The Annals of Statistics
2019, Vol. 47, No. 4, 2320–2347
https://doi.org/10.1214/18-AOS1749

CONVERGENCE COMPLEXITY ANALYSIS OF ALBERT AND CHIB'S ALGORITHM FOR BAYESIAN PROBIT REGRESSION

BY QIAN QIN AND JAMES P. HOBERT[1]

University of Florida

The use of MCMC algorithms in high dimensional Bayesian problems has become routine. This has spurred so-called convergence complexity analysis, the goal of which is to ascertain how the convergence rate of a Monte Carlo Markov chain scales with sample size, n, and/or number of covariates, p. This article provides a thorough convergence complexity analysis of Albert and Chib's [*J. Amer. Statist. Assoc.* **88** (1993) 669–679] data augmentation algorithm for the Bayesian probit regression model. The main tools used in this analysis are drift and minorization conditions. The usual pitfalls associated with this type of analysis are avoided by utilizing centered drift functions, which are minimized in high posterior probability regions, and by using a new technique to suppress high-dimensionality in the construction of minorization conditions. The main result is that the geometric convergence rate of the underlying Markov chain is bounded below 1 both as $n \to \infty$ (with p fixed), and as $p \to \infty$ (with n fixed). Furthermore, the first computable bounds on the total variation distance to stationarity are byproducts of the asymptotic analysis.

1. Introduction. Markov chain Monte Carlo (MCMC) has become an indispensable tool in Bayesian analysis, and it is now well known that the ability to utilize an MCMC algorithm in a principled manner (e.g., with regard to choosing the Monte Carlo sample size) requires an understanding of the convergence properties of the underlying Markov chain [see, e.g., Flegal, Haran and Jones (2008)]. Taking this a step further, in modern high dimensional problems it is also important to understand how the convergence properties of the chain change as the sample size, n, and/or number of covariates, p, increases. Denote the data (i.e., responses and covariates) by $\mathcal{D}$. If we imagine n or p (or both) increasing, this leads to consideration of a sequence of data sets, $\{\mathcal{D}_j\}$, and corresponding sequences of posterior distributions and Monte Carlo Markov chains. A natural question to ask is "What can we say about the convergence properties of the Markov chains as $j \to \infty$?" There is currently a great deal of interest in questions like this in

Received December 2017; revised April 2018.
[1]Supported by NSF Grant DMS-15-11945.
Primary 60J05; secondary 65C05.
Key words and phrases. Drift condition, geometric ergodicity, high dimensional inference, large p-small n, Markov chain Monte Carlo, minorization condition.

the MCMC community [see, e.g., Durmus and Moulines (2016), Johndrow et al. (2016), Rajaratnam and Sparks (2015), Yang and Rosenthal (2017), Yang, Wainwright and Jordan (2016)]. Rajaratnam and Sparks (2015) call this the study of *convergence complexity*, and we will follow their lead.

Asymptotic analysis of the convergence properties of a sequence of Markov chains associated with increasingly large *finite* state spaces has a long history in the computer science literature, dating back at least to Sinclair and Jerrum (1989). While the techniques developed in computer science have been successfully applied to a few problems in statistics [see, e.g., Yang, Wainwright and Jordan (2016)], they are generally not applicable in situations where the state space is high-dimensional and uncountable, which is the norm for Monte Carlo Markov chains in Bayesian statistics. In this paper, we employ methods based on drift and minorization conditions to analyze the convergence complexity of one such Monte Carlo Markov chain.

Let $\pi : \mathsf{X} \to [0, \infty)$ denote an intractable probability density function (pdf), where $\mathsf{X} \subset \mathbb{R}^d$, and let $\Pi(\cdot)$ denote the corresponding probability measure, that is, for measurable C, $\Pi(C) = \int_C \pi(x)\,\mathrm{d}x$. Let $K(x, \cdot)$, $x \in \mathsf{X}$, denote the Markov transition function (Mtf) of an irreducible, aperiodic, Harris recurrent Markov chain with invariant probability measure Π. [See Meyn and Tweedie (2009) for definitions.] The chain is called *geometrically ergodic* if there exist $M : \mathsf{X} \to [0, \infty)$ and $\rho \in [0, 1)$ such that

$$\|K^m(x, \cdot) - \Pi(\cdot)\|_{\mathrm{TV}} \le M(x)\rho^m \qquad \text{for all } x \in \mathsf{X} \text{ and all } m \in \mathbb{N}, \tag{1}$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes total variation norm, and $K^m(x, \cdot)$ is the m-step Mtf. The important practical benefits of basing one's MCMC algorithm on a geometrically ergodic Markov chain have been well documented by, for example, Roberts and Rosenthal (1998), Jones and Hobert (2001), Flegal, Haran and Jones (2008) and Łatuszyński, Miasojedow and Niemiro (2013). Define the *geometric convergence rate* of the chain as

$$\rho_* = \inf\{\rho \in [0, 1] : (1) \text{ holds for some } M : \mathsf{X} \to [0, \infty)\}.$$

Clearly, the chain is geometrically ergodic if and only if $\rho_* < 1$.

Establishing the convergence rate of a practically relevant Monte Carlo Markov chain can be quite challenging. A key tool for this purpose has been the technique developed by Rosenthal (1995), which allows for the construction of an upper bound on ρ_* using drift and minorization conditions [see also Baxendale (2005), Hairer and Mattingly (2011), Meyn and Tweedie (1994), Roberts and Tweedie (1999)]. This method, which is described in detail in Section 2, has been used to establish the geometric ergodicity of myriad Monte Carlo Markov chains [see, e.g., Fort et al. (2003), Marchev and Hobert (2004), Roy and Hobert (2010), Vats (2017)]. Since methods based on drift and minorization (hereafter, d&m) are still the most (and arguably the only) reliable tools for bounding ρ_* for practically relevant Monte Carlo Markov chains on uncountable state spaces, it is important to

know whether they remain useful in the context of convergence complexity analysis. Unfortunately, it turns out that most of the upper bounds on ρ_* that have been produced using techniques based on d&m converge to 1 (often exponentially fast), becoming trivial, as n and/or p grow [see, e.g., Rajaratnam and Sparks (2015)]. One example of this is Roy and Hobert's (2007) analysis of Albert and Chib's (1993) data augmentation algorithm for the Bayesian probit model, which establishes geometric ergodicity of the underlying Markov chain, but also leads to an upper bound on ρ_* that converges to 1 as $n \to \infty$.

There are, of course, many possible explanations for why the d&m-based upper bounds on ρ_* converge to 1. It could be that the associated Monte Carlo Markov chains actually have poor asymptotic properties, or, if not, perhaps d&m-based methods are simply not up to the more delicate task of convergence complexity analysis. We show that, in the case of Albert and Chib's (1993) chain, neither of these potential explanations is correct. Indeed, our careful d&m analysis of Albert and Chib's (1993) chain (hereafter, A&C's chain) leads to upper bounds on ρ_* that are bounded away from 1 in both the large n, small p case, and the large p, small n case. We believe that this is the first successful convergence complexity analysis of a practically relevant Monte Carlo Markov chain using d&m. The key ideas we use to establish our results include "centering" the drift function to a region in the state space that the chain visits frequently, and suppressing high-dimensionality in the construction of minorization conditions. In particular, for two-block Gibbs chains, we introduce a technique for constructing asymptotically stable minorization conditions that is based on the well-known fact that the two marginal Markov chains and the joint chain all share the same geometric convergence rate.

Recently, Yang and Rosenthal (2017) used a modified version of Rosenthal's (1995) technique to successfully analyze the convergence complexity of a Gibbs sampler for a simple Bayesian linear mixed model. We note that, because one of the variance components in their model is assumed known, it is actually straightforward to sample directly from the posterior distribution using a univariate rejection sampler [Jones (2001), Section 3.9]. Thus, while Yang and Rosenthal's (2017) results are impressive, and their methods may suggest a way forward, the Monte Carlo Markov chain that they analyzed is not practically relevant.

Before describing our results for A&C's chain, we introduce an alternative definition of geometric ergodicity. Let $L^2(\Pi)$ denote the set of signed measures μ that are absolutely continuous with respect to Π, and satisfy $\int_{\mathsf{X}} (\mathrm{d}\mu/\mathrm{d}\Pi)^2 \, \mathrm{d}\Pi < \infty$. As in Roberts and Rosenthal (1997), we say that the Markov chain with Mtf K is L^2-*geometrically ergodic* if there exists $\rho < 1$ such that for each probability measure $\nu \in L^2(\Pi)$, there exists a constant $M_\nu < \infty$ such that

$$\left\| \nu K^m(\cdot) - \Pi(\cdot) \right\|_{\mathrm{TV}} \leq M_\nu \rho^m \qquad \text{for all } m \in \mathbb{N},$$

where $\nu K^m(\cdot) = \int_{\mathsf{X}} K^m(x, \cdot) \nu(\mathrm{d}x)$. We define the L^2-*geometric convergence rate*, ρ_{**}, to be the infimum of all $\rho \in [0, 1]$ that satisfy this definition. Not surprisingly, ρ_* and ρ_{**} are closely related [see, e.g., Roberts and Tweedie (2001)].

We now provide an overview of our results for Albert and Chib's (1993) Markov chain, starting with a brief description of their algorithm. Let $\{X_i\}_{i=1}^n$ be a set of p-dimensional covariate vectors, and let $\{Y_i\}_{i=1}^n$ be a corresponding sequence of binary random variables such that $Y_i|X_i, B \sim \text{Bernoulli}(\Phi(X_i^T B))$ independently, where B is a $p \times 1$ vector of unknown regression coefficients, and Φ is the standard normal distribution function. Consider a Bayesian analysis using a prior density for B given by

$$\omega(\beta) \propto \exp\left\{-\frac{1}{2}(\beta - v)^T Q(\beta - v)\right\}, \tag{2}$$

where $v \in \mathbb{R}^p$, and $Q \in \mathbb{R}^{p\times p}$ is either a positive definite matrix (proper Gaussian prior), or a zero matrix (flat improper prior). Assume for now that the posterior is proper. (Propriety under a flat prior is discussed in Section 3.) As usual, let X denote the $n \times p$ matrix whose ith row is X_i^T, and let $Y = (Y_1\ Y_2\ \cdots\ Y_n)^T$ denote the vector of responses. The intractable posterior density is given by

$$\pi_{B|Y,X}(\beta|Y, X) \propto \left\{\prod_{i=1}^n \big(\Phi\big(X_i^T\beta\big)\big)^{Y_i}\big(1 - \Phi\big(X_i^T\beta\big)\big)^{1-Y_i}\right\}\omega(\beta). \tag{3}$$

The standard method for exploring (3) is the classical data augmentation algorithm of Albert and Chib (1993), which simulates a Harris ergodic (irreducible, aperiodic and Harris recurrent) Markov chain, $\{B_m\}_{m=0}^\infty$, that has invariant density $\pi_{B|Y,X}$. In order to state the algorithm, we require a bit of notation. For $\theta \in \mathbb{R}$, $\sigma > 0$, and $i \in \{0, 1\}$, let $\text{TN}(\theta, \sigma^2; i)$ denote the $\text{N}(\theta, \sigma^2)$ distribution truncated to $(0, \infty)$ if $i = 1$, and to $(-\infty, 0)$ if $i = 0$. The matrix $\Sigma := X^T X + Q$ is necessarily nonsingular because of propriety. If the current state of A&C's chain is $B_m = \beta$, then the new state, B_{m+1}, is drawn using the two steps in Algorithm 1.

The convergence rate of A&C's chain has been studied by several authors. Roy and Hobert (2007) proved that when $Q = 0$, the chain is always geometrically ergodic. (Again, we are assuming posterior propriety.) A similar result for proper normal priors was established by Chakraborty and Khare (2017). Both results were established using a technique that does not require construction of a minorization condition [see Meyn and Tweedie (2009), Lemma 15.2.8], and consequently, does not yield an explicit upper bound on ρ_*. Thus, neither paper addresses the issue of convergence complexity. However, in Section 5 we prove that Roy and Hobert's

Algorithm 1 Iteration $m + 1$ of the data augmentation algorithm

1. Draw $\{Z_i\}_{i=1}^n$ independently with $Z_i \sim \text{TN}(X_i^T\beta, 1; Y_i)$, and let $Z = (Z_1\ Z_2\ \cdots\ Z_n)^T$.
2. Draw

$$B_{m+1} \sim \text{N}_p\big(\Sigma^{-1}\big(X^T Z + Qv\big), \Sigma^{-1}\big).$$

(2007) drift function *cannot* be used to construct an upper bound on ρ_* that is bounded away from 1 as $n \to \infty$. Johndrow et al. (2016) recently established a convergence complexity result for the intercept only version of the model ($p = 1$ and $X_i = 1$ for $i = 1, 2, \ldots, n$) with a proper (univariate) normal prior, under the assumption that *all* the responses are successes ($Y_i = 1$ for $i = 1, 2, \ldots, n$). Their results, which are based on Cheeger's inequality, imply that $\rho_{**} \to 1$ as $n \to \infty$, indicating that the algorithm is inefficient for large samples.

The results established herein provide a much more complete picture of the convergence behavior of A&C's chain. Three different regimes are considered: (i) fixed n and p, (ii) large n, small p and (iii) large p, small n. Our analysis is based on two different drift functions that are both appropriately centered (at the posterior mode). One of the two drift functions is designed for regime (ii), and the other for regime (iii). We establish d&m conditions for both drift functions, and these are used in conjunction with Rosenthal's (1995) result to construct two explicit upper bounds on ρ_*. They are also used to construct two computable upper bounds on the total variation distance to stationarity [as in (1)], which improves upon the analyses of Roy and Hobert (2007) and Chakraborty and Khare (2017).

The goal in regime (ii) is to study the asymptotic behavior of the geometric convergence rate as $n \to \infty$, when p is fixed. To this end, we consider a sequence of data sets, $\mathcal{D}_n := \{(X_i, Y_i)\}_{i=1}^n$. So, each time n increases by 1, we are are given a new $p \times 1$ covariate vector, X_i, and a corresponding binary response, Y_i. To facilitate the asymptotic study, we assume that the (X_i, Y_i) pairs are generated according to a *random* mechanism that is governed by very weak assumptions (that are consistent with the probit regression model). We show that there exists a constant $\rho < 1$ such that, almost surely, $\limsup_{n\to\infty} \rho_*(\mathcal{D}_n) \le \rho$. Apart from this general result, we are also able to show that, in the intercept only model considered by Johndrow et al. (2016), the A&C chain is actually quite well behaved as long as the proportion of successes is bounded away from 0 and 1. To be specific, let $\{Y_i\}_{i=1}^\infty$ denote a *fixed* sequence of binary responses, and let $\hat{p}_n = n^{-1} \sum_{i=1}^n Y_i$. Our results imply that, as long as $0 < \liminf_{n\to\infty} \hat{p}_n \le \limsup_{n\to\infty} \hat{p}_n < 1$, there exists $\rho < 1$ such that both $\rho_*(\mathcal{D}_n)$ and $\rho_{**}(\mathcal{D}_n)$ are eventually bounded above by ρ, and there is a closed form expression for ρ.

In regime (iii), n is fixed and $p \to \infty$. There are several important differences between regimes (ii) and (iii). First, in regime (ii), since p is fixed, only a single prior distribution need be considered. In contrast, when $p \to \infty$, we must specify a sequence of priors, $\{(Q_p, v_p)\}_{p=1}^\infty$, where v_p is a $p \times 1$ vector, and Q_p is a $p \times p$ positive definite matrix. (When $p > n$, a positive definite Q_p is required for posterior propriety.) Also, in regime (iii), there is a fixed vector of responses (of length n), and it is somewhat unnatural to consider the new columns of X to be random. Let $\{\mathcal{D}_p\} := \{(v_p, Q_p, X_{n\times p}, Y)\}$ denote a fixed sequence of priors and data sets, where Y is a fixed $n \times 1$ vector of responses, and $X_{n\times p}$ is an $n \times p$ matrix. We show that, under a natural regularity condition on $X_{n\times p} Q_p^{-1} X_{n\times p}^T$, there exists a $\rho < 1$ such that $\rho_*(\mathcal{D}_p) \le \rho$ for all p.

The remainder of the paper is laid out as follows. In Section 2, we formally introduce the concept of *stable* d&m conditions, and describe techniques that we employ for constructing such. The centered drift functions that are used in our analysis of A&C's chain are described in Section 3. In Section 4, we provide results for A&C's chain in the case where n and p are both fixed. Two sets of d&m conditions are established, and corresponding exact total variation bounds on the distance to stationarity are provided. The heart of the paper is Section 5 where it is shown that the geometric convergence rate of A&C's chain is bounded away from 1 as $n \to \infty$ for fixed p, and as $p \to \infty$ for fixed n. A good deal of technical material is relegated to the Supplementary Material [Qin and Hobert (2018)].

2. Asymptotically stable drift and minorization. Let X be a set equipped with a countably generated σ-algebra $\mathcal{B}(\mathsf{X})$. Suppose that $K : \mathsf{X} \times \mathcal{B}(\mathsf{X}) \to [0, 1]$ is an Mtf with invariant probability measure $\Pi(\cdot)$, so that $\Pi(C) = \int_{\mathsf{X}} K(x, C)\Pi(\mathrm{d}x)$ for all $C \in \mathcal{B}(\mathsf{X})$. Assume that the corresponding Markov chain is Harris ergodic. Recall the definitions of geometric ergodicity and geometric convergence rate from Section 1. The following result has proven extremely useful for establishing geometric ergodicity in the context of Monte Carlo Markov chains used to study complex Bayesian posterior distributions.

THEOREM 1 [Rosenthal (1995)]. *Suppose that $K(x, \cdot)$ satisfies the drift condition*

$$\int_{\mathsf{X}} V(x')K(x, \mathrm{d}x') \le \lambda V(x) + L, \qquad x \in \mathsf{X} \tag{4}$$

for some $V : \mathsf{X} \to [0, \infty)$, $\lambda < 1$ and $L < \infty$. Suppose that it also satisfies the minorization condition

$$K(x, \cdot) \ge \varepsilon Q(\cdot) \qquad \text{whenever } V(x) < d \tag{5}$$

for some $\varepsilon > 0$, probability measure $Q(\cdot)$ on X, and $d > 2L/(1 - \lambda)$. Then assuming the chain is started according to the probability measure $\nu(\cdot)$, for any $0 < r < 1$, we have

$$\begin{aligned}\left\|\nu K^m(\cdot) - \Pi(\cdot)\right\|_{\mathrm{TV}} \le{}& (1 - \varepsilon)^{rm} \\ &+ \left(1 + \frac{L}{1 - \lambda} + \int_{\mathsf{X}} V(x)\nu(\mathrm{d}x)\right) \\ &\times \left[\left(\frac{1 + 2L + \lambda d}{1 + d}\right)^{1-r} \{1 + 2(\lambda d + L)\}^r\right]^m.\end{aligned}$$

The function V is called the drift (or Lyapunov) function, and $\{x \in \mathsf{X} : V(x) < d\}$ is called the small set associated with V. We will refer to ε as the minorization number. Manipulation of the total variation bound in Theorem 1 leads to

$$\left\|\nu K^m(\cdot) - \Pi(\cdot)\right\|_{\mathrm{TV}} \le \left(2 + \frac{L}{1 - \lambda} + \int_{\mathsf{X}} V(x)\nu(\mathrm{d}x)\right)\hat{\rho}^m,$$

where

$$\hat{\rho} := (1-\varepsilon)^r \vee \left(\frac{1+2L+\lambda d}{1+d}\right)^{1-r} \{1+2(\lambda d+L)\}^r, \tag{6}$$

and $r \in (0,1)$ is arbitrary. Then $\hat{\rho}$ is an upper bound on the geometric convergence rate ρ_*. It is easy to verify that when $\lambda < 1$, $L < \infty$ and $\varepsilon > 0$, there exists $r \in (0,1)$ such that $\hat{\rho} < 1$.

The bound $\hat{\rho}$ has a reputation for being too conservative. This is partly due to the fact that there are toy examples where the true ρ_* is known, and $\hat{\rho}$ is quite far off [see, e.g., Rosenthal (1995)]. There also exist myriad analyses of practical Monte Carlo Markov chains where the d&m conditions (4) and (5) have been established (proving that the underlying chain is indeed geometrically ergodic), but the total variation bound of Theorem 1 is useless because $\hat{\rho}$ is so near unity. Of course, the quality of the bound $\hat{\rho}$ depends on the choice of drift function, and the sharpness of (4) and (5). Our results for the A&C chain suggest that poorly chosen drift functions and/or loose inequalities in the d&m conditions are to blame for (at least) some of the unsuccessful applications of Theorem 1. We now introduce the concept of asymptotically stable d&m.

Consider a sequence of geometrically ergodic Markov chains, $\{\Psi^{(j)}\}_{j=1}^{\infty}$, with corresponding geometric convergence rates given by $\rho_*^{(j)}$. (In practice, j is usually the sample size, n, or number of covariates, p.) We are interested in the asymptotic behavior of the rate sequence. For example, we might want to know if it is bounded away from 1. Suppose that for each chain, $\Psi^{(j)}$, we have d&m conditions defined through $\lambda^{(j)}$, $L^{(j)}$, and $\varepsilon^{(j)}$, and thus an upper bound on the convergence rate, $\hat{\rho}^{(j)} \in [\rho_*^{(j)}, 1)$. The following simple result (whose proof is left to the reader) provides conditions under which these upper bounds are *unstable*, that is, $\limsup_{j\to\infty} \hat{\rho}^{(j)} = 1$.

PROPOSITION 2. *Suppose that there exists a subsequence of* $\{\Psi^{(j)}\}_{j=0}^{\infty}$, *call it* $\{\Psi^{(j_l)}\}_{l=0}^{\infty}$, *that satisfies one or more of the following three conditions*: (i) $\lambda^{(j_l)} \to 1$, (ii) $L^{(j_l)} \to \infty$, *while* $\varepsilon^{(j_l)}$ *is bounded away from* 1 *and* $\lambda^{(j_l)}$ *is bounded away from* 0, (iii) $\varepsilon^{(j_l)} \to 0$. *Then the corresponding subsequence of the upper bounds*, $\hat{\rho}^{(j_l)}$, *converges to* 1.

If one (or more) of the conditions in Proposition 2 holds for some subsequence of $\{\Psi^{(j)}\}_{j=0}^{\infty}$, then we say that the d&m conditions are (asymptotically) unstable (in j). On the other hand, if (i′) $\lambda^{(j)}$ is bounded away from 1, (ii′) $L^{(j)}$ is bounded above and (iii′) $\varepsilon^{(j)}$ is bounded away from 0, then the sequence $\hat{\rho}^{(j)}$ can be bounded away from 1, thus giving an asymptotically nontrivial upper bound on $\rho_*^{(j)}$. We say that the drift conditions are stable if (i′) and (ii′) hold, and likewise, that the minorization conditions are stable if (iii′) holds.

Before moving on to describe the techniques that we use to develop stable d&m for the A&C chain, we note that elementary linear transformations of the drift function can affect the quality of $\hat{\rho}$, and even stability. It is easy to show that, while multiplying the drift function by a scale factor will affect L, it will not affect the quality of the minorization inequality (5) in any nontrivial way. Subtracting a positive number from V (while preserving its nonnegativity) will, on the other hand, always lead to an improved bound $\hat{\rho}$. Needless to say, we will only deal with instability that cannot be prevented by these trivial transformations. In particular, throughout the article, we consider only drift functions whose infimums are 0, that is, we make the elementary transformation $V(x) \mapsto V(x) - \inf_{x' \in \mathsf{X}} V(x')$.

To obtain stable d&m for the A&C chain, we will exploit the notion of "centered" drift functions. Theorem 1 is based on a coupling construction, in which two copies of the Markov chain coalesce with probability ε each time they (both) enter the small set. The total variation distance between the two chains at time m is then bounded above by 1 minus the probability of coalescence in m iterations. Thus, loosely speaking, we want the chains to visit the small set as often as possible, without making the small set too large. [Larger small sets usually result in smaller ε, as indicated by (5).] So, it makes sense to use a drift function that is centered in the sense that it takes small values in the part of the state space where the chain spends the bulk of its time. Of course, if the chain is well suited to the problem, then it should linger in the high posterior probability regions of X.

The idea of centering is not new, and has been employed without emphasis by many authors. In this article, we illustrate the importance of centering to stable d&m, especially when n is large. Indeed, in Section 5 it is shown that, for A&C's chain, in the large n, small p regime, the uncentered drift function employed by Roy and Hobert (2007) cannot possibly lead to stable d&m, while a centered version of the same drift function does. The intuition behind these results is as follows. By (4), $d > 2L/(1-\lambda) \geq 2\Pi V$, where $\Pi V := \int_{\mathsf{X}} V(x)\Pi(dx)$. Hence, ΠV controls the volume of the small set. In Bayesian models, as n increases, the posterior is likely to concentrate around a single point in the state space. Consider a sequence of posterior distributions and drift functions, $\{(\Pi^{(n)}, V^{(n)})\}$, such that $\Pi^{(n)}$ concentrates around a point x_0. Heuristically, we expect $\Pi^{(n)}V^{(n)}$ to be close to $V^{(n)}(x_0)$ for large n. Therefore, when n is large, if the drift functions are minimized at or near x_0, then we will have a better chance of controlling the volumes of the small sets, and bounding the minorization numbers away from 0.

Another technique we use to achieve stable d&m for the A&C chain is a dimension reduction trick that is designed specifically for two-block Gibbs samplers and data augmentation algorithms. We begin by describing a common difficulty encountered in the convergence analysis of such algorithms. Suppose that $\mathsf{X} \subset \mathbb{R}^p$, and $K(x, \cdot)$ is associated with a Markov transition density (Mtd) of the form

$$k(x, x') = \int_{\mathbb{R}^n} s(x'|z)h(z|x)\,\mathrm{d}z, \tag{7}$$

where n is the sample size, $z = (z_1\ z_2\ \cdots\ z_n)^T \in \mathbb{R}^n$ is a vector of latent data, and $s : \mathsf{X} \times \mathbb{R}^n \to [0, \infty)$ and $h : \mathbb{R}^n \times \mathsf{X} \to [0, \infty)$ are conditional densities associated with some joint density on $\mathsf{X} \times \mathbb{R}^n$. Assume that $h(z|x)$ can be factored as follows:

$$h(z|x) = \prod_{i=1}^{n} h_i(z_i|x),$$

where, for each i, $h_i : \mathbb{R} \times \mathsf{X} \to [0, \infty)$ is a univariate (conditional) pdf. Usually, $s(\cdot|z)$ and $h_i(\cdot|x)$ are tractable, but there is no closed form for $k(x, x')$. However, there is a well-known argument for establishing a minorization condition in this case. Suppose that whenever $V(x) < d$,

$$h_i(t|x) > \epsilon_i \nu_i(t), \qquad i = 1, 2, \ldots, n$$

where $\epsilon_i > 0$ and $\nu_i : \mathbb{R} \to [0, \infty)$ is a pdf. Then, whenever $V(x) < d$, we have

$$k(x, x') > \left(\prod_{i=1}^{n} \epsilon_i\right) \int_{\mathbb{R}^n} s(x'|z) \prod_{i=1}^{n} \nu_i(z_i)\, \mathrm{d}z. \tag{8}$$

Since $\int_{\mathbb{R}^n} s(x'|z) \prod_{i=1}^{n} \nu_i(z_i)\, \mathrm{d}z$ is a pdf on X, (8) gives a minorization condition with $\varepsilon = \prod_{i=1}^{n} \epsilon_i$. Unfortunately, this quantity will almost always converge to 0 as $n \to \infty$. Consequently, if our sequence of Markov chains are indexed by n, then we have unstable minorization. This problem is well known [see, e.g., Rajaratnam and Sparks (2015)].

The instability of the minorization described above is due to the fact that the dimension of z is growing with n. However, it is often the case that $s(x|z)$ depends on z only through $f(z)$, where $f : \mathbb{R}^n \to \mathsf{Y}$ is a function into some fixed space Y, say $\mathsf{Y} \subset \mathbb{R}^q$, where q does not depend on n. Then integrating along $f(z) = \gamma$ in (7) yields

$$k(x, x') = \int_{\mathsf{Y}} \tilde{s}(x'|\gamma)\tilde{h}(\gamma|x)\, \mathrm{d}\gamma, \tag{9}$$

where $\tilde{s}(x'|f(z)) = s(x'|z)$, and

$$\int_C \tilde{h}(\gamma|x)\, \mathrm{d}\gamma = \int_{\{z : f(z) \in C\}} h(z|x)\, \mathrm{d}z$$

for all $x \in \mathsf{X}$ and any measurable $C \subset \mathsf{Y}$. Note that this new representation of $k(x, x')$ no longer contains n explicitly, and the high dimensionality problem for z is resolved. However, we now have a new problem. Namely, $\tilde{h}(\gamma|x)$ is likely to be quite intractable. Fortunately, the following result provides a way to circumvent this difficulty.

PROPOSITION 3. *Assume that we have a drift condition for $k(x, x')$, that is,*

$$\int_{\mathsf{X}} V(x')k(x, x')\, \mathrm{d}x' \le \lambda V(x) + L, \tag{10}$$

where $V : \mathsf{X} \to [0, \infty)$, $\lambda \in [0, 1)$, *and* L *is finite. Assume further that* $k(x, x')$ *can be written in the form* (9). *Define a Mtd* $\tilde{k} : \mathsf{Y} \times \mathsf{Y} \to [0, \infty)$ *as follows*:

$$\tilde{k}(\gamma, \gamma') = \int_{\mathsf{X}} \tilde{h}(\gamma'|x)\tilde{s}(x|\gamma)\,\mathrm{d}x.$$

If $\tilde{V}(\gamma) = \int_{\mathsf{X}} V(x)\tilde{s}(x|\gamma)\,\mathrm{d}x + c$ *is finite and nonnegative for all* $\gamma \in \mathsf{Y}$, *where* c *is some constant, then the following drift condition holds for* $\tilde{k}$,

$$\int_{\mathsf{Y}} \tilde{V}(\gamma')\tilde{k}(\gamma, \gamma')\,\mathrm{d}\gamma' \le \lambda \tilde{V}(\gamma) + \tilde{L}, \tag{11}$$

where $\tilde{L} = L + c(1 - \lambda)$.

PROOF. Our assumptions imply that

$$\int_{\mathsf{X}} V(x') \int_{\mathsf{Y}} \tilde{s}(x'|\gamma')\tilde{h}(\gamma'|x)\,\mathrm{d}\gamma'\,\mathrm{d}x' \le \lambda V(x) + L.$$

Multiplying both sides of the inequality by $\tilde{s}(x|\gamma)$, and integrating out x yields the result. □

REMARK 4. Note that if we set $c \le 0$ (while preserving the nonnegativity of $\tilde{V}$) in the above proposition, then (11) is stable whenever the original drift (10) is stable.

As we now explain, Proposition 3 has important implications. Indeed, it is well known that the Markov chains driven by k and $\tilde{k}$ (which we call the "flipped" version of k) share the same geometric convergence rate [see, e.g., Diaconis, Khare and Saloff-Coste (2008), Roberts and Rosenthal (2001)]. Thus, we can analyze k indirectly through the flipped chain. Now, as mentioned above, $\tilde{s}(x|f(z)) = s(x|z)$ is often tractable. Suppose that there exists some $\tilde{\varepsilon} > 0$ and pdf $\tilde{\nu} : \mathsf{X} \to [0, \infty)$ such that $\tilde{s}(x|\gamma) \ge \tilde{\varepsilon}\tilde{\nu}(x)$ when $\tilde{V}(\gamma) < \tilde{d}$, where $\tilde{d} > 2\tilde{L}/(1 - \lambda)$. Then we have the following minorization condition for the flipped chain:

$$\tilde{k}(\gamma, \gamma') \ge \tilde{\varepsilon} \int_{\mathsf{X}} \tilde{h}(\gamma'|x)\tilde{\nu}(x)\,\mathrm{d}x \qquad \text{whenever } \tilde{V}(\gamma) < \tilde{d},$$

which is stable as long as $\tilde{\varepsilon}$ is bounded away from 0 as $n \to \infty$. This, along with (11), allows us to construct potentially stable bounds on $\rho_*^{(n)}$ for the flipped chains, and thus for the original chains on X as well. This is exactly how we analyze A&C's chain in the large n, small p regime. It turns out that the flipped chain argument can also be used to establish d&m conditions that are stable in p, and we will exploit this in our analysis of A&C's chain in the large p, small n regime.

We end this section with a result that allows us to use information about a flipped chain to get total variation bounds for the original. The following result follows immediately from Proposition 27, which is stated and proven in Section 7.1 of the Supplementary Material [Qin and Hobert (2018)].

COROLLARY 5. *Suppose we have d&m conditions for a flipped chain (which is driven by $\tilde{k}$). Let λ, $\tilde{L}$ and $\tilde{\varepsilon}$ denote the drift and minorization parameters, and let $\hat{\rho}_f$ denote the corresponding bound on the geometric convergence rate obtained through Theorem 1. Then, if the original chain is started at $x \in \mathsf{X}$, we have, for $m \geq 1$,*

$$\|K^m(x,\cdot) - \Pi(\cdot)\|_{\mathrm{TV}} \leq \left(2 + \frac{\tilde{L}}{1-\lambda} + \int_{\Upsilon} \tilde{V}(\gamma)\tilde{h}(\gamma|x)\,\mathrm{d}\gamma\right)\hat{\rho}_f^{m-1}.$$

In the next section, we begin our analysis of the A&C chain.

3. Albert and Chib's Markov chain and the centered drift functions.

3.1. *Basics.* Let $\{X_i\}_{i=1}^n$, $\{Y_i\}_{i=1}^n$, and $B \in \mathbb{R}^p$ be defined as in the Introduction, so that $Y_i|X_i, B \sim \text{Bernoulli}(\Phi(X_i^T B))$ independently. Suppose that, having observed the data, $\mathcal{D} := \{(X_i, Y_i)\}_{i=1}^n$, we wish to perform a Bayesian analysis using a prior density for B given by (2). Recall that X and Y denote, respectively, the design matrix and vector of responses. The posterior density (3) is proper precisely when

$$\int_{\mathbb{R}^p} \prod_{i=1}^n \bigl(\Phi(X_i^T\beta)\bigr)^{Y_i}\bigl(1-\Phi(X_i^T\beta)\bigr)^{1-Y_i}\omega(\beta)\,d\beta < \infty.$$

When Q is positive definite, $\omega(\beta)$ is a proper normal density, and the posterior is automatically proper. If $Q = 0$, then propriety is not guaranteed. Define X_* as the $n \times p$ matrix whose ith row is $-X_i^T$ if $Y_i = 1$, and X_i^T if $Y_i = 0$. Chen and Shao (2001) proved that when the prior is flat, that is, $Q = 0$, the following two conditions are necessary and sufficient for posterior propriety:

(C1) X has full column rank;

(C2) There exists a vector $a = (a_1\ a_2\ \cdots\ a_n)^T \in \mathbb{R}^n$ such that $a_i > 0$ for all i, and $X_*^T a = 0$.

Until further notice, we will assume that the posterior is proper.

A&C's algorithm to draw from the intractable posterior is based on the following latent data model. Given X and B, let $\{(Y_i, Z_i)\}_{i=1}^n$ be a sequence of independent random vectors such that

$$Y_i|Z_i, X, B \text{ is a point mass at } 1_{\mathbb{R}_+}(Z_i),$$

$$Z_i|X, B \sim N(X_i^T B, 1).$$

Clearly, under this hierarchical structure, $Y_i|X_i, B \sim \text{Bernoulli}(\Phi(X_i^T B))$, which is consistent with the original model. Thus, if we let $\pi_{B,Z|Y,X}(\beta, z|Y, X)$ denote the corresponding (augmented) posterior density [where $Z = (Z_1\ Z_2\ \cdots\ Z_n)^T$], then it is clear that

$$\int_{\mathbb{R}^n} \pi_{B,Z|Y,X}(\beta, z|Y, X)\,\mathrm{d}z = \pi_{B|Y,X}(\beta|Y, X),$$

which is the target posterior from (3). Albert and Chib's algorithm is simply a two-variable Gibbs sampler based on $\pi_{B,Z|Y,X}(\beta, Z|Y, X)$. Indeed, the Mtd, $k_{\mathrm{AC}} : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$, is defined as

$$k_{\mathrm{AC}}(\beta, \beta') := k_{\mathrm{AC}}(\beta, \beta'; Y, X) = \int_{\mathbb{R}^n} \pi_{B|Z,Y,X}(\beta'|z, Y, X)\pi_{Z|B,Y,X}(z|\beta, Y, X)\,\mathrm{d}z.$$

As pointed out by Albert and Chib (1993),

$$B|Z, Y, X \sim \mathrm{N}_p\big(\Sigma^{-1}(X^T Z + Qv), \Sigma^{-1}\big),$$

where, again, $\Sigma = X^T X + Q$. Moreover, the density $\pi_{Z|B,Y,X}(z|\beta, Y, X)$ is a product of n univariate densities, where

$$Z_i|B, Y, X \sim \mathrm{TN}\big(X_i^T B, 1; Y_i\big).$$

Obviously, these are the conditional densities that appear in the algorithm described in the Introduction.

3.2. *A centered drift function.* Roy and Hobert (2007) and Chakraborty and Khare (2017) both used the drift function $V_0(\beta) = \|\Sigma^{1/2}\beta\|^2$. While this drift function is certainly amenable to analysis, it is not "centered" in any practical sense. Indeed, $V_0(\beta)$ takes on its minimum when $\beta = 0$, but, in general, there is no reason to expect A&C's chain to make frequent visits to the vicinity of the origin. This heuristic is borne out by the result in Section 5 showing that V_0 *cannot* lead to stable d&m in the large n, small p regime. As an alternative to $V_0(\beta)$, we consider drift functions of the form

$$V(\beta) = \big\|M(\beta - \beta^*)\big\|^2, \tag{12}$$

where $M = M(X, Y)$ is a matrix with p columns, and $\beta^* = \beta^*(X, Y)$ is a point in $\mathbb{R}^p$ that is "attractive" to A&C's chain. A candidate for β^* would be the posterior mode $\hat{B}$, which uniquely exists because of the well-known fact that the posterior density $\pi_{B|Y,X}$ is log-concave. Setting $\beta^* = \hat{B}$ is, of course, not the only viable centering scheme, and any β^* in a close vicinity of $\hat{B}$ would be equally effective. However, the following proposition shows that the posterior mode has a nice feature that will be exploited in the sequel.

PROPOSITION 6. *The posterior mode,* $\hat{B}$, *satisfies the following equation*:

$$\int_{\mathbb{R}^p} \beta k_{\mathrm{AC}}(\hat{B}, \beta)\,\mathrm{d}\beta = \hat{B}. \tag{13}$$

PROOF. $\hat{B}$ is the solution to the following equation:

$$\frac{d}{d\beta}\big(\log\omega(\beta) + \log\pi_{Y|B,X}(Y|\beta, X)\big) = 0.$$

This implies that

$$\sum_{i=1}^{n}\left(\frac{\phi(X_i^T\hat{B})}{\Phi(X_i^T\hat{B})}1_{\{1\}}(Y_i)-\frac{\phi(X_i^T\hat{B})}{1-\Phi(X_i^T\hat{B})}1_{\{0\}}(Y_i)\right)X_i-(Q\hat{B}-Qv)=0, \tag{14}$$

where $\phi(\cdot)$ is the pdf of the standard normal distribution. On the other hand, it follows from (27) in Section 6.2 of the Supplementary Material [Qin and Hobert (2018)] that

$$\mathbb{E}(Z_i|B=\hat{B},Y,X)=X_i^T\hat{B}+\frac{\phi(X_i^T\hat{B})}{\Phi(X_i^T\hat{B})}1_{\{1\}}(Y_i)-\frac{\phi(X_i^T\hat{B})}{1-\Phi(X_i^T\hat{B})}1_{\{0\}}(Y_i).$$

This, along with (14), implies that

$$Qv+\sum_{i=1}^{n}X_i\mathbb{E}(Z_i|B=\hat{B},Y,X)=\Sigma\hat{B}.$$

But this is equivalent to

$$\int_{\mathbb{R}^p}\int_{\mathbb{R}^n}\beta'\pi_{B|Z,Y,X}(\beta'|z,Y,X)\pi_{Z|B,Y,X}(z|\hat{B},Y,X)\,\mathrm{d}z\,\mathrm{d}\beta'=\hat{B},$$

which is precisely (13). □

REMARK 7. We should emphasize that (13), while interesting, is not essential to the proofs of our main results. It merely simplifies the process of establishing a drift condition.

We will consider two different versions of (12), both centered at $\hat{B}$. The first, which will be used in the large n-small p regime, is simply a centered version of V_0 given by

$$V_1(\beta)=\|\Sigma^{1/2}(\beta-\hat{B})\|^2.$$

In the large p-small n regime, we assume that Q is positive definite (which is necessary for posterior propriety) and that X is full row rank, and we use the following drift function:

$$V_2(\beta)=\|(X\Sigma^{-1}X^T)^{-1/2}X(\beta-\hat{B})\|^2.$$

In the next section, we establish two sets of d&m conditions for the A&C chain based on V_1 and V_2.

4. Results for the Albert and Chib chain Part I: Fixed n and p.

4.1. *Drift inequalities for V_1 and V_2.* Define $g : \mathbb{R} \to \mathbb{R}$ as

$$g(\theta) = \frac{\theta\phi(\theta)}{\Phi(\theta)} + \left(\frac{\phi(\theta)}{\Phi(\theta)}\right)^2.$$

For any $\beta \in \mathbb{R}^p$, let $D(\beta)$ denote an $n \times n$ diagonal matrix with ith diagonal element

$$1 - g(X_i^T\beta)1_{\{1\}}(Y_i) - g(-X_i^T\beta)1_{\{0\}}(Y_i).$$

LEMMA 8. *If $V(\beta) = \|M(\beta - \hat{B})\|^2$, $\beta \in \mathbb{R}^p$, where M is any matrix with p columns, then*

$$\int_{\mathbb{R}^p} V(\beta')k_{\mathrm{AC}}(\beta, \beta')\,\mathrm{d}\beta' \le \sup_{t\in(0,1)} \left\|M\Sigma^{-1}X^T D\big(\hat{B} + t(\beta - \hat{B})\big)X(\beta - \hat{B})\right\|^2 + 2\operatorname{tr}\big(M\Sigma^{-1}M^T\big).$$

PROOF. Note that

$$\int_{\mathbb{R}^p} V(\beta')\pi_{B|Z,Y,X}(\beta'|z, Y, X)\,\mathrm{d}\beta' = \left\|M\Sigma^{-1}\big(X^T z + Qv\big) - M\hat{B}\right\|^2 + \operatorname{tr}\big(M\Sigma^{-1}M^T\big).$$

Moreover,

$$\begin{aligned}\int_{\mathbb{R}^n} \left\|M\Sigma^{-1}\big(X^T z + Qv\big) - M\hat{B}\right\|^2 \pi_{Z|B,Y,X}(z|\beta, Y, X)\,\mathrm{d}z &= \left\|M\Sigma^{-1}\big\{X^T\mathbb{E}(Z|B = \beta, Y, X) + Qv\big\} - M\hat{B}\right\|^2 \\ &\quad + \operatorname{tr}\big\{M\Sigma^{-1}X^T \operatorname{var}(Z|B = \beta, Y, X)X\Sigma^{-1}M^T\big\}. \end{aligned} \tag{15}$$

For two symmetric matrices of the same size, M_1 and M_2, we write $M_1 \le M_2$ if $M_2 - M_1$ is nonnegative definite. By Lemma 26 in Section 6.2 of the Supplementary Material [Qin and Hobert (2018)], $\operatorname{var}(Z|B = \beta, Y, X) \le I_n$. It follows that

$$\begin{aligned} M\Sigma^{-1}X^T \operatorname{var}(Z|B = \beta, Y, X)X\Sigma^{-1}M^T &\le M\Sigma^{-1}X^T X\Sigma^{-1}M^T \le M\Sigma^{-1}M^T. \end{aligned} \tag{16}$$

Therefore,

$$\begin{aligned}\int_{\mathbb{R}^p} V(\beta')k_{\mathrm{AC}}(\beta, \beta')\,\mathrm{d}\beta' &= \int_{\mathbb{R}^p} V(\beta') \int_{\mathbb{R}^n} \pi_{B|Z,Y,X}(\beta'|z, Y, X)\pi_{Z|B,Y,X}(z|\beta, Y, X)\,\mathrm{d}z\,\mathrm{d}\beta' \\ &\le \left\|M\Sigma^{-1}\big\{X^T\mathbb{E}(Z|B = \beta, Y, X) + Qv\big\} - M\hat{B}\right\|^2 + 2\operatorname{tr}\big(M\Sigma^{-1}M^T\big). \end{aligned} \tag{17}$$

Now, for $\alpha \in \mathbb{R}^p$, define

$$\mu(\alpha) = M\Sigma^{-1}\{X^T\mathbb{E}(Z|B = \hat{B} + \alpha, Y, X) + Qv\} - M\hat{B}.$$

By Proposition 6, we have

$$\mu(0) = M\int_{\mathbb{R}^p} \beta k_{\mathrm{AC}}(\hat{B}, \beta)\,\mathrm{d}\beta - M\hat{B} = 0.$$

By the mean value theorem for vector-valued functions [see, e.g., Rudin (1976), Theorem 5.19], for any $\alpha \in \mathbb{R}^p$,

$$\|\mu(\alpha)\|^2 \le \Big(\sup_{t\in(0,1)} \Big\| \frac{\partial \mu(t\alpha)}{\partial t} \Big\|\Big)^2.$$

Now, by results on truncated normal distributions in Section 6.2 of the Supplementary Material [Qin and Hobert (2018)],

$$\begin{aligned}
\frac{\partial \mu(t\alpha)}{\partial t} &= M\Sigma^{-1}\sum_{i=1}^n X_i \frac{\partial}{\partial t}\mathbb{E}(Z_i|B = \hat{B} + t\alpha, Y, X) \\
&= M\Sigma^{-1}\sum_{i=1}^n X_i \frac{\partial}{\partial t}\{X_i^T(\hat{B} + t\alpha)\} \\
&\quad \times \frac{d}{d\theta}\bigg(\theta + 1_{\{1\}}(Y_i)\frac{\phi(\theta)}{\Phi(\theta)} - 1_{\{0\}}(Y_i)\frac{\phi(\theta)}{1-\Phi(\theta)}\bigg)\bigg|_{\theta = X_i^T(\hat{B}+t\alpha)} \\
&= M\Sigma^{-1}X^T D(\hat{B} + t\alpha)X\alpha.
\end{aligned}$$

Hence,

$$\|\mu(\alpha)\|^2 \le \sup_{t\in(0,1)} \|M\Sigma^{-1}X^T D(\hat{B} + t\alpha)X\alpha\|^2. \tag{18}$$

The result then follows from (17) and (18) by taking $\alpha = \beta - \hat{B}$. □

We now use Lemma 8 to establish explicit drift inequalities for V_1 and V_2. We begin with V_1.

PROPOSITION 9. *For all $\beta \in \mathbb{R}^p$, we have*

$$\begin{aligned}
&\int_{\mathbb{R}^p} V_1(\beta')k_{\mathrm{AC}}(\beta, \beta')\,\mathrm{d}\beta' \\
&\quad \le \Big(\sup_{t\in(0,1)}\sup_{\alpha\neq 0} \frac{\|\Sigma^{-1/2}X^T D(\hat{B} + t\alpha)X\alpha\|^2}{\|\Sigma^{1/2}\alpha\|^2}\Big)V_1(\beta) + 2p.
\end{aligned}$$

PROOF. Taking $M = \Sigma^{1/2}$ in Lemma 8 gives the result. □

In Section 7.2 of the Supplementary Material [Qin and Hobert (2018)], we prove that

$$\sup_{t\in(0,1)}\sup_{\alpha\neq 0}\frac{\|\Sigma^{-1/2}X^T D(\hat{B}+t\alpha)X\alpha\|^2}{\|\Sigma^{1/2}\alpha\|^2}<1.$$

For a symmetric matrix M, let $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ denote the smallest and largest eigenvalues of M, respectively. Here is the analogue of Proposition 9 for V_2.

PROPOSITION 10. *Assume that X has full row rank. Then for all $\beta\in\mathbb{R}^p$, we have*

$$\int_{\mathbb{R}^p} V_2(\beta')k_{\mathrm{AC}}(\beta,\beta')\,\mathrm{d}\beta' \le \{\lambda^2_{\max}(X\Sigma^{-1}X^T)\}V_2(\beta)+2n.$$

PROOF. Taking $M=(X\Sigma^{-1}X^T)^{-1/2}X$ in Lemma 8 and applying Lemma 26 yields

$$\begin{aligned}
&\int_{\mathbb{R}^p} V_2(\beta')k_{\mathrm{AC}}(\beta,\beta')\,\mathrm{d}\beta'\\
&\quad\le \sup_{t\in(0,1)}\big\|(X\Sigma^{-1}X^T)^{1/2}D\big(\beta+t(\beta-\hat{B})\big)X(\beta-\hat{B})\big\|^2+2n\\
&\quad\le \sup_{t\in(0,1)}\lambda^2_{\max}\big\{(X\Sigma^{-1}X^T)^{1/2}D\big(\hat{B}+t(\beta-\hat{B})\big)(X\Sigma^{-1}X^T)^{1/2}\big\}V_2(\beta)+2n\\
&\quad\le \{\lambda^2_{\max}(X\Sigma^{-1}X^T)\}V_2(\beta)+2n.
\end{aligned}$$

□

4.2. *Drift and minorization for the Albert and Chib chain based on V_1.* In this subsection, we exploit the flipped chain idea described in Section 2. In particular, V_1 is used to establish d&m conditions for a flipped chain that has the same geometric convergence rate as A&C's chain. Later, in Section 5, we will use these results to prove asymptotic stability as $n\to\infty$.

Note that $\pi_{B|Z,Y,X}(\beta|Z,Y,X)$ depends on the n-dimensional vector Z only through X^TZ, which is a one-to-one function of the following p-dimensional vector:

$$\Gamma := \Sigma^{1/2}\{\Sigma^{-1}(X^TZ+Qv)-\hat{B}\}.$$

Hence, we can represent the Mtd of the A&C chain as follows:

$$k_{\mathrm{AC}}(\beta,\beta')=\int_{\mathbb{R}^p}\pi_{B|\Gamma,Y,X}(\beta'|\gamma,Y,X)\pi_{\Gamma|B,Y,X}(\gamma|\beta,Y,X)\,\mathrm{d}\gamma.$$

Recalling the discussion in Section 2, this maneuver seems to represent progress since we have replaced n with p. However, it is difficult to establish a minorization

condition using this version of $k_{\mathrm{AC}}(\beta, \beta')$ because $\pi_{\Gamma|B,Y,X}(\gamma|\beta, Y, X)$ lacks a closed form. On the other hand, this chain has the same geometric convergence rate as the flipped chain defined by the following Mtd:

$$\begin{aligned}\tilde{k}_{\mathrm{AC}}(\gamma, \gamma') &:= \tilde{k}_{\mathrm{AC}}(\gamma, \gamma'; Y, X)\\ &= \int_{\mathbb{R}^p} \pi_{\Gamma|B,Y,X}(\gamma'|\beta, Y, X)\pi_{B|\Gamma,Y,X}(\beta|\gamma, Y, X)\,\mathrm{d}\beta.\end{aligned}$$

Constructing a minorization condition for this Mtd is much less daunting since

$$B|\Gamma, Y, X \sim N\bigl(\Sigma^{-1/2}\Gamma + \hat{B}, \Sigma^{-1}\bigr).$$

Here is the main result of this subsection.

PROPOSITION 11. *Let $\tilde{V}_1(\gamma) = \|\gamma\|^2$. The Mtd $\tilde{k}_{\mathrm{AC}}$ satisfies the drift condition*

$$\int_{\mathbb{R}^p} \tilde{V}_1(\gamma')\tilde{k}_{\mathrm{AC}}(\gamma, \gamma')\,\mathrm{d}\gamma' \le \lambda\tilde{V}_1(\gamma) + L,$$

where $L = p(1 + \lambda)$, and

$$\lambda = \sup_{t\in(0,1)}\sup_{\alpha\neq 0} \frac{\|\Sigma^{-1/2}X^T D(\hat{B} + t\alpha)X\alpha\|^2}{\|\Sigma^{1/2}\alpha\|^2}.$$

Moreover, for $d > 2L/(1 - \lambda)$, $\tilde{k}_{\mathrm{AC}}$ satisfies

$$\tilde{k}_{\mathrm{AC}}(\gamma, \gamma') \ge \varepsilon q(\gamma'),$$

where $q : \mathbb{R}^p \to [0, \infty)$ is a pdf, and $\varepsilon = 2^{-p/2}e^{-d}$.

PROOF. We begin with the drift. It's easy to verify that

$$\tilde{V}_1(\gamma) = \int_{\mathbb{R}^p} V_1(\beta)\pi_{B|\Gamma,Y,X}(\beta|\gamma, Y, X)\,\mathrm{d}\beta - p.$$

We now use the techniques described at the end of Section 2 to convert the drift inequality in Proposition 9 into a drift inequality for the flipped chain. We know that

$$\int_{\mathbb{R}^p} V_1(\beta')k_{\mathrm{AC}}(\beta, \beta')\,\mathrm{d}\beta' \le \lambda V_1(\beta) + 2p.$$

Then taking $c = -p$ in Proposition 3 yields

$$\int_{\mathbb{R}^p} \tilde{V}_1(\gamma')\tilde{k}_{\mathrm{AC}}(\gamma, \gamma')\,\mathrm{d}\gamma' \le \lambda\tilde{V}_1(\gamma) + p(1 + \lambda).$$

As explained in Section 2, to establish the minorization condition, it suffices to show that there exists a pdf $\nu(\beta) := \nu(\beta|Y, X)$ such that

$$\pi_{B|\Gamma,Y,X}(\beta|\gamma, Y, X) \ge \varepsilon\nu(\beta) \tag{19}$$

whenever $\tilde{V}_1(\gamma) \le d$. Recall that

$$B|\Gamma, Y, X \sim N\big(\Sigma^{-1/2}\Gamma + \hat{B}, \Sigma^{-1}\big).$$

Define

$$\begin{aligned}\nu_1(\beta) &= \inf_{\gamma:\tilde{V}_1(\gamma)\le d} \pi_{B|\Gamma,Y,X}(\beta|\gamma, Y, X)\\ &= \inf_{\gamma:\tilde{V}_1(\gamma)\le d} \frac{|\Sigma|^{1/2}}{(2\pi)^{p/2}} \exp\biggl\{-\frac{1}{2}\|\Sigma^{1/2}(\beta - \hat{B} - \Sigma^{-1/2}\gamma)\|^2\biggr\}.\end{aligned}$$

Then $\nu(\beta) = \nu_1(\beta)/\int_{\mathbb{R}^p} \nu_1(\beta')\,d\beta'$ is a pdf, and whenever $\tilde{V}_1(\gamma) \le d$,

$$\pi_{B|\Gamma,Y,X}(\beta|\gamma, Y, X) \ge \biggl(\int_{\mathbb{R}^p} \nu_1(\beta')\,d\beta'\biggr)\nu(\beta).$$

This is (19) with

$$\varepsilon = \int_{\mathbb{R}^p} \nu_1(\beta)\,d\beta = (2\pi)^{-p/2}\int_{\mathbb{R}^p} \inf_{\gamma:\|\gamma\|^2\le d} \exp\biggl(-\frac{1}{2}\|\beta - \gamma\|^2\biggr)\,d\beta.$$

Finally, since $\|\beta - \gamma\|^2 \le 2(\|\beta\|^2 + \|\gamma\|^2)$,

$$\varepsilon \ge (2\pi)^{-p/2}\int_{\mathbb{R}^p} \inf_{\gamma:\|\gamma\|^2\le d} \exp\bigl(-\|\beta\|^2 - \|\gamma\|^2\bigr)\,d\beta = 2^{-p/2}e^{-d}.$$

□

Mainly, Proposition 11 will be used to establish asymptotic stability results for A&C's chain in the large n, small p regime. Indeed, since the Markov chains defined by k_{AC} and $\tilde{k}_{\mathrm{AC}}$ have the same geometric convergence rate, $\hat{\rho}$ calculated using (6) with λ, L, and ε from Proposition 11 is an upper bound on $\rho_* = \rho_*(X, Y)$ for the A&C chain. On the other hand, Proposition 11 can also be used in conjunction with Theorem 1 and Corollary 5 to get computable bounds on the total variation distance to stationarity for the A&C chain for fixed n and p. In order to state the result, we require a bit of notation. For an integer $m \ge 1$, let $k_{\mathrm{AC}}^{(m)} : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$ be the chain's m-step Mtd. For $\beta \in \mathbb{R}^p$, let $\varphi(\beta) = \mathbb{E}(Z|B = \beta, Y, X) \in \mathbb{R}^n$. Then results in Section 6.2 of the Supplementary Material [Qin and Hobert (2018)] show that the ith element of $\varphi(\beta)$ is given by

$$X_i^T\beta + \frac{\phi(X_i^T\beta)}{\Phi(X_i^T\beta)}1_{\{1\}}(Y_i) - \frac{\phi(X_i^T\beta)}{1 - \Phi(X_i^T\beta)}1_{\{0\}}(Y_i).$$

PROPOSITION 12. *If $\hat{\rho}$ is calculated using (6) with λ, L, and ε from Proposition 11, then for $m \ge 1$ and $\beta \in \mathbb{R}^p$,*

$$\int_{\mathbb{R}^p} \bigl|k_{\mathrm{AC}}^{(m)}(\beta, \beta'; Y, X) - \pi_{B|Y,X}(\beta'|Y, X)\bigr|\,d\beta' \le H(\beta)\hat{\rho}^{m-1},$$

where

$$H(\beta) = 2 + \frac{L}{1-\lambda} + \mathrm{tr}\bigl(X\Sigma^{-1}X^T\bigr) + \|\Sigma^{1/2}\{\Sigma^{-1}(X^T\varphi(\beta) + Qv) - \hat{B}\}\|^2.$$

PROOF. We simply apply Corollary 5. Putting $M = \Sigma^{1/2}$ in (15), we have

$$\begin{aligned}
&\int_{\mathbb{R}^p} \tilde{V}_1(\gamma)\pi_{\Gamma|B,Y,X}(\gamma|\beta, Y, X)\,\mathrm{d}\gamma \\
&\quad = \int_{\mathbb{R}^n} \bigl\|\Sigma^{1/2}\bigl\{\Sigma^{-1}\bigl(X^T z + Qv\bigr) - \hat{B}\bigr\}\bigr\|^2 \pi_{Z|B,Y,X}(z|\beta, Y, X)\,\mathrm{d}z \\
&\quad = \bigl\|\Sigma^{1/2}\bigl\{\Sigma^{-1}\bigl(X^T\varphi(\beta) + Qv\bigr) - \hat{B}\bigr\}\bigr\|^2 \\
&\qquad + \mathrm{tr}\bigl\{\Sigma^{-1/2}X^T \operatorname{var}(Z|B = \beta, Y, X)X\Sigma^{-1/2}\bigr\}.
\end{aligned}$$

A calculation similar to (16) shows that

$$\mathrm{tr}\bigl\{\Sigma^{-1/2}X^T \operatorname{var}(Z|B = \beta, Y, X)X\Sigma^{-1/2}\bigr\} \le \mathrm{tr}\bigl(X\Sigma^{-1}X^T\bigr),$$

and the result follows. □

Calculating λ in Proposition 11 calls for maximization of a function on $(0, 1) \times \mathbb{R}^p$, which may be difficult. Here, we provide an upper bound on λ that is easy to compute when p is small. Let $\{S_j\}_{j=1}^{2^p}$ denote the open orthants of $\mathbb{R}^p$. For instance, if $p = 2$, then S_1, S_2, S_3 and S_4 are the open quadrants of the real plane. Define $W(S_j) = W(S_j; X, Y)$ as follows:

$$W(S_j) = \sum_{X_i \in S_j} X_i 1_{\{0\}}(Y_i)X_i^T + \sum_{X_i \in -S_j} X_i 1_{\{1\}}(Y_i)X_i^T.$$

The following result is proven in Section 7.3 of the Supplementary Material [Qin and Hobert (2018)].

PROPOSITION 13. *An upper bound on $\lambda^{1/2}$ in Proposition 11 is*

$$\lambda_{\max}\bigl(\Sigma^{-1/2}X^T X\Sigma^{-1/2}\bigr) - \frac{2}{\pi}\min_{1\le j\le 2^p} \lambda_{\min}\bigl(\Sigma^{-1/2}W(S_j)\Sigma^{-1/2}\bigr).$$

If this upper bound is strictly less than 1 (which is always true when Q is positive definite), then one can replace λ with the square of this bound in Proposition 11.

4.3. *Drift and minorization for the Albert and Chib chain based on V_2.* The A&C chain has the same convergence rate as the flipped chain defined by the following Mtd:

$$\begin{aligned}
\check{k}_{\mathrm{AC}}(z, z') &:= \check{k}_{\mathrm{AC}}(z, z'; Y, X) \\
&= \int_{\mathbb{R}^p} \pi_{Z|B,Y,X}\bigl(z'|\beta, Y, X\bigr)\pi_{B|Z,Y,X}(\beta|z, Y, X)\,\mathrm{d}\beta.
\end{aligned}$$

In this subsection, we use V_2 to establish d&m conditions for this chain, and these will be used later to prove asymptotic stability as $p \to \infty$. First, for $z \in \mathbb{R}^n$, define

$$w(z) = \bigl(X\Sigma^{-1}X^T\bigr)^{-1/2} X\bigl\{\Sigma^{-1}\bigl(X^T z + Qv\bigr) - \hat{B}\bigr\}.$$

Now define $\check{V}_2 : \mathbb{R}^n \to [0, \infty)$ as $\check{V}_2(z) = \|w(z)\|^2$.

PROPOSITION 14. *The Mtd $\check{k}_{\mathrm{AC}}$ satisfies the drift condition*

$$\int_{\mathbb{R}^n} \check{V}_2(z')\check{k}_{\mathrm{AC}}(z, z')\,\mathrm{d}z' \le \lambda \check{V}_2(z) + L, \tag{20}$$

where $\lambda = \lambda^2_{\max}\{X\Sigma^{-1}X^T\}$ and $L = n(1+\lambda)$. Moreover, for $d > 2L/(1-\lambda)$, $\check{k}_{\mathrm{AC}}$ satisfies

$$\check{k}_{\mathrm{AC}}(z, z') \ge \varepsilon q(z),$$

where $q : \mathbb{R}^n \to [0, \infty)$ is a pdf, and $\varepsilon = 2^{-n/2}e^{-d}$.

PROOF. It is easy to verify that

$$\check{V}_2(z) = \int_{\mathbb{R}^p} V_2(\beta')\pi_{B|Z,Y,X}(\beta'|z, Y, X)\,\mathrm{d}\beta' - n.$$

We know from Proposition 10 that

$$\begin{aligned}\int_{\mathbb{R}^p} V_2(\beta')\biggl\{\int_{\mathbb{R}^n} \pi_{B|Z,Y,X}(\beta'|z', Y, X)\pi_{Z|B,Y,X}(z'|\beta, Y, X)\,\mathrm{d}z'\biggr\}\mathrm{d}\beta' \\ \le \lambda V_2(\beta) + 2n.\end{aligned}$$

As in Proposition 3, multiplying both sides of the above inequality by the conditional density $\pi_{B|Z,Y,X}(\beta|z, Y, X)$ and integrating with respect to β yields (20).

We now move on the the minorization condition. Note that $\pi_{Z|B,Y,X}(z|B, Y, X)$ depends on B only through XB, which is a one-to-one function of the n-dimensional vector

$$A := \bigl(X\Sigma^{-1}X^T\bigr)^{-1/2} X(B - \hat{B}).$$

Hence, $\check{k}_{\mathrm{AC}}(z, z')$ can be reexpressed as

$$\check{k}_{\mathrm{AC}}(z, z') = \int_{\mathbb{R}^n} \pi_{Z|A,Y,X}(z'|\alpha, Y, X)\pi_{A|Z,Y,X}(\alpha|z, Y, X)\,\mathrm{d}\alpha,$$

where $A|Z, Y, X \sim N(w(Z), I_n)$. To get the minorization condition, we will construct a pdf $\nu(\alpha) := \nu(\alpha|Y, X)$ such that

$$\pi_{A|Z,Y,X}(\alpha|z, Y, X) \ge \varepsilon\nu(\alpha)$$

whenever $\check{V}_2(z) \le d$. Define

$$\nu_1(\alpha) = \inf_{z:\check{V}_2(z)\le d} \pi_{A|Z,Y,X}(\alpha|z, Y, X) = \inf_{w:w^2\le d} (2\pi)^{-n/2} \exp\biggl(-\frac{1}{2}\|\alpha - w\|^2\biggr).$$

Then $\nu(\alpha) = \nu_1(\alpha)/\int_{\mathbb{R}^n} \nu_1(\alpha')\,d\alpha'$ is a pdf, and

$$\varepsilon = (2\pi)^{-n/2} \int_{\mathbb{R}^n} \inf_{w: w^2 \le d} \exp\left(-\frac{1}{2}\|\alpha - w\|^2\right) d\alpha$$
$$\ge (2\pi)^{-n/2} e^{-d} \int_{\mathbb{R}^n} \exp(-\|\alpha\|^2)\,d\alpha = 2^{-n/2} e^{-d}. \qquad \square$$

The next result is the analogue of Proposition 12. The proof is omitted as it is essentially the same as the proof of the said proposition.

PROPOSITION 15. *Assume that X has full row rank. Let λ, L, and ε be as in Proposition 14. If $\hat{\rho}$ is calculated using (6), then for $m \ge 1$ and $\beta \in \mathbb{R}^p$,*

$$\int_{\mathbb{R}^p} |k_{\mathrm{AC}}^{(m)}(\beta, \beta'; Y, X) - \pi_{B|Y,X}(\beta'|Y, X)|\,d\beta' \le H(\beta)\hat{\rho}^{m-1},$$

where

$$H(\beta) = 2 + \frac{L}{1-\lambda} + \operatorname{tr}(X\Sigma^{-1}X^T)$$
$$+ \left\|(X\Sigma^{-1}X^T)^{-1/2} X\{\Sigma^{-1}(X^T\varphi(\beta) + Qv) - \hat{B}\}\right\|^2.$$

Let $\rho_* = \rho_*(X, Y)$ denote the geometric convergence rate of the A&C chain. In the next section, which is the heart of the paper, we develop general convergence complexity results showing that, under weak regularity conditions, ρ_* is bounded away from 1 both as $n \to \infty$ (for fixed p), and as $p \to \infty$ (for fixed n).

5. Results for the Albert and Chib chain Part II: Asymptotics.

5.1. *Large n, small p.* In this section, we consider the case where p is fixed and n grows. In particular, we are interested in what happens to the geometric convergence rate of the A&C chain in this setting. Recall that the prior on B is

$$\omega(\beta) \propto \exp\{-(\beta - v)^T Q(\beta - v)/2\}.$$

Since p is fixed, so is the prior. Hence, the hyperparameters v and Q will remain fixed throughout this subsection. In Section 3.1, we introduced the data set $\mathcal{D} := \{(X_i, Y_i)\}_{i=1}^n$. We now let n vary, and consider a sequence of data sets, $\mathcal{D}_n := \{(X_i, Y_i)\}_{i=1}^n$, $n \ge 1$. So, each time n increases by 1, we are are given a new $p \times 1$ covariate vector and a corresponding binary response. In order to study the asymptotics, we assume that the (X_i, Y_i) pairs are generated according to a random mechanism that is consistent with the probit regression model. In particular, we make the following assumptions:

(A1) The pairs $\{(X_i, Y_i)\}_{i=1}^\infty$ are i.i.d. random vectors such that

$$Y_i|X_i \sim \text{Bernoulli}(G(X_i)),$$

where $G : \mathbb{R}^p \to (0, 1)$ is a measurable function;

(A2) $\mathbb{E}X_1X_1^T$ is finite and positive definite;
(A3) for $j \in \{1, 2, \ldots, 2^p\}$ and $\beta \neq 0$,

$$\mathbb{P}\big(\big(1_{S_j}(X_1) + 1_{S_j}(-X_1)\big)X_1^T\beta \neq 0\big) > 0.$$

Assumption (A1) contains the probit model as a special case. Thus, our results concerning the asymptotic behavior of A&C's Markov chain do not require strict adherence of the data to the probit regression model. The main reason for assuming (A2) is to guarantee that, almost surely, X will eventually have full column rank. This is a necessary condition for posterior propriety when $Q = 0$. While (A3) is rather technical, it is clearly satisfied if X_1 follows a distribution admitting a pdf (with respect to Lebesgue measure) that is positive over some open ball $\{\gamma \in \mathbb{R}^p : \|\gamma\|^2 \leq c\}$, where $c > 0$. It is shown in Section 7.4 of the Supplementary Material [Qin and Hobert (2018)] that (A3) also allows for an intercept. That is, even if the first component of X_1 is 1 (constant), then as long as the remaining $p - 1$ components satisfy the density condition described above, (A3) is still satisfied.

It was assumed throughout Section 3 that the posterior distribution is proper. Of course, for a fixed data set, this is check-able. All we need in the large n, small p regime is a guarantee that the posterior is proper for all large n, almost surely. The following result is proven in Section 7.5 of the Supplementary Material [Qin and Hobert (2018)].

PROPOSITION 16. *Under Assumptions* (A1)–(A3), *almost surely*, *the posterior distribution is proper for all sufficiently large* n.

For fixed n, $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ represents the first n (random) covariate vectors and responses. Let $\rho_*(\mathcal{D}_n)$ denote the (random) geometric convergence rate of the corresponding A&C chain. Here is one of our main results.

THEOREM 17. *If* (A1)–(A3) *hold*, *then there exists a constant* $\rho < 1$ *such that*, *almost surely*,

$$\limsup_{n\to\infty} \rho_*(\mathcal{D}_n) \leq \rho.$$

PROOF. Let $\hat{\rho}(\mathcal{D}_n)$ denote the upper bound on $\rho_*(\mathcal{D}_n)$ that is based on λ, L, and ε from Proposition 11. We prove the result by showing that, almost surely, $\limsup_{n\to\infty} \hat{\rho}(\mathcal{D}_n) \leq \rho < 1$. Note that $L = p(1 + \lambda)$ and $\varepsilon = 2^{-p/2}e^{-d}$ [where $d > 2L/(1 - \lambda)$]. Thus, control over λ provides control over L and ε as well. In particular, to prove the result it suffices to show that there exists a constant $c \in [0, 1)$, such that, almost surely,

$$\limsup_{n\to\infty} \lambda(\mathcal{D}_n) \leq c. \tag{21}$$

Noting that $\Sigma^{-1/2}X^TX\Sigma^{-1/2} \preceq I_p$, we have, by Proposition 13,

$$\begin{aligned}\lambda^{1/2} \le 1 - \frac{2}{\pi} \min_{1\le j\le 2^p} \lambda_{\min}\Bigg\{\left(\frac{\Sigma}{n}\right)^{-1/2}\Bigg(\frac{1}{n}\sum_{X_i\in S_j} X_i 1_{\{0\}}(Y_i)X_i^T \\ + \frac{1}{n}\sum_{X_i\in -S_j} X_i 1_{\{1\}}(Y_i)X_i^T\Bigg)\left(\frac{\Sigma}{n}\right)^{-1/2}\Bigg\}.\end{aligned} \tag{22}$$

Fix $j \in \{1, 2, \ldots, 2^p\}$. By (A1) and the strong law, almost surely,

$$\begin{aligned}&\lim_{n\to\infty}\left(\frac{\Sigma}{n}\right)^{-1/2}\left(\frac{1}{n}\sum_{X_i\in S_j} X_i 1_{\{0\}}(Y_i)X_i^T + \frac{1}{n}\sum_{X_i\in -S_j} X_i 1_{\{1\}}(Y_i)X_i^T\right)\left(\frac{\Sigma}{n}\right)^{-1/2} \\ &\quad = \big(\mathbb{E}X_1X_1^T\big)^{-1/2}\big\{\mathbb{E}X_1X_1^T 1_{S_j}(X_1)\big(1-G(X_1)\big) \\ &\qquad + \mathbb{E}X_1X_1^T 1_{S_j}(-X_1)G(X_1)\big\}\big(\mathbb{E}X_1X_1^T\big)^{-1/2}.\end{aligned}$$

It follows from (26) and Lemma 25 of the Supplementary Material [Qin and Hobert (2018)] that, almost surely,

$$\begin{aligned}&\lim_{n\to\infty} \lambda_{\min}\Bigg\{\left(\frac{\Sigma}{n}\right)^{-1/2}\Bigg(\frac{1}{n}\sum_{X_i\in S_j} X_i 1_{\{0\}}(Y_i)X_i^T \\ &\qquad + \frac{1}{n}\sum_{X_i\in -S_j} X_i 1_{\{1\}}(Y_i)X_i^T\Bigg)\left(\frac{\Sigma}{n}\right)^{-1/2}\Bigg\} \\ &\qquad \ge \lambda_{\max}^{-1}\big(\mathbb{E}X_1X_1^T\big)\lambda_{\min}\big\{\mathbb{E}X_1X_1^T 1_{S_j}(X_1)\big(1-G(X_1)\big) \\ &\qquad\quad + \mathbb{E}X_1X_1^T 1_{S_j}(-X_1)G(X_1)\big\}.\end{aligned} \tag{23}$$

By (A2), $\lambda_{\max}^{-1}(\mathbb{E}X_1X_1^T) > 0$. Hence by (22) and (23), to show that (21) holds, almost surely, it is enough to show that

$$\lambda_{\min}\big\{\mathbb{E}X_1X_1^T 1_{S_j}(X_1)\big(1-G(X_1)\big) + \mathbb{E}X_1X_1^T 1_{S_j}(-X_1)G(X_1)\big\} > 0.$$

By (A3), for any $\beta \ne 0$,

$$\mathbb{P}\big(1_{S_j}(X_1)X_1^T\beta \ne 0 \text{ or } 1_{S_j}(-X_1)X_1^T\beta \ne 0\big) > 0. \tag{24}$$

Since $0 < G(X_1) < 1$, (24) implies that, for any $\beta \ne 0$,

$$\mathbb{P}\big(\beta^T\big\{X_1X_1^T 1_{S_j}(X_1)\big(1-G(X_1)\big)\big\}\beta + \beta^T\big\{X_1X_1^T 1_{S_j}(-X_1)G(X_1)\big\}\beta > 0\big) > 0.$$

As a result, for any $\beta \in \mathbb{R}^p$ such that $\|\beta\|^2 = 1$,

$$\beta^T\big\{\mathbb{E}X_1X_1^T 1_{S_j}(X_1)\big(1-G(X_1)\big) + \mathbb{E}X_1X_1^T 1_{S_j}(-X_1)G(X_1)\big\}\beta > 0. \tag{25}$$

It follows from (A2) that the left-hand side of (25) is continuous in β. Hence, we can take infimum on both sides of the inequality with respect to β and retain the greater-than symbol, which yields

$$\lambda_{\min}\{\mathbb{E}X_1X_1^T1_{S_j}(X_1)(1-G(X_1))+\mathbb{E}X_1X_1^T1_{S_j}(-X_1)G(X_1)\}>0,$$

and the result follows. □

REMARK 18. From the proof of Theorem 17 it's easy to see that the asymptotic bound ρ in the said theorem is unaffected by the precision matrix Q. This is because the effect of the prior is overshadowed by the increasing amount of data as $n\to\infty$.

Theorem 17 shows that, under weak regularity conditions on the random mechanism that generates $\mathcal{D}_n=\{(X_i,Y_i)\}_{i=1}^n$, A&C's MCMC algorithm scales well with n. Johndrow et al. (2016) studied the convergence rate of the A&C chain as $n\to\infty$ for a particular *fixed* sequence of covariate vectors and responses. Suppose that $p=1$, $Q>0$ is a constant, $v=0$, and $X_1=X_2=\cdots=X_n=1$. They showed that if *all* the Bernoulli trials result in success, that is, $Y_1=Y_2=\cdots=Y_n=1$, then $\lim_{n\to\infty}\rho_{**}(\mathcal{D}_n)=1$. That is, in terms of L^2-geometric convergence rate, the convergence is arbitrarily slow for sufficiently large n. As we now explain, our results can be used to show that, in Johndrow et al.'s (2016) setting, almost any other sequence of responses leads to well behaved convergence rates. Let $\{Y_i\}_{i=1}^\infty$ denote a fixed sequence of binary responses, and define $\hat{p}_n=n^{-1}\sum_{i=1}^n Y_i$. It follows from Propositions 11 and 13 that the A&C chain satisfies d&m conditions with

$$\lambda=\left[\frac{n}{n+Q}-\frac{2}{\pi}\frac{n}{n+Q}\{\hat{p}_n\wedge(1-\hat{p}_n)\}\right]^2\le\left[1-\frac{2}{\pi}\{\hat{p}_n\wedge(1-\hat{p}_n)\}\right]^2,$$

$L=1+\lambda$, and $\varepsilon=2^{-1/2}e^{-d}$ for $d>2L/(1-\lambda)$. For any fixed n, suppose that there exist $c_1,c_2\in(0,1)$ such that $c_1\le\hat{p}_n\le c_2$, then [using (6)] one can find $\rho<1$, which depends *only* on $c_1\wedge(1-c_2)$, such that $\rho_*(\mathcal{D}_n)\le\rho$. It now follows that the geometric convergence rates, $\rho_*(\mathcal{D}_n)$, are eventually bounded away from 1 so long as $0<\liminf_{n\to\infty}\hat{p}_n\le\limsup_{n\to\infty}\hat{p}_n<1$. (It is important to note that $\mathcal{D}_n$ is *not* random here.) Moreover, an analogous result holds for $\rho_{**}(\mathcal{D}_n)$. Here is a formal statement.

COROLLARY 19. *For the intercept-only model described above, if*

$$0<\liminf_{n\to\infty}\hat{p}_n\le\limsup_{n\to\infty}\hat{p}_n<1,$$

then for any

$$\delta\in\left(0,1-\left[1-\frac{2}{\pi}\left\{\liminf_{n\to\infty}\hat{p}_n\wedge\left(1-\limsup_{n\to\infty}\hat{p}_n\right)\right\}\right]^2\right),$$

$\limsup_{n\to\infty} \rho_*(\mathcal{D}_n) \le \rho < 1$, *and* $\limsup_{n\to\infty} \rho_{**}(\mathcal{D}_n) \le \rho < 1$, *where* ρ *equals* $\hat{\rho}$ *in* (6) *with*

$$\lambda = \left[1 - \frac{2}{\pi}\left\{\liminf_{n\to\infty} \hat{p}_n \wedge \left(1 - \limsup_{n\to\infty} \hat{p}_n\right)\right\}\right]^2 + \delta,$$

$L = 1 + \lambda$, *and* $\varepsilon = 2^{-1/2}e^{-d}$, *where* $d > 2L/(1-\lambda)$.

PROOF. It suffices to prove the result for ρ_{**}, since the argument for ρ_* has already been provided. Fix an arbitrary δ. By Proposition 12, when n is sufficiently large, for any $\beta \in \mathbb{R}$ and $m \ge 1$,

$$\int_{\mathbb{R}} |k_{\mathrm{AC}}^{(m)}(\beta, \beta'; Y, X) - \pi_{B|Y,X}(\beta'|Y, X)| \, \mathrm{d}\beta' \le H(\beta)\rho^{m-1},$$

where $H(\beta)$ is given in the said proposition. Let Π be the probability measure corresponding to the posterior density, $\pi_{B|Y,X}(\beta|Y, X)$. Then for any probability measure $\nu \in L^2(\Pi)$ and $m \ge 1$,

$$\begin{aligned}
&\int_{\mathbb{R}} \left| \int_{\mathbb{R}} k_{\mathrm{AC}}^{(m)}(\beta, \beta'; Y, X)\nu(\mathrm{d}\beta) - \pi_{B|Y,X}(\beta'|Y, X) \right| \mathrm{d}\beta' \\
&\quad \le \int_{\mathbb{R}} \int_{\mathbb{R}} |k_{\mathrm{AC}}^{(m)}(\beta, \beta'; Y, X) - \pi_{B|Y,X}(\beta'|Y, X)| \, \mathrm{d}\beta' \nu(\mathrm{d}\beta) \\
&\quad \le \left(\int_{\mathbb{R}} H(\beta)\nu(\mathrm{d}\beta) \right) \rho^{m-1}.
\end{aligned}$$

One can verify that $H(\beta)$ can be bounded by polynomial functions. As a result, by Theorem 2.3 in Chen and Shao (2001), $\int_{\mathbb{R}} H^2(\beta)\pi_{B|Y,X}(\beta|Y, X) \, \mathrm{d}\beta < \infty$. Then by Cauchy–Schwarz, $\int_{\mathbb{R}} H(\beta)\nu(\mathrm{d}\beta) < \infty$. Therefore, $\rho_{**}(\mathcal{D}_n) \le \rho$ for all sufficiently large n. □

As mentioned previously, the convergence rate analyses of Roy and Hobert (2007) and Chakraborty and Khare (2017), which establish the geometric ergodicity of the A&C chain for fixed n and p, are based on the *uncentered* drift function, V_0. We end this subsection with a result showing that, while this uncentered drift may be adequate for *nonasymptotic* results, it simply does not match the dynamics of the A&C chain well enough to get a result like Theorem 17. The following result is proven is Section 7.6 of the Supplementary Material [Qin and Hobert (2018)].

PROPOSITION 20. *Assume that* (A1) *and* (A2) *hold, and that there exists* $\beta_* \in \mathbb{R}^p$ *such that* $\beta_* \ne 0$ *and* $G(\gamma) = G_*(\gamma^T \beta_*)$ *for all* $\gamma \in \mathbb{R}^p$, *where* $G_* : \mathbb{R} \to (0, 1)$ *is a strictly increasing function such that* $G_*(0) = 1/2$. *Then, almost surely, any drift and minorization based on* $V_0(\beta) = \|\Sigma^{1/2}\beta\|^2$ *is necessarily unstable in* n.

REMARK 21. In the above proposition, if $G_*(\theta) = \Phi(\theta)$ for all $\theta \in \mathbb{R}$, then the probit model is correctly specified, and the true parameter is β_*.

5.2. *Large p, small n.* In this subsection, we consider the case where n is fixed and p grows. In contrast with the strategy of the previous subsection, here we consider a *deterministic* sequence of data sets. Also, since p is changing, we need to specify a sequence of prior parameters $\{(Q_p, v_p)\}_{p=1}^{\infty}$. Let $\mathcal{D}_p = (v_p, Q_p, X_{n\times p}, Y)$, $p \geq 1$, denote a sequence of priors and data sets, where Y is a fixed $n \times 1$ vector of responses, $X_{n\times p}$ is an $n \times p$ matrix, v_p is a $p \times 1$ vector, and Q_p is a $p \times p$ positive definite matrix. (Note that positive definite-ness of Q_p is required for posterior propriety.) So, each time p increases by 1, we are are given a new $n \times 1$ column vector to add to the current design matrix. For the rest of this subsection, we omit the p and $n \times p$ subscripts. We also assume that the following conditions hold for all p:

(B1) X has full row rank;
(B2) There exists a finite, positive constant c, such that $\lambda_{\max}(XQ^{-1}X^T) < c$.

Assumption (B1) is equivalent to $X\Sigma^{-1}X^T$ being nonsingular. Assumption (B2) regulates the eigenvalues of the prior variance, Q^{-1}. More specifically, it requires that the prior drives B toward v. For illustration, if $X = (1\ 1\ \cdots\ 1)$, then (B2) holds if for some $\tau > 0$, $Q^{-1} = \operatorname{diag}(\tau/p, \tau/p, \ldots, \tau/p)$, or $Q^{-1} = \operatorname{diag}(\tau, \tau/2^2, \ldots, \tau/p^2)$. Assumption (B2) is satisfied by the generalized g-priors used by, for example, Gupta and Ibrahim (2007), Yang and Song (2009), and Baragatti and Pommeret (2012). It can be shown [see, e.g., Chakraborty and Khare (2017)] that (B2) is equivalent to the existence of a constant $c < 1$ such that

$$\lambda_{\max}\big(X\Sigma^{-1}X^T\big) < c.$$

While (B2) may seem like a strong assumption, we will provide some evidence later in this subsection suggesting that it may actually be necessary. Here is our main result concerning the large p, small n case.

THEOREM 22. *If* (B1) *and* (B2) *hold, then there exists a constant $\rho < 1$ such that $\rho_*(\mathcal{D}_p) \leq \rho$ for all p.*

PROOF. The proof is based on Proposition 14. Indeed, as in the proof of Theorem 17, it suffices to show that there exists a $c < 1$ such that

$$\lambda(\mathcal{D}_p) = \lambda_{\max}^2\big(X\Sigma^{-1}X^T\big) < c$$

for all p. But this follows immediately from (B2). □

An important feature of Theorem 22 is that it holds for any sequence of prior means, $\{v_p\}_{p=1}^{\infty}$. This is achieved by adopting a drift function that is centered around a point that adapts to the prior mean. Although Gaussian priors with nonvanishing means are not commonly used in practice, it is interesting to see that Albert and Chib's algorithm can be robust under location shifts in the prior, even when the dimension of the state space is high.

The following result, which is proven in Section 7.7 of the Supplementary Material [Qin and Hobert (2018)], shows that (B2) is not an unreasonable assumption.

PROPOSITION 23. *If* $n = 1$ *and* $v = 0$, *then as* $X\Sigma^{-1}X^T$ *tends to* 1,

$$1 - \rho_{**} = O\big(1 - X\Sigma^{-1}X^T\big).$$

In particular, ρ_{**} *is not bounded away from* 1 *if* (B2) *does not hold.*

Acknowledgment. We thank the Editor and four anonymous reviewers for helpful comments and suggestions.

SUPPLEMENTARY MATERIAL

Supplementary material for "Convergence complexity analysis of Albert and Chib's algorithm for Bayesian probit regression" (DOI: 10.1214/18-AOS1749SUPP; .pdf). Section 6 provides some basic results on Hermitian matrices and truncated normal distributions. Section 7 gives some technical results, and the proofs for Corollary 5, Proposition 13, Proposition 16, Proposition 20 and Proposition 23.

REFERENCES

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

BARAGATTI, M. and POMMERET, D. (2012). A study of variable selection using g-prior distribution with ridge parameter. *Comput. Statist. Data Anal.* **56** 1920–1934. MR2892387

BAXENDALE, P. H. (2005). Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Ann. Appl. Probab.* **15** 700–738. MR2114987

CHAKRABORTY, S. and KHARE, K. (2017). Convergence properties of Gibbs samplers for Bayesian probit regression with proper priors. *Electron. J. Stat.* **11** 177–210. MR3604022

CHEN, M.-H. and SHAO, Q.-M. (2001). Propriety of posterior distribution for dichotomous quantal response models. *Proc. Amer. Math. Soc.* **129** 293–302. MR1694452

DIACONIS, P., KHARE, K. and SALOFF-COSTE, L. (2008). Gibbs sampling, exponential families and orthogonal polynomials (with discussion). *Statist. Sci.* **23** 151–178. MR2446500

DURMUS, A. and MOULINES, E. (2016). High-dimensional Bayesian inference via the unadjusted Langevin algorithm. arXiv:1605.01559.

FLEGAL, J. M., HARAN, M. and JONES, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statist. Sci.* **23** 250–260. MR2516823

FORT, G., MOULINES, E., ROBERTS, G. O. and ROSENTHAL, J. S. (2003). On the geometric ergodicity of hybrid samplers. *J. Appl. Probab.* **40** 123–146. MR1953771

GUPTA, M. and IBRAHIM, J. G. (2007). Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *J. Amer. Statist. Assoc.* **102** 867–880. MR2411650

HAIRER, M. and MATTINGLY, J. C. (2011). Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI. Progress in Probability* **63** 109–117. Birkhäuser/Springer Basel AG, Basel. MR2857021

JOHNDROW, J. E., SMITH, A., PILLAI, N. and DUNSON, D. B. (2018). MCMC for imbalanced categorical data. *J. Amer. Statist. Assoc.* To appear. Available at arXiv:1605.05798.

JONES, G. L. (2001). *Convergence Rates and Monte Carlo Standard Errors for Markov Chain Monte Carlo Algorithms*. ProQuest LLC, Ann Arbor, MI. Ph.D. thesis, Univ. Florida. MR2702583

JONES, G. L. and HOBERT, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statist. Sci.* **16** 312–334. MR1888447

ŁATUSZYŃSKI, K., MIASOJEDOW, B. and NIEMIRO, W. (2013). Nonasymptotic bounds on the estimation error of MCMC algorithms. *Bernoulli* **19** 2033–2066. MR3129043

MARCHEV, D. and HOBERT, J. P. (2004). Geometric ergodicity of van Dyk and Meng's algorithm for the multivariate Student's t model. *J. Amer. Statist. Assoc.* **99** 228–238. MR2054301

MEYN, S. P. and TWEEDIE, R. L. (1994). Computable bounds for geometric convergence rates of Markov chains. *Ann. Appl. Probab.* **4** 981–1011. MR1304770

MEYN, S. and TWEEDIE, R. L. (2009). *Markov Chains and Stochastic Stability*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2509253

QIN, Q. and HOBERT, J. P. (2018). Supplement to "Convergence complexity analysis of Albert and Chib's algorithm for Bayesian probit regression." DOI:10.1214/18-AOS1749SUPP.

RAJARATNAM, B. and SPARKS, D. (2015). MCMC-based inference in the era of big data: A fundamental analysis of the convergence complexity of high-dimensional chains. arXiv:1508.00947.

ROBERTS, G. O. and ROSENTHAL, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electron. Commun. Probab.* **2** 13–25. MR1448322

ROBERTS, G. O. and ROSENTHAL, J. S. (1998). Markov-chain Monte Carlo: Some practical implications of theoretical results (with discussion). *Canad. J. Statist.* **26** 5–31. MR1624414

ROBERTS, G. O. and ROSENTHAL, J. S. (2001). Markov chains and de-initializing processes. *Scand. J. Stat.* **28** 489–504. MR1858413

ROBERTS, G. O. and TWEEDIE, R. L. (1999). Bounds on regeneration times and convergence rates for Markov chains. *Stochastic Process. Appl.* **80** 211–229. MR1682243

ROBERTS, G. O. and TWEEDIE, R. L. (2001). Geometric L^2 and L^1 convergence are equivalent for reversible Markov chains. *J. Appl. Probab.* **38A** 37–41. MR1915532

ROSENTHAL, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **90** 558–566. MR1340509

ROY, V. and HOBERT, J. P. (2007). Convergence rates and asymptotic standard errors for Markov chain Monte Carlo algorithms for Bayesian probit regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 607–623. MR2370071

ROY, V. and HOBERT, J. P. (2010). On Monte Carlo methods for Bayesian multivariate regression models with heavy-tailed errors. *J. Multivariate Anal.* **101** 1190–1202. MR2595301

RUDIN, W. (1976). *Principles of Mathematical Analysis*, 3rd ed. McGraw-Hill Book Co., New York-Auckland-Düsseldorf. MR0385023

SINCLAIR, A. and JERRUM, M. (1989). Approximate counting, uniform generation and rapidly mixing Markov chains. *Inform. and Comput.* **82** 93–133. MR1003059

VATS, D. (2017). Geometric ergodicity of Gibbs samplers in Bayesian penalized regression models. *Electron. J. Stat.* **11** 4033–4064. MR3714307

YANG, J. and ROSENTHAL, J. S. (2017). Complexity results for MCMC derived from quantitative bounds. arXiv:1708.00829.

YANG, A.-J. and SONG, X.-Y. (2009). Bayesian variable selection for disease classification using gene expression data. *Bioinformatics* **26** 215–222.

YANG, Y., WAINWRIGHT, M. J. and JORDAN, M. I. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *Ann. Statist.* **44** 2497–2532. MR3576552

DEPARTMENT OF STATISTICS
UNIVERSITY OF FLORIDA
GAINESVILLE, FLORIDA 32611
USA
E-MAIL: qianqin11@ufl.edu
jhobert@stat.ufl.edu