$  ranges between -5 and 5 when  $A = 3$  and  $V = 2$ . Note that the ASIS algorithm, which corresponds to the value  $c = -1$ , is not perfect. In fact, its convergence rate is 0.1. Perhaps this example could be used to settle the open problem described in Y&M's Section 6 concerning the optimality of AA-SA pairs.

If we replace the proper normal prior on  $X$  in the hierarchical model (11) with a flat (Haar) prior, then the model becomes exactly the one studied in Section 2.1 of Y&M. In that case, the convergence rate becomes

$$\|K_{\text{YM}}\| = \frac{|1 + c| |1 + c(V + 1)|}{(1 + c(V + 1))^2 + V}. \quad (13)$$

(Not surprisingly, (13) is simply the limit of (12) as  $A \rightarrow \infty$ .) As Y&M noted, the ASIS algorithm ( $c = -1$ ) is perfect in this case. Indeed, their Theorem 4 is applicable here because the Haar assumptions are satisfied. Note, however, that one of the GIS algorithms ( $c = -(V + 1)^{-1}$ ) is also

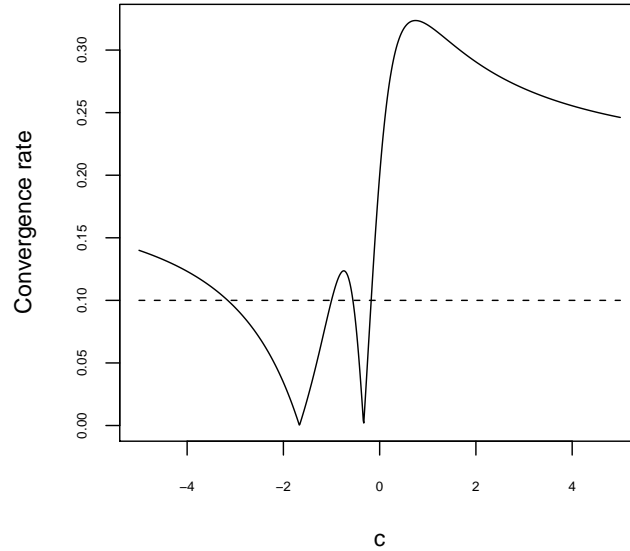


Figure 2: The convergence rate of the GIS algorithm versus  $c$ . The dashed horizontal line at 0.1 is the convergence rate of the ASIS algorithm.

perfect. This shows that, even when a Haar prior is used (and the conditions of Y&M's Theorem 4 are satisfied), there may still exist a GIS algorithm (that is not ASIS) that has the same convergence rate as the optimal ASIS algorithm. Figure 3 shows a plot of the convergence rate of the GIS algorithm as  $c$  ranges between -5 and 5 when  $A = 3$  and  $V = 2$ .

### 3 A Non-reversible GIS Algorithm

Y&M did not provide an example of a non-reversible GIS chain, so we present one here. It was shown in Section 1 of this discussion that the alternating scheme defined by Y&M's equation (2.12) is a GIS algorithm. While this seems like an obvious place to look for an example of a non-reversible GIS algorithm, we decided to go in a different direction. Let  $X = \{1, 2, 3\}$  and take the target mass function to be  $f_X = (0.4 \ 0.3 \ 0.3)^T$ . Take  $Y = X$  and consider the following joint mass function:

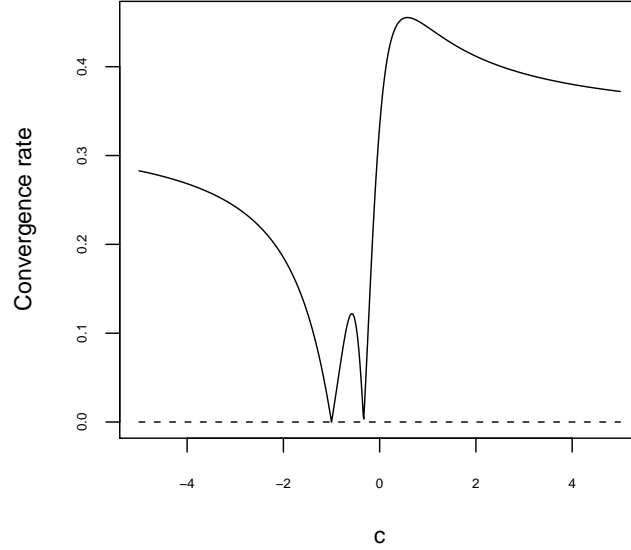


Figure 3: The convergence rate of the GIS algorithm versus  $c$  for the model with a flat (Haar) prior. The dashed horizontal line is at zero, which is the convergence rate of both the ASIS algorithm and the GIS algorithm with  $c = -(V + 1)^{-1}$ .

		Y		
		1	2	3
X	1	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$
	2	$\frac{2}{10}$	$\frac{1}{10}$	0
	3	$\frac{2}{10}$	0	$\frac{1}{10}$

The  $x$ -marginal is clearly  $f_X$ , and the  $y$ -marginal is  $f_Y = (0.6 \ 0.2 \ 0.2)^T$ . The Markov transition matrix (Mtm) of the corresponding DA chain is:

$$K = \begin{pmatrix} \frac{10}{24} & \frac{7}{24} & \frac{7}{24} \\ \frac{7}{18} & \frac{7}{18} & \frac{4}{18} \\ \frac{7}{18} & \frac{4}{18} & \frac{7}{18} \end{pmatrix}.$$

The  $(i, j)$ th entry is the probability that the DA Markov chain moves from  $i$  to  $j$  in one step. Note that  $K$  is reversible with respect to  $f_X$ . Now define a second joint mass function:

		Y		
		1	2	3
X	1	$\frac{3}{10}$	$\frac{1}{10}$	0
	2	0	$\frac{1}{10}$	$\frac{2}{10}$
	3	$\frac{2}{10}$	$\frac{1}{10}$	0

The  $x$ -marginal is again  $f_X$ , but the  $y$ -marginal in this case is  $\tilde{f}_Y = (0.5 \ 0.3 \ 0.2)^T$ . The Mtm of the new DA algorithm is:

$$\tilde{K} = \begin{pmatrix} \frac{32}{60} & \frac{5}{60} & \frac{23}{60} \\ \frac{1}{9} & \frac{7}{9} & \frac{1}{9} \\ \frac{23}{45} & \frac{5}{45} & \frac{17}{45} \end{pmatrix}.$$

Again,  $\tilde{K}$  is reversible with respect to  $f_X$ .

Consider a Mtm on Y given by

$$Q = \begin{pmatrix} \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 0 & 0 \end{pmatrix},$$

and note that  $f_Y^T Q = \tilde{f}_Y^T$ . Thus, (the discrete analogue of) (1) holds, and we can use  $Q$  to construct a GIS algorithm. A simple calculation reveals that the Mtm of this GIS algorithm is

$$K_{YM} = \begin{pmatrix} \frac{49}{120} & \frac{35}{120} & \frac{36}{120} \\ \frac{31}{90} & \frac{35}{90} & \frac{24}{90} \\ \frac{4}{9} & \frac{2}{9} & \frac{3}{9} \end{pmatrix}.$$

Clearly,  $K_{YM}$  is not reversible with respect to  $f_X$ , since, for example,  $\frac{35}{120} \times \frac{4}{9} \neq \frac{31}{90} \times \frac{3}{9}$ . However, it is true that  $f_X^T K_{YM} = f_X^T$ , so  $f_X$  is indeed invariant for the GIS chain.

The eigenvalues of  $K_{YM}$  are  $\{1, 0.1230, 0.0075\}$ , so its spectral radius is 0.123. The eigenvalues of  $K$  and  $\tilde{K}$  are  $\{1, 0.1667, 0.0278\}$  and  $\{1, 0.6835, 0.0054\}$ , respectively. Thus, in this case, the GIS algorithm converges faster than either of the two underlying DA algorithms. Again, it is interesting to compare the exact answer, 0.123, with the upper bound from Y&M's Theorem 1. A straightforward, but somewhat tedious calculation shows that  $\mathcal{R}_{12} = 0.5477$ . Thus, Y&M's Theorem 1 says that the spectral radius of  $K_{YM}$  is bounded above by  $0.5477\sqrt{0.1667 \times 0.6835} = 0.1849$ .

## Acknowledgments

The authors thank Hani Doss and Vivekananda Roy for helpful discussions. The first author's work was supported by NSF Grant DMS-08-05860.

## References

- 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.
- KHARE, K. and HOBERT, J. P. (2010). A spectral analytic comparison of trace-class data augmentation algorithms and their sandwich variants. Tech. rep., University of Florida.
- 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. and WU, Y. N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association*, **94** 1264–1274.
- MENG, X.-L. and VAN DYK, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, **86** 301–320.
- RETFERD, J. R. (1993). *Hilbert Space: Compact Operators and the Trace Theorem*. Cambridge University Press, Cambridge.
- 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 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. (2003). Asymptotic variance and convergence rates of nearly-periodic MCMC algorithms. *Journal of the American Statistical Association*, **98** 169–177.

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.