Electronic Journal of Statistics
Vol. 7 (2013) 2054–2064
ISSN: 1935-7524
DOI: 10.1214/13-EJS837

The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic

Hee Min Choi and James P. Hobert

Department of Statistics
University of Florida
e-mail: heemin@stat.ufl.edu; jhobert@stat.ufl.edu

Abstract: One of the most widely used data augmentation algorithms is Albert and Chib's (1993) algorithm for Bayesian probit regression. Polson, Scott, and Windle (2013) recently introduced an analogous algorithm for Bayesian logistic regression. The main difference between the two is that Albert and Chib's (1993) truncated normals are replaced by so-called Polya-Gamma random variables. In this note, we establish that the Markov chain underlying Polson, Scott, and Windle's (2013) algorithm is uniformly ergodic. This theoretical result has important practical benefits. In particular, it guarantees the existence of central limit theorems that can be used to make an informed decision about how long the simulation should be run.

AMS 2000 subject classifications: Primary 60J27; secondary 62F15.
Keywords and phrases: Polya-Gamma distribution, data augmentation algorithm, minorization condition, Markov chain, Monte Carlo.

Received May 2013.

Contents

1 Introduction 2054
2 Polson, Scott and Windle's algorithm 2056
3 Uniform ergodicity 2058
Acknowledgements 2063
References 2064

1. Introduction

Consider a binary regression set-up in which $Y_1, \ldots, Y_n$ are independent Bernoulli random variables such that $\Pr(Y_i = 1|\beta) = F(x_i^T \beta)$, where x_i is a $p \times 1$ vector of known covariates associated with Y_i, β is a $p \times 1$ vector of unknown regression coefficients, and $F : \mathbb{R} \to (0, 1)$ is a distribution function. Two important special cases are probit regression, where F is the standard normal distribution function, and logistic regression, where F is the standard logistic distribution function, that is, $F(t) = e^t/(1 + e^t)$. In general, the joint mass function of

$Y_1, \ldots, Y_n$ is given by

$$\prod_{i=1}^{n} \Pr(Y_i = y_i|\beta) = \prod_{i=1}^{n} \left[F(x_i^T\beta)\right]^{y_i} \left[1 - F(x_i^T\beta)\right]^{1-y_i} I_{\{0,1\}}(y_i).$$

A Bayesian version of the model requires a prior distribution for the unknown regression parameter, β. If $\pi(\beta)$ is the prior density for β, then the posterior density of β given the data, $y = (y_1, \ldots, y_n)^T$, is defined as

$$\pi(\beta \,|\, y) = \frac{\pi(\beta)}{c(y)} \prod_{i=1}^{n} \left[F(x_i^T\beta)\right]^{y_i} \left[1 - F(x_i^T\beta)\right]^{1-y_i},$$

where $c(y)$ is the normalizing constant, that is,

$$c(y) := \int_{\mathbb{R}^p} \pi(\beta) \prod_{i=1}^{n} \left[F(x_i^T\beta)\right]^{y_i} \left[1 - F(x_i^T\beta)\right]^{1-y_i} d\beta.$$

Regardless of the choice of F, the posterior density is intractable in the sense that expectations with respect to $\pi(\beta \,|\, y)$, which are required for Bayesian inference, cannot be computed in closed form. Moreover, classical Monte Carlo methods based on independent and identically distributed (iid) samples from $\pi(\beta \,|\, y)$ are problematic when the dimension, p, is large. These difficulties have spurred the development of many Markov chain Monte Carlo methods for exploring $\pi(\beta \,|\, y)$. One of the most popular is a simple data augmentation (DA) algorithm for the probit regression case that was developed by Albert and Chib (1993). Each iteration of this algorithm has two steps, the first entails the simulation of n independent univariate truncated normals, and the second requires a draw from the p-variate normal distribution. A convergence rate analysis of the underlying Markov chain can be found in Roy and Hobert (2007).

Since the publication of Albert and Chib (1993), the search has been on for an analogous algorithm for Bayesian logistic regression. Attempts to develop such an analogue by mimicking the missing data argument of Albert and Chib (1993) in the logistic case have not been entirely successful. In particular, the resulting algorithms are much more complicated than that of Albert and Chib (1993), and simplified versions of them are inexact (see, e.g., Holmes and Held, 2006; Frühwirth-Schnatter and Frühwirth, 2010; Marchev, 2011). Using a different approach, Polson, Scott, and Windle (2013) (hereafter PS&W) have developed a real analogue of Albert and Chib's (1993) algorithm for Bayesian logistic regression. The main difference between the two algorithms is that the truncated normals in Albert and Chib's (1993) algorithm are replaced by Polya-Gamma random variables in PS&W's algorithm. We now describe the new algorithm.

In the remainder of this paper, we restrict attention to the logistic regression model with a proper $\mathrm{N}_p(b, B)$ prior on β. As usual, let X denote the $n \times p$ matrix whose ith row is x_i^T, and let $\mathbb{R}_+ = (0, \infty)$. For fixed $w \in \mathbb{R}_+^n$, define $\Sigma(w) = \left(X^T\Omega(w)X + B^{-1}\right)^{-1}$ and $m(w) = \Sigma(w)\left(X^T(y - \frac{1}{2}1_n) + B^{-1}b\right)$, where $\Omega(w)$ is the $n \times n$ diagonal matrix whose ith diagonal element is w_i, and 1_n is

an $n \times 1$ vector of 1s. Finally, let $\text{PG}(1, c)$ denote the Polya-Gamma distribution with parameters 1 and c, which is carefully defined in Section 2. The dynamics of PS&W's Markov chain are defined (implicitly) through the following two-step procedure for moving from the current state, $\beta^{(m)} = \beta$, to $\beta^{(m+1)}$.

Iteration $m + 1$ of PS&W's DA algorithm:

1. Draw $W_1, \ldots, W_n$ independently with
$$W_i \sim \text{PG}(1, |x_i^T \beta|),$$
and call the observed value $w = (w_1, \ldots, w_n)^T$
2. Draw $\beta^{(m+1)} \sim \text{N}_p\big(m(w), \Sigma(w)\big)$

Highly efficient methods of simulating Polya-Gamma random variables are provided by PS&W.

In this paper, we prove that PS&W's Markov chain is uniformly ergodic. This theoretical convergence rate result is extremely important when it comes to using PS&W's algorithm in practice to estimate intractable posterior expectations. Indeed, uniform ergodicity guarantees the existence of central limit theorems for averages such as $m^{-1} \sum_{i=0}^{m-1} g\big(\beta^{(i)}\big)$, where $g : \mathbb{R}^p \to \mathbb{R}$ is square integrable with respect to the target posterior, $\pi(\beta \,|\, y)$ (see, e.g., Tierney, 1994; Roberts and Rosenthal, 2004). It also allows for the computation of a consistent estimator of the associated asymptotic variance, which can be used to make an informed decision about how long the simulation should be run (see, e.g., Flegal, Haran, and Jones, 2008). We also establish the existence of the moment generating function (mgf) of the posterior distribution. It follows that, if $g(\beta_1, \ldots, \beta_p) = \beta_i^a$ for some $i \in \{1, 2, \ldots, p\}$ and $a > 0$, then g is square integrable with respect to the posterior.

The remainder of this paper is organized as follows. Section 2 contains a brief, but careful development of PS&W's DA algorithm. Section 3 contains the proof that the Markov chain underlying this algorithm is uniformly ergodic, as well as the proof that the posterior density, $\pi(\beta|y)$, has an mgf.

2. Polson, Scott and Windle's algorithm

PS&W's DA algorithm is based on a latent data representation of the posterior distribution. To describe it, we need to introduce what PS&W call the Polya-Gamma distribution. Let $\{E_k\}_{k=1}^{\infty}$ be a sequence of iid Exp(1) random variables and define

$$W = \frac{2}{\pi^2} \sum_{k=1}^{\infty} \frac{E_k}{(2k-1)^2}.$$

It is well-known (see, e.g., Biane, Pitman and Yor, 2001) that the random variable W has density

$$g(w) = \sum_{k=0}^{\infty} (-1)^k \frac{(2k+1)}{\sqrt{2\pi w^3}} e^{-\frac{(2k+1)^2}{8w}} I_{(0,\infty)}(w), \tag{2.1}$$

and that its Laplace transform is given by

$$\mathrm{E}\big[e^{-tW}\big] = \cosh^{-1}\big(\sqrt{t/2}\big).$$

(Recall that $\cosh(z) = (e^z + e^{-z})/2$.) PS&W create the Polya-Gamma family of densities through an exponential tilting of the density g. Indeed, consider a parametric family of densities, indexed by $c \geq 0$, that takes the form

$$f(x; c) = \cosh(c/2)\, e^{-\frac{c^2 x}{2}}\, g(x).$$

Of course, when $c = 0$, we recover the original density. A random variable with density $f(x; c)$ is said to have a PG$(1, c)$ distribution. We now describe a latent data formulation that leads to PS&W's DA algorithm. Our development is different, and we believe somewhat more transparent, than that given by PS&W.

Conditional on β, let $\{(Y_i, W_i)\}_{i=1}^n$ be independent random pairs such that Y_i and W_i are also independent, with $Y_i \sim \text{Bernoulli}\big(e^{x_i^T\beta}/(1 + e^{x_i^T\beta})\big)$ and $W_i \sim \text{PG}(1, |x_i^T\beta|)$. Let $W = (W_1, \ldots, W_n)^T$ and denote its density by $f(w|\beta)$. Combining this latent data model with the prior, $\pi(\beta)$, yields the *augmented posterior density* defined as

$$\pi(\beta, w|y) = \frac{\big[\prod_{i=1}^n \Pr(Y_i = y_i|\beta)\big] f(w|\beta)\, \pi(\beta)}{c(y)}.$$

Clearly,

$$\int_{\mathbb{R}_+^n} \pi(\beta, w|y)\, dw = \pi(\beta|y),$$

which is our target posterior density. PS&W's DA algorithm alternates between draws from $\pi(\beta|w, y)$ and $\pi(w|\beta, y)$. The conditional independence of Y_i and W_i implies that $\pi(w|\beta, y) = f(w|\beta)$. Thus, we can draw from $\pi(w|\beta, y)$ by making n independent draws from the Polya-Gamma distribution (as in the first step of the two-step procedure described in the Introduction). The other conditional density is multivariate normal. To see this, note that

$$\begin{aligned}
\pi(\beta|w, y) &\propto \left[\prod_{i=1}^n \Pr(Y_i = y_i|\beta)\right] f(w|\beta)\pi(\beta) \\
&= \left[\prod_{i=1}^n \frac{\big(e^{x_i^T\beta}\big)^{y_i}}{1 + e^{x_i^T\beta}}\right] \left[\prod_{i=1}^n \cosh\left(\frac{|x_i^T\beta|}{2}\right) e^{-\frac{(x_i^T\beta)^2 w_i}{2}}\, g(w_i)\right] \pi(\beta)
\end{aligned}$$

$$\propto \pi(\beta)\prod_{i=1}^{n}\left[\frac{\left(e^{x_i^T\beta}\right)^{y_i}}{1+e^{x_i^T\beta}}\right]\cosh\left(\frac{|x_i^T\beta|}{2}\right)e^{-\frac{(x_i^T\beta)^2 w_i}{2}}$$
$$= 2^{-n}\pi(\beta)\prod_{i=1}^{n}\exp\left\{y_i x_i^T\beta - \frac{x_i^T\beta}{2} - \frac{(x_i^T\beta)^2 w_i}{2}\right\},$$

where the last equality follows from the fact that $\cosh(z) = (1+e^{2z})/(2e^z)$. A routine Bayesian regression-type calculation then reveals that $\beta|w,y \sim \mathrm{N}_p(m(w),\Sigma(w))$, where $m(w)$ and $\Sigma(w)$ are defined in the Introduction.

The Markov transition density (Mtd) of the DA Markov chain, $\Phi = \{\beta^{(m)}\}_{m=0}^{\infty}$, is

$$k\big(\beta\big|\beta'\big) = \int_{\mathbb{R}_+^n} \pi(\beta|w,y)\,\pi(w|\beta',y)\,dw. \tag{2.2}$$

Of course, we do not have to perform the integration in (2.2), but we do need to be able to simulate random vectors with density $k\big(\,\cdot\,\big|\beta'\big)$, and this is exactly what the two-step procedure described in the Introduction does. Note that $k : \mathbb{R}^p \times \mathbb{R}^p \to (0,\infty)$; that is, k is strictly positive. It follows immediately that the Markov chain Φ is irreducible, aperiodic and Harris recurrent (see, e.g., Hobert, 2011).

3. Uniform ergodicity

For $m \in \{1,2,3,\ldots\}$, let $k^m : \mathbb{R}^p \times \mathbb{R}^p \to (0,\infty)$ denote the m-step Mtd of PS&W's Markov chain, where $k^1 \equiv k$. These densities are defined inductively as follows

$$k^m\big(\beta\big|\beta'\big) = \int_{\mathbb{R}^p} k^{m-1}\big(\beta\big|\beta''\big)\,k\big(\beta''\big|\beta'\big)\,d\beta''.$$

The conditional density of $\beta^{(m)}$ given $\beta^{(0)} = \beta'$ is precisely $k^m\big(\,\cdot\,\big|\beta'\big)$. The Markov chain Φ is geometrically ergodic if there exist $M : \mathbb{R}^p \to [0,\infty)$ and $\rho \in [0,1)$ such that, for all m,

$$\int_{\mathbb{R}^p}\Big|k^m\big(\beta\big|\beta'\big) - \pi(\beta|y)\Big|\,d\beta \le M(\beta')\,\rho^m. \tag{3.1}$$

The quantity on the left-hand side of (3.1) is the total variation distance between the posterior distribution and the distribution of $\beta^{(m)}$ (given $\beta^{(0)} = \beta'$). If the function $M(\cdot)$ is bounded above, then the chain is uniformly ergodic.

One method of establishing uniform ergodicity is to construct a *minorization condition* that holds on the entire state space. In particular, if we can find a $\delta > 0$ and a density function $h : \mathbb{R}^p \to [0,\infty)$ such that, for all $\beta, \beta' \in \mathbb{R}^p$,

$$k\big(\beta\big|\beta'\big) \ge \delta\,h(\beta), \tag{3.2}$$

then the chain is uniformly ergodic. In fact, if (3.2) holds, then (3.1) holds with $M \equiv 1$ and $\rho = 1-\delta$ (see, e.g., Jones and Hobert, 2001, Section 3.2).

In order to state our main result, we need a bit more notation. Let $\phi(z;\mu,V)$ denote the multivariate normal density with mean μ and covariance matrix V evaluated at the point z. Recall that the prior on β is $\mathrm{N}_p(b,B)$, and define

$$s = B^{\frac{1}{2}}X^T\Big(y - \frac{1}{2}1_n\Big) + B^{-\frac{1}{2}}b.$$

Here is our main result.

Proposition 3.1. *The Mtd of* Φ *satisfies the following minorization condition*

$$k\big(\beta\big|\beta'\big) \geq \delta\phi(\beta; m_*, \Sigma_*)\ ,$$

where $m_* = \big(\frac{1}{2}X^TX + B^{-1}\big)^{-1}B^{-\frac{1}{2}}s$, $\Sigma_* = \big(\frac{1}{2}X^TX + B^{-1}\big)^{-1}$, *and*

$$\delta = \frac{|\Sigma_*|^{1/2}}{|B|^{1/2}}\exp\left\{\frac{1}{2}m_*^T\Sigma_*^{-1}m_*\right\}e^{-\frac{n}{4}}2^{-n}\exp\left\{-\frac{1}{2}s^Ts\right\}.$$

Hence, PS&W's Markov chain is uniformly ergodic.

Remark 3.1. As discussed in the Introduction, uniform ergodicity guarantees the existence of central limit theorems and consistent estimators of the associated asymptotic variances, and these can be used to make an informed decision about how long the simulation should be run. While it is true that, in many cases, δ will be so close to zero that the bound (3.1) (with $M \equiv 1$ and $\rho = 1-\delta$) will not be useful for getting a handle on the total variation distance, this does not detract from the usefulness of the result. Moreover, it is certainly possible that a different analysis could yield a different minorization condition with a larger δ.

The following easily established lemmas will be used in the proof of Proposition 3.1.

Lemma 3.1. *If* A *is a symmetric nonnegative definite matrix, then all of the eigenvalues of* $(I+A)^{-1}$ *are in* $(0,1]$, *and* $I-(I+A)^{-1}$ *is also nonnegative definite.*

Lemma 3.2. *For* $a,b \in \mathbb{R}$, $\cosh(a+b) \leq 2\cosh(a)\cosh(b)$.

Proof of Proposition 3.1. Recall that $\Sigma = \Sigma(w) = \big(X^T\Omega(w)X + B^{-1}\big)^{-1}$ and $m = m(w) = \Sigma(w)\big(X^T(y - \frac{1}{2}1_n) + B^{-1}b\big)$. We begin by showing that $|\Sigma|^{-\frac{1}{2}} \geq |B|^{-\frac{1}{2}}$. Indeed,

$$\begin{aligned}|\Sigma| &= \big|(X^T\Omega X + B^{-1})^{-1}\big| = \big|\big(B^{-\frac{1}{2}}B^{\frac{1}{2}}X^T\Omega X B^{\frac{1}{2}}B^{-\frac{1}{2}} + B^{-\frac{1}{2}}B^{-\frac{1}{2}}\big)^{-1}\big| \\ &= |B|\big|(\tilde{X}^T\Omega\tilde{X} + I)^{-1}\big|,\end{aligned}$$

where $\tilde{X} = XB^{\frac{1}{2}}$. Now, since $\tilde{X}^T\Omega\tilde{X}$ is nonnegative definite, Lemma 3.1 implies that $|\Sigma| \leq |B|$, and the result follows. Next, we show that $m^T\Sigma^{-1}m \leq s^Ts$. Letting $l = y - \frac{1}{2}1_n$, we have

$$m^T\Sigma^{-1}m = (X^Tl + B^{-1}b)^T(X^T\Omega X + B^{-1})^{-1}(X^Tl + B^{-1}b)$$

$$
\begin{aligned}
&= (X^T l + B^{-1}b)^T B^{\frac{1}{2}} (\tilde{X}^T \Omega \tilde{X} + I)^{-1} B^{\frac{1}{2}} (X^T l + B^{-1}b) \\
&= (\tilde{X}^T l + B^{-\frac{1}{2}}b)^T (\tilde{X}^T \Omega \tilde{X} + I)^{-1} (\tilde{X}^T l + B^{-\frac{1}{2}}b) \\
&\leq (\tilde{X}^T l + B^{-\frac{1}{2}}b)^T (\tilde{X}^T l + B^{-\frac{1}{2}}b) \\
&= s^T s,
\end{aligned}
$$

where the inequality follows from Lemma 3.1. Using these two inequalities, we have

$$
\begin{aligned}
&\pi(\beta|w,y) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\beta - m)^T \Sigma^{-1} (\beta - m) \right\} \\
&= (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}\beta^T (X^T \Omega X)\beta - \frac{1}{2}\beta^T B^{-1}\beta - \frac{1}{2} m^T \Sigma^{-1} m + m^T \Sigma^{-1}\beta \right\} \\
&\geq (2\pi)^{-\frac{p}{2}} |B|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}\beta^T (X^T \Omega X)\beta - \frac{1}{2}\beta^T B^{-1}\beta - \frac{1}{2} s^T s + m^T \Sigma^{-1}\beta \right\} \\
&= (2\pi)^{-\frac{p}{2}} |B|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n} w_i (x_i^T \beta)^2 - \frac{1}{2}\beta^T B^{-1}\beta - \frac{1}{2} s^T s + s^T B^{-\frac{1}{2}}\beta \right\} \\
&= (2\pi)^{-\frac{p}{2}} |B|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}\beta^T B^{-1}\beta - \frac{1}{2} s^T s + s^T B^{-\frac{1}{2}}\beta \right\} \left[\prod_{i=1}^{n} \exp\left\{ -\frac{(x_i^T \beta)^2}{2} w_i \right\} \right].
\end{aligned}
$$

Now since

$$
\pi(w|\beta, y) = \prod_{i=1}^{n} \cosh\left(\frac{|x_i^T \beta|}{2} \right) \exp\left\{ -\frac{(x_i^T \beta)^2}{2} w_i \right\} g(w_i) \,,
$$

it follows that

$$
\begin{aligned}
&\pi(\beta|w,y)\pi(w|\beta',y) \\
&\quad \geq (2\pi)^{-\frac{p}{2}} |B|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}\beta^T B^{-1}\beta - \frac{1}{2} s^T s + s^T B^{-\frac{1}{2}}\beta \right\} \\
&\qquad \times \left[\prod_{i=1}^{n} \cosh\left(\frac{|x_i^T \beta'|}{2} \right) \exp\left\{ -\left[\frac{(x_i^T \beta)^2 + (x_i^T \beta')^2}{2} \right] w_i \right\} g(w_i) \right].
\end{aligned}
$$

Recall that $k(\beta|\beta') = \int_{\mathbb{R}_+^n} \pi(\beta|w,y)\, \pi(w|\beta',y)\, dw$. To this end, note that

$$
\begin{aligned}
&\int_{\mathbb{R}_+} \exp\left\{ -\left[\frac{(x_i^T \beta)^2 + (x_i^T \beta')^2}{2} \right] w_i \right\} g(w_i)\, dw_i \\
&\qquad = \left\{ \cosh\left(\frac{\sqrt{(x_i^T \beta)^2 + (x_i^T \beta')^2}}{2} \right) \right\}^{-1} \\
&\qquad \geq \left\{ \cosh\left(\frac{|x_i^T \beta|}{2} + \frac{|x_i^T \beta'|}{2} \right) \right\}^{-1}
\end{aligned}
$$

$$\geq \left\{ 2\cosh\left(\frac{|x_i^T\beta|}{2}\right)\cosh\left(\frac{|x_i^T\beta'|}{2}\right)\right\}^{-1},$$

where the first inequality is due to the fact that $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$ for non-negative a, b, and the second inequality is by Lemma 3.2. It follows that

$$\begin{aligned}&\int_{\mathbb{R}_+^n}\left[\prod_{i=1}^n \cosh\left(\frac{|x_i^T\beta'|}{2}\right)\exp\left\{-\frac{(x_i^T\beta)^2+(x_i^T\beta')^2}{2}w_i\right\}g(w_i)\right]dw\\ &\quad\geq 2^{-n}\left[\prod_{i=1}^n \cosh\left(\frac{|x_i^T\beta|}{2}\right)\right]^{-1}.\end{aligned}$$

Moreover,

$$\begin{aligned}\left[\prod_{i=1}^n \cosh\left(\frac{|x_i^T\beta|}{2}\right)\right]^{-1} &\geq \left[\prod_{i=1}^n \exp\left(\frac{|x_i^T\beta|}{2}\right)\right]^{-1}\\ &\geq \left[\prod_{i=1}^n \exp\left(\frac{(x_i^T\beta)^2+1}{4}\right)\right]^{-1}\\ &= \exp\left\{-\frac{1}{2}\sum_{i=1}^n\left(\frac{(x_i^T\beta)^2}{2}+\frac{1}{2}\right)\right\}\\ &= e^{-\frac{n}{4}}\exp\left\{-\frac{1}{2}\left(\frac{\beta^T X^T X\beta}{2}\right)\right\},\end{aligned}$$

where the second inequality holds because $|a| \leq (a^2+1)/2$ for any real a. Putting all of this together, we have

$$\begin{aligned}k(\beta|\beta') &= \int_{\mathbb{R}_+^n}\pi(\beta|w,y)\,\pi(w|\beta',y)\,dw\\ &\geq (2\pi)^{-\frac{p}{2}}|B|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\beta^T B^{-1}\beta-\frac{1}{2}s^T s+s^T B^{-\frac{1}{2}}\beta\right\}\\ &\qquad\times 2^{-n}e^{-\frac{n}{4}}\exp\left\{-\frac{1}{2}\left(\frac{\beta^T X^T X\beta}{2}\right)\right\}\\ &= (2\pi)^{-\frac{p}{2}}|B|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\beta^T\left(\frac{X^T X}{2}+B^{-1}\right)\beta+s^T B^{-\frac{1}{2}}\beta\right\}2^{-n}e^{-\frac{n}{4}-\frac{s^T s}{2}}\\ &= (2\pi)^{-\frac{p}{2}}|\Sigma_*|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\beta-m_*)^T\Sigma_*^{-1}(\beta-m_*)\right\}\\ &\qquad\times|\Sigma_*|^{\frac{1}{2}}|B|^{-\frac{1}{2}}2^{-n}e^{-\frac{n}{4}-\frac{s^T s}{2}+\frac{m_*^T\Sigma_*^{-1}m_*}{2}}\\ &= \delta\,\phi(\beta;m_*,\Sigma_*),\end{aligned}$$

and the proof is complete. □

Remark 3.2. It is worth pointing out that our proof circumvents the need to deal directly with the unwieldy density of the PG$(1, 0)$ distribution, which is given in (2.1).

The utility of the Polya-Gamma latent data strategy extends far beyond the basic logistic regression model. Indeed, PS&W use this technique to develop useful Gibbs samplers for a variety of Bayesian hierarchical models in which a logit link is employed. These include Bayesian logistic regression with random effects and negative-binomial regression. Of course, given the results herein, it is natural to ask whether these other Gibbs samplers are also uniformly ergodic, and to what extent our methods could be used to answer this question. We note, however, that the Mtd of the Polya-Gamma Gibbs sampler is relatively simple. For example, the Gibbs sampler for the mixed effects model has more than two-steps, so its Mtd is significantly more complex than that of the Polya-Gamma Gibbs sampler. Thus, the development of (uniform) minorization conditions for the other Gibbs samplers would likely entail more than just a straightforward extension of the proof of Proposition 3.1. On the other hand, it is possible that some of the inequalities in that proof could be recycled.

Finally, we establish that the posterior distribution of β given the data has a moment generating function.

Proposition 3.2. *For any fixed $t \in \mathbb{R}^p$,*

$$\int_{\mathbb{R}^p} e^{\beta^T t}\, \pi(\beta \,|\, y)\, d\beta < \infty.$$

Hence, the moment generating function of the posterior distribution exists.

Remark 3.3. Chen and Shao (2000) develop results for Bayesian logistic regression with a flat (improper) prior on β. In particular, these authors provide conditions on X and y that guarantee the existence of the mgf of the posterior of β given the data. Note that our result holds for all X and y

Proof. Recall that

$$\int_{\mathbb{R}^n_+} \pi(\beta, w|y)\, dw = \pi(\beta|y),$$

where

$$\pi(\beta, w|y) = \frac{\big[\prod_{i=1}^n \Pr(Y_i = y_i|\beta)\big] f(w|\beta)\, \pi(\beta)}{c(y)}.$$

Hence, it suffices to show that

$$\int_{\mathbb{R}^n_+} \int_{\mathbb{R}^p} e^{\beta^T t}\, \pi(\beta, w|y)\, d\beta\, dw < \infty.$$

Straightforward calculations similar to those done in Section 2 show that

$$\pi(\beta, w|y) = \frac{\phi(\beta; m, \Sigma)}{2^n c(y)} \frac{|\Sigma|^{\frac{1}{2}}}{|B|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} b^T B^{-1} b \right\} \exp\left\{ \frac{1}{2} m^T \Sigma^{-1} m \right\} \prod_{i=1}^n g(w_i).$$

Then, since $|\Sigma| \leq |B|$ and $m^T\Sigma^{-1}m \leq s^Ts$, we have

$$\pi(\beta, w|y) \leq \frac{\phi(\beta; m, \Sigma)}{2^n c(y)} \exp\left\{ -\frac{1}{2} b^T B^{-1} b \right\} \exp\left\{ \frac{1}{2} s^T s \right\} \prod_{i=1}^n g(w_i).$$

Now, using the formula for the multivariate normal mgf, it suffices to show that

$$\begin{aligned} &\int_{\mathbb{R}_+^n} \int_{\mathbb{R}^p} e^{\beta^T t}\, \phi(\beta; m, \Sigma) \left[\prod_{i=1}^n g(w_i) \right] d\beta\, dw \\ &= \int_{\mathbb{R}_+^n} \exp\left\{ m^T t + \frac{1}{2} t^T \Sigma t \right\} \left[\prod_{i=1}^n g(w_i) \right] dw < \infty. \end{aligned}$$

We establish this by demonstrating that $m^T t + \frac{1}{2} t^T \Sigma t$ is uniformly bounded in w.

For a matrix A, define $\|A\| = \sup_{\|x\|=1} \|Ax\|$. Of course, if A is non-negative definite, then $\|A\|$ is equal to the largest eigenvalue of A. Now, using Cauchy-Schwartz and properties of the norm, we have

$$|m^T t|^2 \leq \|m\|^2 \|t\|^2 = \|B^{\frac{1}{2}} B^{-\frac{1}{2}} m\|^2 \|t\|^2 \leq \|B^{\frac{1}{2}}\|^2 \|B^{-\frac{1}{2}} m\|^2 \|t\|^2.$$

Now since $B^{-\frac{1}{2}} m = \left(\tilde{X}^T \Omega \tilde{X} + I\right)^{-1} \left(\tilde{X}^T (y - \frac{1}{2} 1_n) + B^{-\frac{1}{2}} b\right)$, we have

$$\begin{aligned} \|B^{-\frac{1}{2}} m\|^2 &= \left\| \left(\tilde{X}^T \Omega \tilde{X} + I\right)^{-1} \left(\tilde{X}^T \left(y - \frac{1}{2} 1_n \right) + B^{-\frac{1}{2}} b \right) \right\|^2 \\ &\leq 2 \left\| \left(\tilde{X}^T \Omega \tilde{X} + I\right)^{-1} \tilde{X}^T \left(y - \frac{1}{2} 1_n \right) \right\|^2 + 2 \left\| \left(\tilde{X}^T \Omega \tilde{X} + I\right)^{-1} B^{-\frac{1}{2}} b \right\|^2 \\ &\leq 2 \left\| \left(\tilde{X}^T \Omega \tilde{X} + I\right)^{-1} \right\|^2 \left\| \tilde{X}^T \left(y - \frac{1}{2} 1_n \right) \right\|^2 \\ &\qquad + 2 \left\| \left(\tilde{X}^T \Omega \tilde{X} + I\right)^{-1} \right\|^2 \left\| B^{-\frac{1}{2}} b \right\|^2 \\ &\leq 2 \left\| \tilde{X}^T \left(y - \frac{1}{2} 1_n \right) \right\|^2 + 2 \left\| B^{-\frac{1}{2}} b \right\|^2, \end{aligned}$$

where the first inequality is due to the fact that $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for any vectors a and b, and the third inequality is due to the fact that $\left\| \left(\tilde{X}^T \Omega \tilde{X} + I\right)^{-1} \right\|^2 \leq 1$ (Lemma 1). Hence, $|m^T t|$ is uniformly bounded in w. Finally, another application of Lemma 1 yields

$$t^T \Sigma t = t^T \left(X^T \Omega X + B^{-1} \right)^{-1} t = t^T B^{\frac{1}{2}} \left(\tilde{X}^T \Omega \tilde{X} + I \right)^{-1} B^{\frac{1}{2}} t \leq t^T B t,$$

so $t^T \Sigma t$ is also uniformly bounded in w, and the result follows. □

Acknowledgements

The second author was supported by NSF Grant DMS-11-06395.

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88** 669–679. MR1224394

Polson, N.G., Scott, J.G. and Windle, J. (2013). Bayesian inference for logistic models using Polya-Gamma latent variables. *Journal of the American Statistical Association* (to appear). arXiv:1205.0310v3.

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. MR2370071

Holmes, C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* **1** 145–168. MR2227368

Frühwirth-Schnatter, S. and Frühwirth, R. (2010). Data augmentation and MCMC for binary and multinomial logit models. In *Statistical Modelling and Regression Structures*, 111–132. Springer-Verlag. MR2664631

Marchev, D. (2011). Markov chain Monte Carlo algorithms for the Bayesian logistic regression model. In *Proceedings of the Annual International Conference on Operations Research and Statistics*, 154–159.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics* **22** 1701–1762. MR1329166

Roberts, G.O. and Rosenthal, J.S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys* **1** 20–71. MR2095565

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. MR2516823

Biane, P., Pitman, J. and Yor, M. (2001). Probability laws related to the Jacobi theta and Riemann zeta functions, and Brownian excursions. *Bulletin of the American Mathematical Society* **38** 435–465. MR1848256

Hobert, J.P. (2011). The data augmentation algorithm: Theory and methodology. In *Handbook of Markov Chain Monte Carlo* (S. Brooks, A. Gelman, G. Jones and X.L. Meng, eds.). Chapman & Hall/CRC Press. MR2742422

Jones, G.L. and Hobert, J.P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science* **16** 312–34. MR1888447

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. MR1694452