Wasserstein-based methods for convergence complexity analysis of MCMC with application to Albert and Chib’s algorithm

Qian Qin and James P. Hobert
Department of Statistics
University of Florida
2018

Abstract

Over the last 25 years, techniques based on drift and minorization (d&m) have been mainstays in the convergence analysis of MCMC algorithms. However, results presented herein suggest that d&m may be less useful in the emerging area of convergence complexity analysis, which is the study of how Monte Carlo Markov chain convergence behavior scales with sample size, \( n \), and/or number of covariates, \( p \). The problem appears to be that minorization becomes a serious liability as dimension increases. Alternative methods of constructing convergence rate bounds (with respect to total variation distance) that do not require minorization are investigated. These methods incorporate both old and new theory on Wasserstein distance and random mappings, and produce bounds that are apparently more robust to increasing dimension than those based on d&m. Indeed, the Wasserstein-based bounds are used to develop strong convergence complexity results for Albert and Chib’s (1993) algorithm in the challenging asymptotic regime where both \( n \) and \( p \) diverge. We note that Qin and Hobert’s (2019+) d&m-based analysis of the same algorithm led to useful results in the cases where \( n \to \infty \) with \( p \) fixed, and \( p \to \infty \) with \( n \) fixed, but these authors provided no results for the case where \( n \) and \( p \) are both large.

1 Introduction

Markov chain Monte Carlo (MCMC) has become an indispensable tool in Bayesian statistics, and it is now well-known that the performance of an MCMC algorithm depends heavily on the convergence properties of the underlying Markov chain. In the era of big data, it is no longer enough to understand the convergence

Key words and phrases. Coupling, Drift condition, Geometric ergodicity, High dimensional inference, Minorization condition, Random mapping
behavior of a given algorithm for each fixed data set. Indeed, there is now growing interest in the so-called convergence complexity of MCMC algorithms, which describes how the convergence rate of the Markov chain scales with the sample size, \( n \), and/or the number of covariates, \( p \), in the associated data set (see, e.g., Rajaratnam and Sparks, 2015; Durmus and Moulines, 2016; Yang et al., 2016; Yang and Rosenthal, 2017; Johndrow et al., 2018; Qin and Hobert, 2019+). While techniques based on drift and minorization (d&m) conditions have been mainstays in the analysis of MCMC algorithms for decades, we show that these techniques are likely to fail in convergence complexity analysis. We consider alternative methods based on Wasserstein distance and coupling, and we demonstrate the potential advantages of such methods over those based on d&m through a convergence complexity analysis of Albert and Chib’s (1993) algorithm for Bayesian probit regression.

Let \( \Pi \) denote an intractable posterior distribution on some state space \( X \), and consider a Monte Carlo Markov chain (with stationary probability measure \( \Pi \)) that is driven by the Markov transition function (Mtf) \( K(\cdot, \cdot) \). For a non-negative integer \( m \) and \( x \in X \), let \( K^m(\cdot, \cdot) \) be the \( m \)-step Mtf, and let \( K_x^m \) be the distribution defined by \( K^m(x, \cdot) \). (See Section 2 for formal definitions.) Convergence of this chain can be assessed by the rate at which \( D(K_x^m, \Pi) \) decreases as \( m \) grows, where \( D \) is some divergence function between probability measures. (Traditionally, the most commonly used \( D \) is the total variation distance.) If there exist \( \rho < 1 \) and \( M : X \to [0, \infty) \) such that

\[
D(K_x^m, \Pi) \leq M(x)\rho^m, \quad m \geq 0, \quad x \in X,
\]

then we say that the chain converges geometrically in \( D \). The rate of convergence, \( \rho_e \in [0, 1] \), is defined to be the infimum of the \( \rho \)s that satisfy (1) for some \( M(\cdot) \). Given a sequence of (growing) data sets, and corresponding sequences of posterior distributions and Monte Carlo Markov chains, we are interested in the asymptotic behavior of \( \rho_e \). Our program entails constructing an upper bound on \( \rho_e \), which we call \( \hat{\rho} \), and then considering the asymptotic behavior of this bound. Of course, the bound should be sufficiently sharp so that it correctly reflects the asymptotic dynamics of \( \rho_e \).

When \( D \) is the total variation distance, \( \hat{\rho} \) is often constructed via drift and minorization (d&m) arguments (see, e.g., Rosenthal, 1995). It is well-known that the resulting bounds are often overly conservative, especially in high dimensional settings (see, e.g., Rajaratnam and Sparks, 2015). While there have been a few successful d&m-based convergence complexity analyses of MCMC algorithms (Yang and Rosenthal, 2017; Qin and Hobert, 2019+), our results strongly suggest that techniques based on d&m are unlikely to be at the forefront of convergence complexity analysis going forward. Indeed, we consider a family of simple autoregressive processes, and show that, in this context, the usual d&m-based methods cannot possibly produce...
sharp bounds on $\rho_s$ as dimension increases. The problems exhibited by these methods in this example are so fundamental that they cannot be avoided no matter how hard one tries to find good d&m conditions. The culprit appears to be the minorization condition, which becomes a serious liability as dimension increases.

Recent results suggest that convergence complexity analysis becomes more tractable when total variation distance is replaced with an appropriate Wasserstein distance (see, e.g., Durmus and Moulines, 2015; Mangoubi and Smith, 2017; Trillos et al., 2017; Hairer et al., 2011). This could be (at least partly) due to the fact that minorization conditions are typically not used to bound Wasserstein distance. In this paper, we explore the possibility of performing convergence complexity analysis of MCMC algorithms under total variation distance by using a technique developed in Madras and Sezer (2010) to convert Wasserstein bounds into total variation bounds. While there is a substantial literature on bounding the Wasserstein distance to stationarity for Markov chains (see, e.g., Steinsaltz, 1999; Ollivier, 2009; Hairer et al., 2011; Durmus and Moulines, 2015; Butkovsky, 2014), the available methods are not directly applicable to typical Monte Carlo Markov chains that arise in Bayesian statistics. We present several new results that extend the range of applicability of the existing methods to such Markov chains. In particular, we provide a method for constructing geometric convergence bounds in a (generalized) Wasserstein distance through a drift condition and associated contraction conditions. This extends results of Durmus and Moulines (2015) and Butkovsky (2014), which deal with Wasserstein distances defined via bounded metrics, to the unbounded case, at the cost of imposing a stronger contraction condition. We also establish a new result that facilitates the application of techniques developed in Steinsaltz (1999) and Ollivier (2009).

Our application of the Wasserstein-to-TV method to Albert and Chib’s (1993) algorithm shows that this approach can produce bounds that are more robust to increasing dimension than those based on d&m. Indeed, Qin and Hobert (2019+) (hereafter, Q&H) recently performed a convergence complexity analysis of the Albert and Chib (A&C) algorithm using d&m methods. Roughly, the two main results (which concern convergence in total variation) are as follows: (1) When $p$ is fixed, under mild conditions on the data structure, $\limsup_{n \to \infty} \rho_s \leq \rho_p$ for some $\rho_p < 1$. (2) When $n$ is fixed, $\limsup_{p \to \infty} \rho_s \leq \rho_n$ for some $\rho_n < 1$, provided that the prior distribution provides sufficiently strong shrinkage. Unfortunately, one can show that $\rho_p \to 1$ as $p \to \infty$, and that $\rho_n \to 1$ as $n \to \infty$. Thus, Q&H were not able to provide useful asymptotic results on $\rho_s$ when both $n$ and $p$ are large. Here we are able to utilize the Wasserstein-to-TV method to get results in this more complex asymptotic regime. Our results for the A&C chain can loosely be described as follows: (1’) When $p$ is fixed, under certain sparsity assumptions on the true regression coefficients (and some regularity conditions), $\limsup_{n \to \infty} \rho_s \leq \rho$ for some $\rho < 1$ independent of $p$. (2’) When the prior provides enough shrinkage, $\rho_s \leq \rho$ for some $\rho < 1$ independent of $n$ and $p$. (3’) When rows of the design
matrix are duplicated, under sparsity assumptions on the true regression coefficients (and some regularity conditions), $\rho^*$ is bounded above by some $\rho < 1$ with high probability as $n$ and $p$ grow simultaneously at some appropriate joint rate. Along the way, we also establish some non-asymptotic results, one of them being that the A&C chain converges geometrically in the Wasserstein distance induced by the Euclidean norm for any finite data set.

Our results for the A&C chain are among the first of their kind, providing strong asymptotic statements on convergence rates for a practically relevant Monte Carlo Markov chain in the case where both $n$ and $p$ are large. While we consider only a single serious example in this paper, we believe that many of our ideas and techniques are potentially applicable to other similar high dimensional problems.

The remainder of this article is structured as follows. In Section 2 we illustrate how d&m-based methods fail in typical high dimensional settings by studying a simple autoregressive process. Alternative methods based on coupling, random mappings and Wasserstein bounds are the topic of Section 3. Our analysis of the A&C chain is presented in Section 4. Some technical details are relegated to an Appendix.

### 2 Minorization Becomes a Liability as Dimension Increases

Let $(X, \mathcal{B})$ be a countably generated measurable space. We consider a discrete-time time-homogeneous Markov chain on $X$ with Mtf $K : X \times \mathcal{B} \rightarrow [0, 1]$. For an integer $m \geq 0$, let $K^m : X \times \mathcal{B} \rightarrow [0, 1]$ be the corresponding $m$-step Mtf, so that for any $x \in X$ and $A \in \mathcal{B}$,

$$K^0(x, A) = 1_{x \in A}, \quad K^{m+1}(x, A) = \int_X K(y, A) K^m(x, dy).$$

Let $K^m_x : x \in X$, denote the probability measure defined by $K^m(x, \cdot)$. We assume that the Markov chain is Harris ergodic, i.e., irreducible, aperiodic, and positive Harris recurrent. Thus, the chain has a unique stationary distribution $\Pi$ to which it converges. The difference between $K^m_x$ and $\Pi$ is most commonly measured by the total variation distance, $\|K^m_x - \Pi\|_{TV}$, which is the supremum of their discrepancy over measurable sets. The associated convergence rate, denoted by $\rho^*_{TV}$, is defined as the infimum of $\rho \in [0, 1]$ that satisfy (P) when $D$ is the TV distance. A standard technique for constructing upper bounds on $\rho^*_{TV}$ is based on d&m conditions. One of the earliest examples of this method is due to Rosenthal (1995), whose result is now stated.

**Proposition 1.** (Rosenthal, 1995) Suppose that

(A1) there exist $\lambda < 1$, $b < \infty$, and a function $V : X \rightarrow [0, \infty)$ such that

$$\int_X V(x') K(x, dx') \leq \lambda V(x) + b$$


for all $x \in X$;

(A2) there exist $d > 2b/(1 - \lambda)$, $\gamma < 1$ and a probability measure $\nu : B \rightarrow [0, 1]$ such that for every $x \in C := \{x' \in X : V(x') \leq d\}$, $K(x, \cdot) \geq (1 - \gamma)\nu(\cdot)$.

Then for all $x \in X$ and $m \geq 0$,

$$\|K_x^m - \Pi\|_{TV} \leq \gamma^m \left(1 + \frac{b}{1 - \lambda} + V(x)\right) \left\{\left(\frac{1 + 2b + \lambda d}{1 + d}\right)^{1-a} \left[1 + 2(\lambda d + b)\right]^a\right\}^m,$$

where $a \in (0, 1)$ is arbitrary.

**Remark 2.** The function $V$ is called a drift function, and $C$ is called a small set. (A1) and (A2) are referred to as the drift and minorization conditions, respectively.

Proposition 1 gives the following upper bound on $\hat{\rho}^*_{TV}$:

$$\hat{\rho}_{Ros} = \gamma^a \vee \left\{\left(\frac{1 + 2b + \lambda d}{1 + d}\right)^{1-a} \left[1 + 2(\lambda d + b)\right]^a\right\}.$$

Note that $(1 + 2b + \lambda d)/(1 + d) < 1$, so there always exists an $a$ such that $\hat{\rho}_{Ros} < 1$. In the remainder of this section, we argue that this method is likely to fail when the Markov chain is high dimensional. The problem appears to be the nonexistence of a minorization condition with a small value of $\gamma$.

**Lemma 3.** Let $C$ be as in Proposition 1. Then $\Pi(C) \geq 1/2$.

**Proof.** Let $V$, $\lambda$, $b$, and $d$ be as in the said proposition. A cut-off argument (see, e.g., Hairer, 2006, Proposition 4.24) shows that (A1) implies $\Pi V \leq b/(1 - \lambda)$. (This inequality is trivial to verify if it is known that $\Pi V < \infty$.) On the other hand, since $V(x) \geq d1_{X \setminus C}(x)$,

$$\Pi V \geq d(1 - \Pi(C)) \geq \frac{2b(1 - \Pi(C))}{1 - \lambda}.$$

The result is then immediate. \qed

As dimension increases, $\Pi$ tends to “spread out,” and the requirement that $\Pi(C) \geq 1/2$ forces $C$ to be a large subset of $X$. As a result, typically, (A2) can only hold when $\gamma$ is very close to 1, and this in turn leads to an upper bound on $\rho^*_{TV}$ that is very close to 1. We now illustrate this phenomenon using a family of simple autoregressive processes.

Let $X = \mathbb{R}^p$, and let $K(x, \cdot)$ be the probability measure associated with the $N(x/2, 3I_p/4)$ distribution. This Mtf defines a simple, well-behaved autoregressive process. It is Harris ergodic, its invariant distribution $\Pi$ is $N(0, I_p)$, and it is known that the convergence rate is $\rho^*_{TV} = 1/2$ for all $p$. Now consider using
Proposition 1 to construct an upper bound on $\hat{\rho}_{\text{ros}}$. In particular, for each $p$, we choose a drift function $V : \mathbb{R}^p \to [0, \infty)$ and an associated small set $C$, which together yield an upper bound, $\hat{\rho}_{\text{ros}}$. The following result shows that the sequence of bounds constructed using Proposition 1 is necessarily quite badly behaved.

**Proposition 4.** Let $K(x, \cdot), \ x \in \mathbb{R}^p$, be the probability measure associated with the $N(x/2, 3I_p/4)$ distribution. For any sequence of d&m conditions, $\hat{\rho}_{\text{ros}} \to 1$ at an exponential rate as $p \to \infty$.

Before proving Proposition 4, we state a general result concerning condition $(A2)$. The proof is left to the reader.

**Lemma 5.** Suppose that $K(x, \cdot), \ x \in X$, admits a density function $k(x, \cdot)$ with respect to some reference measure $\mu$. If $(A2)$ holds, then for all $x, y \in C$,

$$\frac{1}{2} \int_X |k(x, x') - k(y, x')| \mu(dx') \leq \gamma.$$ 

**Proof of Proposition 4.** Let $C$ be as in Proposition 1. It is easy to show that, in order for $\Pi(C) \geq 1/2$, the diameter of $C$ must be at least $2\sqrt{m_p}$, where $m_p$ is the median of a $\chi^2_p$ distribution. Hence, letting $k(x, \cdot), x \in \mathbb{R}^p$, be the density associated with $K$, we have

$$\sup_{x,y \in C} \int_X |k(x, x') - k(y, x')| \mu(dx') \geq \sqrt{\frac{2}{3\pi}} \int_R \left| \exp \left[ -\frac{2}{3} \left( x - \frac{\sqrt{m_p}}{2} \right)^2 \right] - \exp \left[ -\frac{2}{3} \left( x + \frac{\sqrt{m_p}}{2} \right)^2 \right] \right| dx$$

$$= 2 - 4\Phi \left( -\sqrt{\frac{m_p}{3}} \right).$$

Now, $m_p$ is of order $p$ and $\Phi(-\sqrt{m_p}/3)$ goes to 0 exponentially fast as $m_p \to \infty$. Hence, it follows from Lemma 5 that $\gamma \to 1$ at an exponential rate as $p \to \infty$, which in turn implies (see, e.g., Q&H, Proposition 2) that $\hat{\rho}_{\text{ros}} \to 1$ at an exponential rate.

We conclude that, no matter how hard we work to find good d&m conditions, Proposition 1 cannot possibly yield a reasonable asymptotic bound on the convergence rate for our simple autoregressive process as $p \to \infty$. Moreover, it seems unlikely that the situation would be any better for more complex Markov chains, like those used in MCMC. The reader may wonder whether the problems described above extend to other d&m-based bounds. The answer is “yes.” Indeed, Proposition 4 continues to hold if we replace $\hat{\rho}_{\text{ros}}$ with the corresponding bound from Hairer and Mattingly (2011) (and the proof is essentially the same). Also, in Section A of the Appendix, we show that the bounds developed by Roberts and Tweedie (1999) and Baxendale (2005) behave similarly in our toy example. Note that the d&m conditions used in Baxendale (2005) do not impose a constant positive lower bound on $\Pi(C)$, as in Lemma 3 yet the resultant convergence bound still suffers from high-dimensionality. Intuitively, a good d&m-based bound requires a minorization
inequality with a small set $C$ that is large enough to be visited frequently by the chain, but simultaneously small enough that $\gamma$ is not too close to 1. However, at least for our toy example, $\gamma$ is highly susceptible to the growing size of $C$. As a result, in high dimensional settings where $\Pi$ tends to spread out, the “Goldilocks” small set doesn’t exist.

Finally, it should be mentioned that, if one is able to establish $d&m$ for $K^m$ rather than $K$ itself, where $m$ is a sufficiently large integer, then it’s possible to avoid the problems described above. However, in practical examples, it is rarely possible to establish sharp minorization inequalities for multi-step Mtfs.

Q&H did make effective use of $\hat{\rho}_{\text{ros}}$ in a convergence complexity analysis of A&C’s Markov chain in two asymptotic regimes: $n \to \infty$ with $p$ fixed, and $p \to \infty$ with $n$ fixed. This fact seems at odds with the arguments given above. However, Q&H were able to circumvent the problems associated with high-dimensionality by showing that the Gibbs chain underlying the A&C algorithm, which has support $\mathbb{R}^p \times \mathbb{R}^n$, is equivalent, in terms of convergence rate, to two other chains, one on $\mathbb{R}^p \times \mathbb{R}^p$, and the second on $\mathbb{R}^n \times \mathbb{R}^n$. In each of the two asymptotic regimes considered, one of the two variables ($n$ or $p$) is fixed, which allowed the authors to (effectively) avoid increasing dimension. On the other hand, not surprisingly, these authors were unable to use $\hat{\rho}_{\text{ros}}$ to establish any positive results in the more interesting asymptotic regime in which both $n$ and $p$ diverge.

In the next section, we provide a different type of general bound on $\hat{\rho}_{\text{ros}}^G$ that does not require a minorization condition. This bound, which is developed using Wasserstein distance and coupling, is apparently more robust to increasing dimension than $\hat{\rho}_{\text{ros}}$. Indeed, we employ the new bound in Section 4 to show that the rate of convergence of A&C’s Markov chain is bounded below 1 in three different asymptotic regimes where $n$ and $p$ both diverge.

3 Total Variation Bounds via Coupling-based Wasserstein Bounds

Let $\psi : X \times X \to [0, \infty)$ be a measurable metric, i.e., distance function. For two probability measures on $(X, \mathcal{B})$, $\mu$ and $\nu$, their Wasserstein divergence induced by $\psi$ is defined as

$$W_\psi(\mu, \nu) = \inf_{\psi \in \Upsilon(\mu, \nu)} \int_{X \times X} \psi(x, y) \, \nu(dx, dy),$$

where $\Upsilon(\mu, \nu)$ is the set of all couplings of $\mu$ and $\nu$, i.e., the set of all probability measures on the product space $(X \times X, \mathcal{B} \times \mathcal{B})$ whose marginals are respectively $\mu$ and $\nu$. (At this point, we do not require $W_\psi$ to be a formal distance between probability measures, so we are not making the usual moment and Polish space assumptions, but we will do so in Section 4.) Our goal is to bound $W_\psi(K^m_x, \Pi)$ for $x \in X$ and $m \geq 0$. 7
The associated convergence rate, $\rho_*(\psi) \in [0, 1]$, is the infimum of the $\rho$s that satisfy (1) when $D$ is the Wasserstein divergence $W_\psi$.

A natural way of bounding the Wasserstein divergence between $K^m_x$ and $K^m_y$ is to construct a pair of coupled Markov chains governed by an Mtf $\tilde{K} : (X \times X) \times (B \times B) \to [0, 1]$ such that $\tilde{K}_{(x,y)}^1 \in \Upsilon(K^1_x, K^1_y)$ for all $x, y \in X$. Then, for any $m \geq 0$, $\tilde{K}^m_{(x,y)} \in \Upsilon(K^m_x, K^m_y)$, and it follows that

$$W_\psi(K^m_x, K^m_y) \leq \int_{X\times X} \psi(x', y') \tilde{K}^m((x, y), (dx', dy')) .$$

We call $\tilde{K}$ a coupling kernel of $K$. Based on this construction, one can arrive at the following well-known result (see, e.g., Ollivier, 2009, Corollary 21).

**Proposition 6.** Suppose that $c(x) := \int_X \psi(x, y) K(x, dy) < \infty$ for all $x \in X$. Suppose further that there exists a coupling kernel of $K$, denoted by $\tilde{K}$, and $\gamma < 1$ such that for every $x, y \in X$,

$$\int_{X\times X} \psi(x', y') \tilde{K}((x, y), (dx', dy')) \leq \gamma \psi(x, y).$$

(2)

Then for each $x \in X$ and $m \geq 0$,

$$W_\psi(K^m_x, \Pi) \leq \frac{c(x)}{1 - \gamma^m} .$$

We provide a proof (based on our minimal set of assumptions) for completeness.

**Proof.** It follows immediately from (2) that for any $x, y \in X$ and $m \geq 0$,

$$\int_{X\times X} \psi(x', y') \tilde{K}^m((x, y), (dx', dy')) \leq \psi(x, y) \gamma^m .$$

Now, fix $x \in X$, let $y \sim \Pi$, and let $(\tilde{X}, \tilde{Y})|(x, y) \sim \tilde{K}^m_{(x,y)}$. Then $\tilde{X} \sim K^m_x$, and $\tilde{Y} \sim \Pi$. Thus,

$$W_\psi(K^m_x, \Pi) \leq \mathbb{E}\psi(\tilde{X}, \tilde{Y}) = \int_X \int_X \psi(x', y') \tilde{K}^m((x, y), (dx', dy')) \Pi(dy) \leq \int_X \psi(x, y) \Pi(dy) \gamma^m .$$

It remains to show that $\int_X \psi(x, y) \Pi(dy) \leq c(x)/(1 - \gamma)$. Note that for any $x, y \in X$, by the triangle inequality,

$$\int_X \psi(x, y') K(y, dy') \leq \int_{X\times X} \left[ \psi(x', y') + \psi(x, x') \right] \tilde{K}((x, y), (dx', dy'))$$

$$\leq \gamma \psi(x, y) + c(x) .$$

By Hairer's (2006) Proposition 4.24, $\int_X \psi(x, y) \Pi(dy) \leq c(x)/(1 - \gamma)$. □

In practice, it is often impossible to find a coupling that yields (2) for all $x, y \in X$. However, if the underlying Markov chain satisfies additional conditions, then (2) need not hold on all of $X \times X$. Indeed, Durmus and Moulines (2015) show that, if the chain satisfies a drift condition, then it is enough that (2)
holds on the subset of $X \times X$ where the (joint) drift function takes small values (see also, Jarner and Tweedie, 2001; Butkovsky, 2014). Unfortunately, Durmus and Moulines’s (2015) result assumes that the metric $\psi$ is bounded. Of course, in most cases, the natural metric is unbounded. Below we establish a new version of Durmus and Moulines’s (2015) result in which the bounded metric assumption is removed at the cost of replacing their average contraction condition with a stronger point-wise contraction condition. While strong, our point-wise contraction condition on the coupling does hold in some MCMC applications, e.g., it holds for the coupling that we use to analyze A&C’s algorithm. To state our result, we must first introduce the notion of a random mapping, which is a common tool for constructing couplings and studying Wasserstein divergences.

Let $(\Omega, F, \mathbb{P})$ be a probability space. Let $\theta : \Omega \rightarrow \Theta$ be a random element that assumes values in some measurable space $\Theta$, and let $\tilde{f} : X \times \Theta \rightarrow X$ be a measurable function. Set $f(x) = \tilde{f}(x, \theta)$ for $x \in X$. Then $f$ is called a random mapping on $X$. If $f(x) \sim K_x$ for all $x \in X$, then $f$ is said to induce $K(\cdot, \cdot)$. Assume that this is the case and let $\{f_m\}_{m=1}^{\infty}$ be iid copies of $f$. Denote $f_m \circ f_{m-1} \circ \cdots \circ f_1$ by $F_m$, and define $F_0 : X \rightarrow X$ to be the identity function. Then, for every $x, y \in X$, $\{(F_m(x), F_m(y))\}_{m=0}^{\infty}$ is a time-homogeneous Markov chain such that the joint distribution of $F_m(x)$ and $F_m(y)$ lies in $\Upsilon(K^m_x, K^m_y)$ for $m \geq 0$. Obviously, the conditional distribution of $(f(x), f(y))$ given $(x, y)$ defines a coupling kernel $\tilde{K}((x, y), \cdot)$. As an example, consider again the autoregressive process from Section 2. For $x \in \mathbb{R}^p$, let $f(x) = x/2 + \sqrt{3/4} N$ where $N \sim N(0, I_p)$. Then $f(x) \sim N(x/2, 3I_p/4)$, so $f$ induces $K$. Furthermore, letting $\| \cdot \|_2$ be the Euclidean norm, we have, for any $x, y \in \mathbb{R}^p$,

$$
\mathbb{E}\|f(x) - f(y)\|_2 = \mathbb{E}\left\| \frac{x}{2} + \sqrt{\frac{3}{4}} N - y - \frac{3}{4} N \right\|_2 = \frac{1}{2}\|x - y\|_2.
$$

Thus, taking $\psi$ to be Euclidean distance in Proposition 6, we have $\rho_\psi(\psi) \leq 1/2$ for all $p$.

Again, establishing (2) for all $x, y \in X$ is typically not feasible when dealing with practical Monte Carlo Markov chains. The following result provides some relief. Its proof, which is similar to that of Durmus and Moulines’s (2015) Theorem 1, is relegated to the appendix.

**Proposition 7.** Suppose that $K(\cdot, \cdot)$ is induced by a random mapping $f$. Suppose further that

(A1) there exist $\lambda < 1$, $b < \infty$, and a function $V : X \rightarrow [0, \infty)$ such that

$$
\int_X V(x')K(x, dx') \leq \lambda V(x) + b
$$

for all $x \in X$;
(A2') there exist $d > 2b/(1 - \lambda)$ and $\gamma < 1$ such that for all $x, y \in C := \{x \in X : V(x) \leq d\}$,

$$\mathbb{E}\psi(f(x), f(y)) \leq \gamma \psi(x, y) ;$$

(A3) for every $x, y \in X$,

$$\psi(f(x), f(y)) \leq \psi(x, y) .$$

Then for all $x \in X$ and $m \geq 0$,

$$W_{\psi}(K^m_x, \Pi) \leq c_1(x)\gamma^{am} + c_2(x) \left\{ \left( \frac{1 + 2b + \lambda d}{1 + d} \right)^{1-a} [1 + 2(\lambda d + b)]^a \right\}^m , \quad (3)$$

where $c_1(x) = \int_X \psi(x, y) \Pi(dy)$,

$$c_2(x) = \int_X \psi(x, y) (1 + V(y) + V(x)) \Pi(dy) ,$$

and $a \in (0, 1)$ is arbitrary.

Note that when $\Pi$ is intractable, the convergence bound (3) is not fully computable because $c_1$ and $c_2$ involve integration with respect to $\Pi$. However, when $c_1, c_2 < \infty$, it does give an explicit bound on the convergence rate, $\rho_*(\psi)$, namely,

$$\hat{\rho}_0 = \gamma^a \vee \left\{ \left( \frac{1 + 2b + \lambda d}{1 + d} \right)^{1-a} [1 + 2(\lambda d + b)]^a \right\} .$$

While this formula is exactly the same as that for $\hat{\rho}_{\text{Ros}}$, the two bounds are not the same because $\gamma$ is defined differently here. Moreover, this $\gamma$ is typically much more robust to the size of the small set $C$. For example, the coupling we constructed for the autoregressive process satisfies (A2') with $\gamma = 1/2$ for any $C$ (and any $p \in \mathbb{N}$).

Remark 8. In Durmus and Moulines (2015) and Butkovsky (2014), it is assumed that the metric is bounded, and geometric convergence is established under (A1), (A2') and an average contraction condition that is much weaker than (A3), roughly stated as $\mathbb{E}\psi(f(x), f(y)) \leq \psi(x, y)$ for all $x, y \in X$.

Contraction conditions like (A2') and (A3) can be difficult to verify in practice. We now use ideas from Steinsaltz (1999) to show that, if $X$ is a nice Euclidean space, then we can establish (A2') and (A3) by regulating the local behavior of $f$. Assume that $X$ is a convex subset of a Euclidean space, and that $\psi(x, y) = \|x - y\|$, $x, y \in X$, where $\| \cdot \|$ is a norm (not necessarily the Euclidean norm). For a random mapping $f$ on $X$, define the local Lipschitz constant of $f$ at $x \in X$ to be

$$D_x f = \limsup_{y \to x} \frac{\|f(y) - f(x)\|}{\|y - x\|} .$$
Assume that there is a measurable function \( \varphi : X \times \Theta \to [0, \infty) \) such that \( \tilde{D}_x f := \varphi(x, \theta) \geq D_x f \) for each \((x, \theta) \in X \times \Theta\). The following two lemmas can be useful for establishing \((A2')\) and \((A3)\). A proof of Lemma \(10\) is provided in the Appendix.

**Lemma 9.** Suppose that \( X \) is a convex subset of a Euclidean space, and that \( \psi \) is induced by a norm. Let \( f \) be a random mapping on \( X \). If \( D_x f \leq 1 \) for every \( x \in X \), then for each \( x, y \in X \), \( \psi(f(x), f(y)) \leq \psi(x,y) \).

**Lemma 10.** Suppose that \( X \) is a convex subset of a Euclidean space, and that \( \psi \) is induced by a norm. Let \( f \) be a random mapping on \( X \), and let \( C \in \mathcal{B} \) be convex. If for each \( x, y \in C \), \( \tilde{D}_{x+t(y-x)} f \), as a function of \( t \in [0, 1] \), is Riemann integrable, almost surely in \( \Omega \), then \( \mathbb{E} \psi(f(x), f(y)) \leq \gamma \psi(x,y) \) for every \( x, y \in C \), where \( \gamma = \sup_{x \in C} \mathbb{E} \tilde{D}_x f \).

Combining Lemma \(10\) and Proposition \(9\) yields the following result.

**Corollary 11.** Suppose that \( X \) is a convex subset of a Euclidean space, and that \( \psi \) is induced by a norm. Suppose further that \( c(x) := \int_X \psi(x, y) K(x, dy) < \infty \) for all \( x \in X \), and that \( K \) is induced by a random mapping \( f \). Assume that for each \( x, y \in X \), \( \tilde{D}_{x+t(y-x)} f \), as a function of \( t \in [0, 1] \), is Riemann integrable, almost surely in \( \Omega \), and that

\[
\gamma := \sup_{x \in X} \mathbb{E} \tilde{D}_x f < 1 .
\]

Then for each \( x \in X \) and \( m \geq 0 \),

\[
W_\psi(K_x^m, \Pi) \leq \frac{c(x)}{1 - \gamma} \gamma^m .
\]

**Remark 12.** In the case where \( f \) is Lipschitz and \( \| \cdot \| \) is the Euclidean norm (neither of which is assumed here), Corollary \(11\) is a direct consequence of Steinsaltz’s (1999) Theorem 1.

We now turn our attention to the conversion of Wasserstein divergence bounds into total variation bounds. Here is our main tool.

**Proposition 13.** (Madras and Sezer, 2010) Suppose that \( K(x, \cdot) \), \( x \in X \), admits a density function \( k(x, \cdot) \) with respect to some reference measure \( \mu \). Suppose further that there exists \( c \geq 0 \) such that for all \( x, y \in X \),

\[
\int_X |k(x, x') - k(y, x')| \mu(dx') \leq c \psi(x,y) . \tag{4}
\]

Then for all \( m \geq 1 \) and \( x \in X \),

\[
\|K_x^m - \Pi\|_{TV} \leq \frac{c}{2} W_\psi(K_x^{m-1}, \Pi) .
\]

Madras and Sezer (2010) assume that \( (X, \psi) \) is a Polish metric space and that \( \mathcal{B} \) is the associated Borel \( \sigma \)-algebra, and they use these assumptions in their proof. Again, we have not made any Polish space assumptions, so we provide an alternative proof here.
Proof. An argument from Madras and Sezer’s (2010) proof of their Lemma 13 shows that for two probability measures \( \eta : B \to [0, 1] \) and \( \nu : B \to [0, 1] \),
\[
\| \eta K - \nu K \|_{TV} \leq \frac{1}{2} \int_{X \times X} \int_X |k(x', x'') - k(y', x'')| \mu(dx'') \nu(dx', dy'),
\]
where \( \eta K(\cdot) = \int_X K(x', \cdot) \eta(dx') \), \( \nu K \) is defined analogously, and \( \nu \in \Upsilon(\eta, \nu) \). Now, fix \( x \in X \), let \( \eta = K_{\frac{m-1}{m}} x \), and let \( \nu = \Pi \). Then by (4),
\[
\| K_{\frac{m-1}{m}} x - \Pi \|_{TV} \leq \frac{c}{2} \int_{X \times X} \psi(x', y') \nu(dx', dy').
\]
Since \( \nu \in \Upsilon(\eta, \nu) \) is arbitrary,
\[
\| K_{\frac{m-1}{m}} x - \Pi \|_{TV} \leq \inf_{\nu \in \Upsilon(K_{\frac{m-1}{m}} x, \Pi)} \frac{c}{2} \int_{X \times X} \psi(x', y') \nu(dx', dy') = \frac{c}{2} W_{\psi}(K_{\frac{m-1}{m}} x, \Pi).
\]

Clearly, if the conditions of Proposition 13 are satisfied, then an upper bound on \( \rho_*(\psi) \) also serves as an upper bound on \( \rho^TV_*(\psi) \). For example, it is straightforward to show that (4) holds for the aforementioned autoregressive process when \( \psi \) is the Euclidean distance. Since we know from previous calculations that \( \rho_*(\psi) \leq 1/2 \) for all \( p \), it follows immediately that \( \rho^TV_*(\psi) \leq 1/2 \) for all \( p \). Hence, in this case, a total variation bound converted from a Wasserstein bound is sharp. This is, of course, just a toy example, but recall how poor the d&m bounds are for this toy. In the next section, we use the results developed in this section to analyze the convergence complexity of the A&C algorithm.

4 Convergence Analysis of the Albert and Chib Chain

4.1 Preliminaries

Let \( x_1, x_2, \ldots, x_n \in \mathbb{R}^p \), and let \( y_1, y_2, \ldots, y_n \) be independent binary responses such that for each \( i \in \{1, 2, \ldots, n\} \), \( y_i \sim \text{Bernoulli}(\Phi(x_i^T \beta)) \), where \( \beta \in \mathbb{R}^p \) is an unknown regression coefficient. The design matrix \( X \) is defined to be the \( n \times p \) matrix whose \( i \)th row is \( x_i^T \). Denote the observed data by \( y := (y_1 y_2 \ldots y_n)^T \). Consider a Bayesian analysis based on the prior

\[
\omega(\beta) \propto \exp \left[ -\frac{1}{2}(\beta - v)^T Q(\beta - v) \right], \quad \beta \in \mathbb{R}^p,
\]
where \( v \in \mathbb{R}^p \), and \( Q \) is a \( p \times p \) symmetric matrix that is either positive definite (Gaussian prior) or vanishing (flat prior). While the posterior distribution is automatically proper if \( Q \) is positive-definite, additional...
conditions (on X and y) are required to ensure propriety when Q = 0 (Chen and Shao, 2000). For the time being, we assume that the posterior is proper. The posterior density is, of course, given by

$$\pi(\beta | X, y) \propto n \prod_{i=1}^{n} \Phi(x_i^T \beta)^{y_i}(1 - \Phi(x_i^T \beta))^{1-y_i} \omega(\beta), \beta \in \mathbb{R}^p.$$ 

A standard method of exploring this intractable posterior density is Albert and Chib’s (1993) data augmentation algorithm, which is one of the most well-known MCMC algorithms in Bayesian statistics. We now state the algorithm.

Let $$\Sigma = X^T X + Q$$. Posterior propriety implies that $$\Sigma$$ is non-singular. For $$\mu \in \mathbb{R}, \tau > 0, \text{ and } a \in \{0, 1\}$$, let $$TN(\mu, \tau^2; a)$$ be a normal distribution $$N(\mu, \tau^2)$$ that is truncated to $$(-\infty, 0)$$ if $$a = 0$$, and to $$(0, \infty)$$ if $$a = 1$$. If the current state of the A&C Markov chain is $$\beta_m = \beta$$, then the next state is drawn according to the following procedure.

**Iteration m + 1 of the data augmentation algorithm:**

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

$$\beta_{m+1} \sim N \left( \Sigma^{-1} (X^T Z + Q v), \Sigma^{-1} \right).$$

The Markov transition density of the chain that is simulated by this algorithm is given by

$$k(\beta, \beta') = \int_{\mathbb{R}^n_+} \pi_1(\beta' | z, X, y) \pi_2(z | \beta, X, y) \, dz,$$

where the exact forms of $$\pi_1(\beta' | z, X, y)$$ and $$\pi_2(z | \beta, X, y)$$ can be gleaned from the algorithm. It’s clear that this chain is reversible with respect to the posterior density $$\pi(\cdot | X, y)$$, and that the chain is Harris ergodic. Throughout this section, we will use $$\Pi$$ to denote the stationary measure defined by $$\pi(\cdot | X, y)$$, and $$K(\cdot, \cdot)$$ to denote the Mtfs defined through (5).

Let us now construct a random mapping that induces the A&C Markov chain. For any $$\mu \in \mathbb{R}$$ and $$a \in \{0, 1\}$$, let $$H(\cdot, \mu, a)$$ be the inverse cumulative distribution function of $$TN(\mu, 1; a)$$. Routine calculation shows that

$$H(u, \mu, a) = \mu - \Phi^{-1}(\Phi(\mu)(1 - u)) 1_{\{1\}}(a) + \Phi^{-1}(\Phi(-\mu)u) 1_{\{0\}}(a).$$

Let $$(\Omega, \mathcal{F}, P)$$ be a probability space. Let $$N : \Omega \to \mathbb{R}^p$$ be a p-dimensional standard normal vector, and independently, let $$U : \Omega \to (0, 1)^n$$ be a vector of iid Uniform(0, 1) variables. For $$i \in \{1, 2, \ldots, n\}$$, denote
the $i$th element of $U$ by $U_i$. Consider the random mapping
\[ f(\beta) = \sum_{i=1}^{n} x_i H(U_i, x_i^T \beta, y_i) + \Sigma^{-1} Q v + \Sigma^{-1/2} N, \beta \in \mathbb{R}^p. \]

Clearly, for fixed $\beta$, $f(\beta)$ follows the distribution defined by (5). Hence, $f$ induces $K(\cdot, \cdot)$.

Let $\| \cdot \|_2$ and $\psi_2(\cdot, \cdot)$ denote the Euclidean norm and the corresponding distance, respectively. We will find it convenient to work with a normalized version of $\| \cdot \|_2$. Throughout this section, let $\| \beta \| = (\beta^T \Sigma \beta)^{1/2}$, $\beta \in \mathbb{R}^p$, and denote by $\psi$ the distance function that this norm induces. As described in the previous section, we can use $\psi$ to define a Wasserstein divergence between probability measures on $\mathbb{R}^p$. Furthermore, if we restrict attention to a certain subset of these probability measures, then $W_\psi(\cdot, \cdot)$ is an actual distance. Indeed, let $B$ denote the Borel $\sigma$-algebra on $\mathbb{R}^p$, and define $\mathcal{P}_\psi(\mathbb{R}^p)$ to be the set of probability measures, $\mu$, on $(\mathbb{R}^p, B)$ such that, for some (and hence all) $x \in \mathbb{R}^p$, we have
\[ \int_{\mathbb{R}^p} \psi(x, y) \mu(dy) < \infty. \]

Since $(\mathbb{R}^p, \psi)$ forms a Polish metric space, and $B$ is the associated Borel $\sigma$-algebra, it follows that $W_\psi(\cdot, \cdot)$ restricted to $\mathcal{P}_\psi(\mathbb{R}^p) \times \mathcal{P}_\psi(\mathbb{R}^p)$ is itself a metric, known as the Wasserstein distance. We note that Chen and Shao’s (2000) Theorem 2.3 implies that $\Pi \in \mathcal{P}_\psi(\mathbb{R}^p)$. Later in this section, we will establish that $K_m^{\beta} \in \mathcal{P}_\psi(\mathbb{R}^p)$ for all $m \in \mathbb{N}$ and all $\beta \in \mathbb{R}^p$. In the context of this section, $\rho_*(\psi)$ represents the convergence rate of the A&C Markov chain with respect to the distance $W_\psi$. Note that for any pair of probability measures $\mu$ and $\nu$ in $\mathcal{P}_\psi(\mathbb{R}^p)$, we have
\[ \lambda_{\max}(\Sigma)^{-1/2} W_\psi(\mu, \nu) \leq W_{\psi_2}(\mu, \nu) \leq \lambda_{\min}(\Sigma)^{-1/2} W_\psi(\mu, \nu), \]
where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ give the largest and smallest eigenvalues of a Hermitian matrix (defined through $\| \cdot \|_2$), respectively. Hence, $\rho_*(\psi) = \rho_*(\psi_2)$.

We end this subsection with a result relating Wasserstein and TV bounds for the A&C chain. The proof of this result, which is given in the Appendix, is based on Proposition 13.

**Proposition 14.** For the A&C chain,
\[ \| K_m^{\beta} - \Pi \|_{TV} \leq \frac{1}{\sqrt{2\pi}} W_\psi(K_m^{\beta-1}, \Pi) \]
for all $m \in \mathbb{N}$ and all $\beta \in \mathbb{R}^p$.

An immediate consequence of Proposition 14 is that an upper bound on $\rho_*(\psi)$ is also an upper bound on $\rho_*^{TV}$. This fact will be crucial in our convergence complexity analysis.
4.2 Analysis of the random mapping

We begin our analysis by bounding

\[ D_\beta f := \limsup_{\beta' \to \beta} \frac{\|f(\beta') - f(\beta)\|}{\|\beta' - \beta\|}, \; \beta \in \mathbb{R}^p. \]

Let \( \alpha, \beta \in \mathbb{R}^p \), and assume that \( \alpha \neq 0 \). Then by the mean-value theorem for vector-valued functions,

\[
\frac{\|f(\beta + \alpha) - f(\beta)\|}{\|\alpha\|} = \frac{\left\| \Sigma^{-1/2} \sum_{i=1}^n x_i [H(U_i, x_i^T (\beta + \alpha), y_i) - H(U_i, x_i^T \beta, y_i)] \right\|_2}{\|\Sigma^{-1/2}\alpha\|_2} \\
\leq \left\| \Sigma^{-1/2}\alpha\right\|_2^{-1} \sup_{t \in (0,1)} \left\| \Sigma^{-1/2} \sum_{i=1}^n x_i \frac{\partial H(U_i, \mu, y_i)}{\partial \mu} \bigg|_{\mu=x_i^T(\beta+t\alpha)} x_i^T \alpha \right\|_2. \tag{6}
\]

Now, for \( a \in \{0, 1\} \), we have

\[
\frac{\partial H(u, \mu, a)}{\partial \mu} = s(1 - u, \mu) 1_{\{1\}}(a) + s(u, -\mu) 1_{\{0\}}(a),
\]

where \( s : (0, 1) \times \mathbb{R} \to \mathbb{R} \) is defined as follows:

\[
s(u, \mu) = 1 - \frac{u \phi(\mu)}{\phi(\Phi^{-1}(\Phi(\mu)u))}.
\]

Let \( S(\beta, U) \) be a diagonal matrix whose \( i \)th diagonal component is \( \partial H(U_i, \mu, y_i)/\partial \mu|_{\mu=x_i^T \beta} \). Then \( S(\beta, U) \) is continuous in \( \beta \). It follows from (6) that

\[
D_\beta f = \limsup_{\alpha \to 0} \frac{\|f(\beta + \alpha) - f(\beta)\|}{\|\alpha\|} \\
\leq \limsup_{\alpha \to 0} \sup_{t \in (0,1)} \lambda_{\text{max}} \left[ \Sigma^{-1/2} X^T S(\beta + t\alpha, U) XX^{-1/2} \right] \\
= \lambda_{\text{max}} \left[ \Sigma^{-1/2} X^T S(\beta, U) XX^{-1/2} \right] =: \hat{D}_\beta f.
\]

Remark 15. Note that \( \hat{D}_\beta f \) is continuous in \( \beta \) for each value of \((U, N)\). (In fact, \( \hat{D}_\beta f \) does not depend on \( N \).) Therefore, as a function of \( t \in [0, 1] \), \( \hat{D}_{\beta+t(\beta-\alpha)} f \) is Riemann integrable for all \( \alpha, \beta \in \mathbb{R}^p \), almost surely.

The following lemma, which is proved in the Appendix, will help us develop a bound on \( \hat{D}_\beta f \).

Lemma 16. For \( u \in (0, 1) \) and \( \mu \in \mathbb{R} \), \( s(u, \mu) \in (0, 1) \). Moreover, if \( \mu \leq 0 \), \( s(u, \mu) \leq 1 - u \).

The next lemma shows that \( D_\beta f \) and \( \hat{D}_\beta f \) are bounded by 1.

Lemma 17. For any \( \beta \in \mathbb{R}^p \) and \( \bar{u} \in (0, 1)^n \),

\[
\lambda_{\text{max}} \left( \Sigma^{-1/2} X^T S(\beta, \bar{u}) XX^{-1/2} \right) < 1.
\]

Thus, \( D_\beta f \leq \hat{D}_\beta f < 1 \) for all values of \((U, N)\) and \( \beta \).
Proof. By Lemma 16, all the diagonal components of \( S(\beta, \bar{u}) \) are strictly less than 1. For two Hermitian matrices of the same size, \( M_1 \) and \( M_2 \), write \( M_1 \succ M_2 \), or equivalently, \( M_2 \preceq M_1 \) if \( M_1 - M_2 \) is nonnegative definite. Similarly, write \( M_2 \prec M_1 \) if \( M_1 - M_2 \) is positive definite. We consider two cases: (i) \( Q = 0 \), and (ii) \( Q \) is positive definite. If \( Q = 0 \), then \( \Sigma = X^T X \), which is positive definite (because of propriety). It follows that

\[
\Sigma^{-1/2} X^T S(\beta, \bar{u}) X \Sigma^{-1/2} \prec \Sigma^{-1/2} X^T X \Sigma^{-1/2} = I_p.
\]

If \( Q \) is positive definite, then

\[
\Sigma^{-1/2} X^T S(\beta, \bar{u}) X \Sigma^{-1/2} \preceq \Sigma^{-1/2} X^T X \Sigma^{-1/2} \prec I_p.
\]

Thus, in either case, \( \Sigma^{-1/2} X^T S(\beta, \bar{u}) X \Sigma^{-1/2} \prec I_p \), and the result follows from Weyl’s inequality.

We now develop an approximation of \( \mathbb{E}\tilde{D}_\beta f = \mathbb{E}\lambda_{\max} \left( \Sigma^{-1/2} X^T S(\beta, U) X \Sigma^{-1/2} \right) \) using

\[
\hat{\gamma}(\beta) := \lambda_{\max} \left( \Sigma^{-1/2} X^T \mathbb{E}S(\beta, U) X \Sigma^{-1/2} \right),
\]

which is a much more tractable quantity. A proof of the following proposition is given in the Appendix.

**Proposition 18.** For \( \beta \in \mathbb{R}^p \), \( \mathbb{E}\tilde{D}_\beta f \) is bounded above by \( \hat{\gamma}(\beta) + \sigma(2 \log p)^{1/2} \), where

\[
\sigma = \lambda_{\max}^{1/2} \left[ \sum_{i=1}^{n} (\Sigma^{-1/2} x_i x_i^T \Sigma^{-1/2})^2 \right].
\]

As a result, \( \sup_{\beta \in \mathbb{R}^p} \mathbb{E}\tilde{D}_\beta f \) is bounded above by \( \sup_{\beta \in \mathbb{R}^p} \hat{\gamma}(\beta) + \sigma(2 \log p)^{1/2} \).

**Remark 19.** The error term \( \sigma(2 \log p)^{1/2} \) is typically small when \( n \) is a lot larger than \( p \). As an illustration, suppose that there exist positive constants \( \ell_1 \) and \( \ell_2 \) such that \( \|x_i\|_2^2 \leq \ell_1 p \) for all \( i \), and \( \lambda_{\min}(\Sigma) \geq \ell_2 n \). Then

\[
\sigma^2 = \lambda_{\max} \left[ \sum_{i=1}^{n} (x_i^T \Sigma^{-1} x_i^T \Sigma^{-1} x_i^T \Sigma^{-1}) \right] \leq \frac{\ell_1 p}{\ell_2 n} \lambda_{\max}(\Sigma^{-1/2} X^T X \Sigma^{-1/2}) \leq \frac{\ell_1 p}{\ell_2 n}.
\]

For \( \beta \in \mathbb{R}^p \), \( \mathbb{E}S(\beta, U) \) is an \( n \times n \) diagonal matrix whose \( i \)th diagonal element is \( \int_0^1 s(u, x_i^T \beta) \, du \) if \( y_i = 1 \), and \( \int_0^1 s(u, -x_i^T \beta) \, du = y_i = 0 \). A bit of calculation reveals that for \( \mu \in \mathbb{R} \), we have

\[
\int_0^1 s(u, \mu) \, du = 1 - \frac{\phi(\mu)}{\Phi^2(\mu)} \int_{-\infty}^{\mu} \Phi(x) \, dx.
\]

The following result, which is a direct consequence of Lemma 16, will be used extensively later in this section.
Lemma 20. For \( \mu \in \mathbb{R} \), \( \int_0^1 s(u, \mu) \, du \in (0, 1) \). Moreover, if \( \mu \leq 0 \), then \( \int_0^1 s(u, \mu) \, du \in (0, 1/2] \).

We are now ready to examine the convergence properties of the A&C Markov chain. In Subsection 4.3 we use Proposition 7 to establish a bound on the chain’s Wasserstein distance to stationarity, and in Subsection 4.4 we use Corollary 11 to perform a convergence complexity analysis.

4.3 A bound on the Wasserstein distance for a fixed data set

Q&H developed a drift condition for the A&C chain that is a substantial improvement upon previous such conditions established in Roy and Hobert (2007) and Chakraborty and Khare (2017). Let \( V : \mathbb{R}^p \to [0, \infty) \) be defined as
\[
V(\beta) = \| \beta - \hat{\beta} \|^2,
\]
where \( \hat{\beta} \) is the (unique) mode of \( \pi(\beta | X, y) \). Define \( \Lambda(\beta), \beta \in \mathbb{R}^p \), to be an \( n \times n \) diagonal matrix whose \( i \)th diagonal element is
\[
1 - \left[ \frac{x_i^T \phi(x_i^T \beta)}{\Phi(x_i^T \beta)} + \left( \frac{\phi(x_i^T \beta)}{\Phi(x_i^T \beta)} \right)^2 \right] \mathbf{1}_{\{y_i\}} - \left[ \frac{-x_i^T \phi(x_i^T \beta)}{1 - \Phi(x_i^T \beta)} + \left( \frac{\phi(x_i^T \beta)}{1 - \Phi(x_i^T \beta)} \right)^2 \right] \mathbf{1}_{\{0\}}(y_i).
\]

Here is the drift condition.

Lemma 21. (Qin and Hobert, 2019+) For every \( \beta \in \mathbb{R}^p \),
\[
\int_{\mathbb{R}^p} V(\beta') k(\beta, \beta') \, d\beta' \leq \lambda V(\beta) + 2p,
\]
where
\[
\lambda = \sup_{t \in (0, 1)} \sup_{\alpha \in \mathbb{R}^p \setminus \{0\}} \frac{\| \Sigma^{-1} X^T \Lambda(\hat{\beta} + t\alpha) X \alpha \|^2}{\|\alpha\|^2} < 1.
\]

With this drift condition in hand, we can now employ Proposition 7 to get an upper bound on the Wasserstein distance to stationarity. But first, we use the drift condition to demonstrate that \( K^m_{\beta} \in \mathcal{P}_\psi(\mathbb{R}^p) \) for \( m \in \mathbb{N} \) and all \( \beta \in \mathbb{R}^p \). Indeed, fix \( m \geq 1 \) and \( \beta \in \mathbb{R}^p \), and note that
\[
\int_{\mathbb{R}^p} \psi(\beta', \hat{\beta}) K^m(\beta, d\beta') \leq \left( \int_{\mathbb{R}^p} V(\beta') k^m(\beta, d\beta') \right)^{1/2} \leq \left( \lambda^m V(\beta) + \frac{2p}{1 - \lambda} \right)^{1/2} < \infty,
\]
where the first inequality is Cauchy-Schwarz, and the second follows from Lemma 21. Here is the main result of this subsection.

Proposition 22. The A&C Markov chain converges geometrically in \( W_\psi \) (and thus in \( W_{\psi^2} \) as well). More specifically, for a chain starting at \( \beta \in \mathbb{R}^p \),
\[
W_\psi(K^m_{\beta}, \Pi) \leq c_1(\beta) \gamma^{am} + c_2(\beta) \left\{ \left( \frac{1 + 4p + \lambda d}{1 + d} \right)^{1-a} \{1 + 2(\lambda d + 2p)\}^a \right\}^m
\]
for each \( m \geq 0 \) and any \( a \in (0, 1) \), where \( \lambda \) is given in Lemma 21, \( d > 4p/(1 - \lambda) \),

\[
\gamma = \sup_{\{\beta : \|\beta - \hat{\beta}\|^2 \leq d\}} \mathbb{E} \lambda_{\max} \left( \Sigma^{-1/2} X^T S(\beta, U) X \Sigma^{-1/2} \right) < 1 ,
\]

\[
c_1(\beta) = \int_{\mathbb{R}^p} \|\beta - \beta'||\pi(\beta'|X, y)\ d\beta' < \infty, \text{ and}
\]

\[
c_2(\beta) = \int_{\mathbb{R}^p} \left( 1 + \|\beta - \hat{\beta}\|^2 + \|\beta' - \hat{\beta}\|^2 \right) \|\beta - \beta'||\pi(\beta'|X, y)\ d\beta' < \infty .
\]

**Proof.** We simply make use of Proposition 7. Condition \((A1)\) is implied by Lemma 21. Let \( C = \{\beta \in \mathbb{R}^p : \|\beta - \hat{\beta}\|^2 \leq d\} \). By Lemma 17 and the compactness of \( C \),

\[
\gamma = \sup_{\beta \in C} \mathbb{D}\beta f \leq \mathbb{E} \sup_{\beta \in C} \mathbb{D}\beta f < 1 .
\]

Thus, \((A2')\) follows from Lemma 10 (and Remark 15). Condition \((A3)\) follows from Lemmas 9 and 17. Finally, \( c_1(\beta), c_2(\beta) < \infty \) for all \( \beta \in \mathbb{R}^p \) by Chen and Shao’s (2000) Theorem 2.3.

Of course, Proposition 22 gives a non-trivial upper bound on \( \rho_*(\psi) \), namely

\[
\hat{\rho}_0 = \gamma^a \vee \left( \frac{1 + 4p + \lambda d}{1 + d} \right)^{1-a} \left[ 1 + 2(\lambda d + 2p) \right]^a ,
\]

where \( \gamma, \lambda, d, \) and \( a \) are given in the said proposition.

### 4.4 Convergence complexity analysis

In the previous three subsections, we studied the A&C Markov chain associated with an arbitrary fixed data set \((X, y)\). Posterior propriety was simply assumed, which is reasonable because there is a well-known set of easily checked conditions that are necessary and sufficient for posterior propriety (Chen and Shao, 2000). In the current subsection, we will consider the asymptotic behavior of the convergence rates of an infinite sequence of A&C Markov chains corresponding to an infinite sequence of data sets (that are growing in size). As we shall see, some of the data sets early in the sequence may have design matrices, \( X \), that are rank deficient. Of course, if \( Q = 0 \) and \( X \) is rank deficient, then not only is the corresponding posterior improper, but the A&C algorithm is not even defined (since \( \Sigma \) is singular). Fortunately, all we require for our asymptotic analysis is that, \textit{eventually}, \( \Sigma \) is non-singular and the posterior is proper. We now state a version of Corollary 11 that is specific to the A&C Markov chain.

**Proposition 23.** Suppose that \( \Sigma \) is non-singular, and suppose further that

\[
\hat{\rho}_1 := \sup_{\beta \in \mathbb{R}^p} \mathbb{E} \lambda_{\max} \left( \Sigma^{-1/2} X^T S(\beta, U) X \Sigma^{-1/2} \right) < 1 .
\]
Then the posterior is proper, and for each $\beta \in \mathbb{R}^p$ and $m \geq 0$,

$$W_\psi(K_\beta^m, \Pi) \leq \frac{c(\beta)}{1 - \hat{\rho}_1^m},$$

where $c(\beta) = \int_{\mathbb{R}^p} \|\beta - \beta'\| K(\beta, d\beta') < \infty$.

**Proof.** It suffices to prove that $\hat{\rho}_1 < 1$ implies posterior propriety. The rest follows immediately from Corollary 11. Without assuming posterior propriety, it can be checked that $c(\alpha) < \infty$ for all $\alpha \in \mathbb{R}^p$. Now fix $\alpha \in \mathbb{R}^p$. As in the proof of Proposition 6, for each $\beta \in \mathbb{R}^p$,

$$\int_{\mathbb{R}^p} \psi(\beta', \alpha) K(\beta, d\beta') \leq \hat{\rho}_1 \psi(\beta, \alpha) + c(\alpha).$$

Now, $\hat{\rho}_1 < 1$, $c(\alpha) < \infty$, and the function $\tilde{V}(\cdot) = \psi(\cdot, \alpha) + 1$ is unbounded off petite sets. Therefore, the inequality above constitutes a geometric drift condition. It follows that the Markov chain is geometrically ergodic (Meyn and Tweedie, 2012, Theorem 15.2.8), and this in turn implies that its invariant measure (defined by the unnormalized posterior density) has finite mass. \qed

We now consider three different asymptotic regimes.

### 4.4.1 Letting $n$ diverge with fixed but arbitrarily large $p$

Consider an array of data sets $\{D_{p,n}\}_{p,n}$, where $D_{p,n}$ consists of a design matrix $X_{n \times p}$ and a vector of binary responses $y_{n \times 1}$. As in Section 5 of Q&H, we assume that these data sets are generated according to a random mechanism that is consistent with the probit regression model. Indeed, for each fixed $p$, suppose that $\{D_{p,n}\}_{n}$ is generated sequentially as follows. Let $\{x_i\}_{i=1}^\infty$ be a sequence of $p$-dimensional vectors that are independently generated according to some common distribution. Given $\{x_i\}_{i=1}^\infty$, generate $\{y_i\}_{i=1}^\infty$ independently according to $y_i \sim \text{Bernoulli}(G(x_i^T \beta_*)$, where $G : \mathbb{R} \to (0, 1)$ is an inverse link function that is independent of $p$, and $\beta_* \in \mathbb{R}^p$ is the true coefficient. Finally, set the $i$th row of $X_{n \times p}$ to be $x_i^T$, and set the $i$th element of $y_{n \times 1}$ to be $y_i$. Thus, when $p$ is fixed, whenever $n$ is increased by 1, a new $p \times 1$ covariate vector and a corresponding binary response are added to the data set. Of course, it is assumed that these random data sets are completely unrelated to the random vectors used to construct the random mappings described in Subsection 4.1. To make this clear, we use $\tilde{P}$ and $\tilde{E}$ to denote probability and expectation for the random data sets.

Now for each $p$, assign a prior with parameters $Q = Q_p$ and $v = v_p$. Then, for every given $p$, we have a sequence of posterior distributions and a corresponding sequence of A&C chains indexed by $n$. Q&H used d&m to prove that for every $p$, under mild conditions, $\limsup_{n \to \infty} \rho_1^{-n} \leq \rho_p$ for some $\rho_p < 1$, almost surely.
Hence, for any fixed $p$, the convergence rate of the A&C chain is bounded below 1 as $n$ grows. However, similar to what happens in Proposition 4, their asymptotic upper bounds $\rho_p$ converge to 1 exponentially fast as $p \to \infty$. In what follows, we use the Wasserstein-to-TV approach to extend their result by providing an upper bound on $\limsup_{n \to \infty} \rho^*_n$ that, under the appropriate conditions, does not depend on $p$.

We now state our assumptions concerning the random data sets.

(B1) For each $p$, $0 < \lambda_{\min}(\bar{E}x_1x_1^T) \leq \lambda_{\max}(\bar{E}x_1x_1^T) < \infty$.

(B2) For each $p$, $\bar{E}\|x_1\|^2 < \infty$.

(B3) For each $p$, $\lim_{\delta \to 0} \sup_{\alpha \in \mathbb{R}^p \setminus \{0\}} \bar{P}(|x_1^T \alpha| \leq \delta \|x_1\|) = 0$.

(B4) There exists a constant $\ell > 0$ (that does not depend on $p$) such that for each $p$, 

$$\lambda_{\min}\left[\left(\bar{E}x_1x_1^T\right)^{-1/2}(\bar{E}x_1x_1^TG(x_1^T \beta_*)^\dagger(\bar{E}x_1x_1^T)^{-1/2}\right) \geq \ell,$$

where $\bar{G}(x) = G(x) \wedge (1 - G(x))$ for $x \in \mathbb{R}$.

Assumptions (B1) and (B2) are standard. (B3) ensures that the distribution of $x_1$ is not overly concentrated on any $(p - 1)$-dimensional subspace (that passes through the origin). Note that this condition is satisfied by most common spherically symmetric distributions, e.g., $\mathcal{N}(0, I_p)$. If $G$ is a monotone function such that $G(0) = 1/2$ and $G(-\mu) = 1 - G(\mu)$ for all $\mu \in \mathbb{R}$, then $\bar{G}(x_1^T \beta_*)$ is decreasing in $|x_1^T \beta_*$, and (B4) serves as a sparsity condition on $\beta_*$. For illustration, suppose that $x_1 \sim \mathcal{N}(0, I_p)$, and that only the first $k$, $k < p$, elements of $\beta_* \in \mathbb{R}^p$ are non-vanishing. Denote the first $k$ elements of $\beta_*$ by $\beta_{(k)} \in \mathbb{R}^k$, and the first $k$ elements of $x_1$ by $x_{(k)} \in \mathbb{R}^k$. Then $x_{(k)} \sim \mathcal{N}(0, I_k)$, and

$$(\bar{E}x_1x_1^T)^{-1/2}(\bar{E}x_1x_1^TG(x_1^T \beta_*))((\bar{E}x_1x_1^T)^{-1/2} = \text{Diag}\left(\bar{E}x_{(k)}x_{(k)}^T \bar{G}\left(x_{(k)}^T \beta_{(k)}\right), \bar{E}G\left(x_{(k)}^T \beta_{(k)}\right) I_{p-k}\right).$$

This shows that 

$$\lambda_{\min}\left[\left(\bar{E}x_1x_1^T\right)^{-1/2}(\bar{E}x_1x_1^TG(x_1^T \beta_*)^\dagger(\bar{E}x_1x_1^T)^{-1/2}\right] = \min\left\{\lambda_{\min}\left[\bar{E}x_{(k)}x_{(k)}^T \bar{G}\left(x_{(k)}^T \beta_{(k)}\right)\right], \bar{E}G\left(x_{(k)}^T \beta_{(k)}\right) I_{p-k}\right\}. \quad (7)$$

One can show that $\bar{E}x_{(k)}x_{(k)}^T \bar{G}(x_{(k)}^T \beta_{(k)})$ is positive definite. If $\beta_{(k)}$ is fixed for all $p$, then the right hand side of (7) is a positive constant, and (B4) is satisfied.

**Proposition 24.** Suppose that (B1) – (B4) hold. Then there exists a constant $\rho < 1$ that does not depend on $p$ and is such that, for each $p$, almost surely,

$$\limsup_{n \to \infty} \rho_*(\psi) \leq \rho .$$

By Proposition 4, the result continues to hold if $\rho_*(\psi)$ is replaced by $\rho^*_n$. 

20
Proof. By Propositions 23 and 18, if Σ is non-singular and \( \hat{\rho}_2 := \sup_{\beta \in \mathbb{R}^p} \hat{\gamma}(\beta) + \sigma(2 \log p)^{1/2} < 1 \), then \( \Pi \) is a proper probability distribution, and \( \rho_*(\psi) \leq \hat{\rho}_2 \). Here, \( \sigma = \lambda_{\text{max}}^{1/2} \left[ \sum_{i=1}^n (\Sigma^{-1/2} x_i x_i^T \Sigma^{-1/2})^2 \right] \)

\[
\hat{\gamma}(\beta) = \lambda_{\text{max}} \left( \Sigma^{-1/2} \left( \sum_{i=1}^n x_i \xi(x_i^T \beta, y_i) x_i^T \right) \Sigma^{-1/2} \right),
\]

and

\[
\xi(\mu, v) = \int_0^1 s(u, \mu) \, du \, 1_{\{1\}}(v) + \int_0^1 s(u, -\mu) \, du \, 1_{\{0\}}(v), \quad \mu \in \mathbb{R}, \ v \in \{0, 1\}.
\]

For the rest of this proof, let \( p \) be arbitrarily fixed. By condition \((B1)\), almost surely, for all large \( n \), \( \Sigma \) will be positive-definite. We will first show that

\[
\limsup_{n \to \infty} \sigma(2 \log p)^{1/2} = 0,
\]

almost surely. This will ensure that with probability 1, for large enough \( n \), \( \Pi \) is proper, and \( \rho_*(\psi) \) is bounded above by any number strictly greater than \( \rho \).

Note that

\[
\sigma^2 \leq \text{trace} \left[ \sum_{i=1}^n (\Sigma^{-1/2} x_i x_i^T \Sigma^{-1/2})^2 \right]
\]

\[
\leq \lambda_{\text{min}}^{-2}(\Sigma) \sum_{i=1}^n \|x_i\|^4_2
\]

\[
= \frac{1}{n} \lambda_{\text{min}}^{-2} \left( \frac{1}{n} \sum_{i=1}^n x_i x_i^T + \frac{Q}{n} \right) \frac{1}{n} \sum_{i=1}^n \|x_i\|^4_2.
\]

By the law of large numbers, \((B1)\), and \((B2)\), almost surely,

\[
\limsup_{n \to \infty} n \sigma^2 \leq \lambda_{\text{min}}^{-2}(\mathbb{E} x_1 x_1^T) \mathbb{E} \|x_1\|^4_2 < \infty.
\]

As a result, with probability 1, \( \lim_{n \to \infty} \sigma(2 \log p)^{1/2} = 0 \).

Assume \( n \) is large enough so that \( \Sigma_0 := X^T X \) is positive-definite. For \( \delta \in [0, 1] \) and \( \alpha \in \mathbb{R}^p \setminus \{0\} \), let \( J(\alpha, \delta) = \{ \beta \in \mathbb{R}^p : \alpha^T \beta \geq \delta \|\alpha\|_2 \|\beta\|_2 \} \). For every \( \delta \in (0, 1] \), one can pick a finite number of vectors, say \( \alpha_1, \alpha_2, \ldots, \alpha_m(\delta) \), such that \( \mathbb{R}^p = \bigcup_{j=1}^m J(\alpha_j, \sqrt{1 - \delta^2}) \). By Lemma 35 in the appendix, if \( \alpha \in J(\alpha_j, \sqrt{1 - \delta^2}) \) and \( \beta \in J(\alpha_j, \delta) \) for some \( j \), then \( \alpha^T \beta \geq 0 \). Using this fact and Lemma 20 we see
that for $\delta \in (0, 1]$,

$$
\sup_{\beta \in \mathbb{R}^p} \hat{\gamma}(\beta) = \sup_{1 \leq j \leq m(\delta)} \max_{\beta \in J_j} \left\{ \frac{\sum_{i=1}^n x_i \xi(x_i^T \beta, y_i) x_i^T \sum_{i=1}^n x_i^T (1 - \xi(x_i^T \beta, y_i)) x_i \Sigma_0^{-1/2}}{\sum_{i=1}^n x_i^T 1_{[0, \delta \|x_i\|_2 \|\alpha_j\|_2]} |x_i^T \alpha_j|} \right\}.
$$

Note that we have applied Weyl’s inequality multiple times. For $i = 1, 2, \ldots$ and $j = 1, 2, \ldots, m(\delta)$,

$$
\tilde{E}x_i x_i^T \left[ 1_{(-\infty, 0)}(x_i^T \alpha_j) 1_{\{1\}}(y_i) + 1_{[0, \infty)}(x_i^T \alpha_j) 1_{\{0\}}(y_i) \right] = \tilde{E}x_i x_i^T \left\{ 1_{(-\infty, 0)}(x_i^T \alpha_j) G(x_i^T \beta_s) + 1_{[0, \infty)}(x_i^T \alpha_j) \left[ 1 - G(x_i^T \beta_s) \right] \right\}.
$$

Applying the strong law reveals that almost surely,

$$
\limsup_{n \to \infty} \limsup_{\beta \in \mathbb{R}^p} \hat{\gamma}(\beta) \leq 1 - \frac{1}{2} \lambda_{\min} (\tilde{E}x_1 x_1) - \frac{1}{2} \lambda_{\min} \left( \tilde{E}x_1 x_1^T \tilde{G}(x_1^T \beta_s) \right) \max_{1 \leq j \leq m(\delta)} \lambda_{\max} \left( \tilde{E}x_1 x_1^T 1_{[0, \delta \|x_1\|_2 \|\alpha_j\|_2]} |x_1^T \alpha_j| \right)
$$

where the last line follows from Cauchy-Schwarz and the fact that for every $j$,

$$
\lambda_{\max} \left( \tilde{E}x_1 x_1^T 1_{[0, \delta \|x_1\|_2 \|\alpha_j\|_2]} |x_1^T \alpha_j| \right) \leq \text{trace} \left\{ \tilde{E} x_1 x_1^T 1_{[0, \delta \|x_1\|_2 \|\alpha_j\|_2]} |x_1^T \alpha_j| \right\} = \tilde{E} \left\| x_1 \right\|_2^2 1_{[0, \delta \|x_1\|_2 \|\alpha_j\|_2]} |x_1^T \alpha_j| \right\}.
$$

Let $\delta \to 0$. Then it follows from (B1) – (B3) that almost surely,

$$
\limsup_{n \to \infty} \limsup_{\beta \in \mathbb{R}^p} \hat{\gamma}(\beta) \leq 1 - \frac{1}{2} \lambda_{\min} \left( (\tilde{E}x_1 x_1^T G(x_1^T \beta_s)) (\tilde{E}x_1 x_1) \right).
$$

Since $p$ is arbitrary, the result follows from (B4). □
4.4.2 Arbitrary $n$ and $p$ with shrinkage priors

Consider another array of data sets, $\{D_{p,n}\}_{p,n}$, where each one has a corresponding prior with parameters $(Q_{p,n}, v_{p,n})$. This time we make no assumptions on the structure of the data, and study the effect that a shrinkage prior on $\beta$ has on $\rho^*(\psi)$. For convenience, we will suppress the subscripts of $X_{n \times p}$, $Q_{p,n}$, and $v_{p,n}$.

We study the behavior of $\rho^*(\psi)$ under the following assumptions.

(C1) For every $p$ and $n$, $Q$ is positive-definite.

(C2) There exists a constant $\ell < \infty$ such that $\lambda_{\max}(XQ^{-1}X^T) \leq \ell$ for all $p$ and $n$.

Condition (C1) ensures posterior propriety in all cases. Condition (C2) holds when the precision matrix of the Gaussian prior, $Q$, has sufficiently large eigenvalues, which means that the prior provides a sufficiently strong shrinkage towards its mean $v$. When (C1) holds, (C2) is equivalent to the existence of a constant $\ell' < 1$ such that $\lambda_{\max}(X\Sigma^{-1}X^T) \leq \ell'$ (Chakraborty and Khare, 2017). Examples of priors that satisfy (C1) and (C2) can be found in, e.g., Gupta and Ibrahim (2007), Yang and Song (2009), and Baragatti and Pommeret (2012).

Using d&m-based convergence rate bounds from Q&H, one can show that under (C1) and (C2), for every fixed $n$, $\limsup_{p \to \infty} \rho^*_n \leq \rho_n$ for some $\rho_n < 1$. However, it’s easy to verify that these bounds converge to 1 as $n \to \infty$. Taking the Wasserstein-to-TV approach, we have the following improved result.

**Proposition 25.** If (C1) and (C2) hold, then there exists a constant $\rho < 1$, such that for all $n$ and $p$, $\rho^*(\psi) \leq \rho$. By Proposition 14 the result continues to hold if $\rho^*(\psi)$ is replaced by $\rho^*_n$.

**Proof.** By Proposition 23,

$$\rho^*(\psi) \leq \sup_{\beta \in \mathbb{R}^p} \mathbb{E}\lambda_{\max}\left(\Sigma^{-1/2}X^T S(\beta, U) X \Sigma^{-1/2}\right).$$

It was shown in Lemma 16 that $S(\beta, U)$ is a diagonal matrix whose diagonal components are all less than 1 for every value of $\beta$ and $U$. Hence,

$$\sup_{\beta \in \mathbb{R}^p} \mathbb{E}\lambda_{\max}\left(\Sigma^{-1/2}X^T S(\beta, U) X \Sigma^{-1/2}\right) \leq \lambda_{\max}\left(\Sigma^{-1/2}X^T X \Sigma^{-1/2}\right).$$

Note that $\Sigma^{-1/2}X^T X \Sigma^{-1/2}$ has the same eigenvalues as $X \Sigma^{-1}X^T$. Therefore,

$$\sup_{\beta \in \mathbb{R}^p} \mathbb{E}\lambda_{\max}\left(\Sigma^{-1/2}X^T S(\beta, U) X \Sigma^{-1/2}\right) \leq \lambda_{\max}(X \Sigma^{-1}X^T),$$

and the result follows from (C2).
The argument above shows that $\rho_*(\psi)$ is bounded above by $\lambda_{\text{max}}(\Sigma^{-1/2}X^TX\Sigma^{-1/2})$, which is always less than 1 when $Q$ is positive definite. Loosely speaking, as the eigenvalues of $Q$ increase, this bound tends to decrease, suggesting that priors providing strong shrinkage tend to yield fast convergence.

### 4.4.3 Letting $n$ and $p$ both diverge in a repeated measures design

We now consider the case where $n$ and $p$ grow simultaneously for a particular type of data. Let $q$ be an integer no less than $p$, and let $r$ be a positive integer. Assume that there are $q$ distinct rows in the design matrix $X_{n \times p}$, and that each distinct row is repeated at least $r$ times. Let $r_i$ be the number of repetitions for the $i$th distinct row, which we denote by $\tilde{x}_i^T$, so $\Sigma_{i=1}^q r_i = n$. Now, consider a deterministic sequence of design matrices $\{X_{n(k) \times p(k)}\}_{k=1}^\infty$ that satisfy the above assumptions, where $k$ is an index such that $q(k), r(k) \to \infty$ as $k$ tends to infinity. For each $k$, generate a random vector of binary responses as follows. Independently, for each $i = 1, 2, \ldots, q$ and $j = 1, 2, \ldots, r_i$, let $y_{ij} \sim \text{Bernoulli}(G(\tilde{x}_i^T \beta_*))$ be the response corresponding to the $j$th copy of the $i$th distinct row of $X_{n \times p}$, where $G : \mathbb{R} \to (0, 1)$ is a function independent of $k$, and $\beta_* \in \mathbb{R}^p$ is the true regression coefficient. Finally, for each $k$, assign a prior with parameters $Q = Q_k$ and $v = v_k$. We will study the asymptotic behavior of $\rho_*(\psi)$ as $k \to \infty$.

For two variables $a := a(k)$ and $b := b(k)$ such that $a, b > 0$, write $a \gg b$, or equivalently, $b \ll a$ if $b/a \to 0$ as $k \to \infty$. The assumptions that we make are as follows.

$(D1)$ For every $k$, $\Sigma$ is non-singular.

$(D2)$ $\lim \sup_{k \to \infty} \sigma(2 \log p)^{1/2} = 0$, where $\sigma$ is defined in Proposition [18]

$(D3)$ $r \gg \log q$.

$(D4)$ For every $k$, $\max_{1 \leq i \leq q} |\tilde{x}_i^T \beta_*| < \ell$, where $\ell$ is a constant (independent of $k$) such that

$$\inf_{|\mu| < \ell} |G(\mu) \wedge (1 - G(\mu))| > 0.$$  

Condition $(D1)$ ensures that the A&C chain is well-defined. It follows from Remark [19] that if, for every $k$, $\{||\tilde{x}_i||_2^2/p\}_{i=1}^q$ are bounded above by a positive constant $\ell_1$, and $\lambda_{\text{min}}(\Sigma)/n$ is bounded below by a positive constant $\ell_2$, then $(D2)$ is satisfied if $n \gg p \log p$. Note that since $q \geq p$, and $n \geq qr$, the condition $n \gg p \log p$ is a minimal requirement for $(D3)$. $(D4)$ serves as a sparsity condition.

**Proposition 26.** If $(D1)$ – $(D4)$ hold, then there exists a constant $\rho < 1$ that does not depend on $k$ and is such that, for any $\epsilon > 0$,

$$\lim_{k \to \infty} \tilde{P}(\rho_*(\psi) > \rho + \epsilon) = 0.$$
By Proposition 14, the result continues to hold if \( \rho_*(\psi) \) is replaced by \( \rho_*^{TV} \).

**Remark 27.** Note that, in contrast with Section 4.4.4, here the design matrix \( X \) is deterministic, and the randomness in \( \rho_*(\psi) \) is entirely due to the response vector \( y \).

**Proof of Proposition 26.** By Condition \((D1)\) along with Propositions 23 and 18, if \( \hat{\rho}_2 := \sup_{\beta \in \mathbb{R}^p} \hat{\gamma}(\beta) + \sigma (2 \log p)^{1/2} < 1 \), then \( \Pi \) is a proper probability distribution, and \( \rho_*(\psi) \leq \hat{\rho}_2 \). Since \((D2)\) holds, it suffices to show that there exists a constant \( \rho < 1 \) such that for any \( \epsilon > 0 \),

\[
\lim_{k \to \infty} \hat{P}\left( \sup_{\beta \in \mathbb{R}^p} \hat{\gamma}(\beta) > \rho + \epsilon \right) = 0.
\]

Note that for \( \beta \in \mathbb{R}^p \),

\[
\hat{\gamma}(\beta) = \lambda_{\text{max}} \left[ \Sigma^{-1/2} \left( \sum_{i=1}^{q} \sum_{j=1}^{r_i} \bar{x}_i \xi(\bar{x}_i^T \beta, y_{ij}) \bar{x}_i^T \right) \Sigma^{-1/2} \right],
\]

where as before,

\[
\xi(\mu, v) = \int_0^1 s(u, \mu) \, du \mathbb{1}_{(1)}(v) + \int_0^1 s(u, -\mu) \, du \mathbb{1}_{(0)}(v), \quad \mu \in \mathbb{R}, \quad v \in \{0, 1\}.
\]

For \( i = 1, 2, \ldots, q \), let

\[
v_i = \left( \frac{1}{r_i} \sum_{j=1}^{r_i} y_{ij} \right) \land \left( 1 - \frac{1}{r_i} \sum_{j=1}^{r_i} y_{ij} \right).
\]

By Lemma 20, for each \( i \) and \( \beta \in \mathbb{R}^p \),

\[
\sum_{j=1}^{r_i} \xi(\bar{x}_i^T \beta, y_{ij}) \leq r_i - \frac{1}{2} v_i r_i.
\]

As a result,

\[
\Sigma^{-1/2} \left( \sum_{i=1}^{q} \sum_{j=1}^{r_i} \bar{x}_i \xi(\bar{x}_i^T \beta, y_{ij}) \bar{x}_i^T \right) \Sigma^{-1/2} \lesssim \Sigma^{-1/2} \left[ \sum_{i=1}^{q} \left( 1 - \frac{1}{2} v_i \right) \left( r_i \bar{x}_i \bar{x}_i^T + \frac{r_i}{n} Q \right) \right] \Sigma^{-1/2} \lesssim \left( 1 - \frac{1}{2} \min_{1 \leq i \leq q} v_i \right) I_p.
\]

Thus, \( \sup_{\beta \in \mathbb{R}^p} \hat{\gamma}(\beta) \leq (1 - \min_{1 \leq i \leq q} v_i/2) \).
Now, for any \( \epsilon > 0 \), by Hoeffding’s inequality,

\[
\tilde{P} \left( \min_{1 \leq i \leq q} \frac{1}{r_i} \sum_{j=1}^{r_i} y_{ij} - G(\tilde{x}_i^T \beta_*) \leq \epsilon \right) \geq \prod_{i=1}^{q} \tilde{P} \left( \frac{1}{r_i} \sum_{j=1}^{r_i} y_{ij} - G(\tilde{x}_i^T \beta_*) \leq \epsilon \right) \geq \prod_{i=1}^{q} [1 - 2 \exp(-2r_i \epsilon^2)] = [1 - 2 \exp(-2r \epsilon^2)]^q.
\]

Note that \([1 - 2 \exp(-2r \epsilon^2)]^q \to 1\) if \(\log q - 2\epsilon^2 r \to -\infty\), which holds because of \((D3)\). By \((D4)\),

\[
\min_{1 \leq i \leq q} \left[ G(\tilde{x}_i^T \beta_*) \wedge (1 - G(\tilde{x}_i^T \beta_*)) \right] \geq \inf_{|\mu| < \ell} [G(\mu) \wedge (1 - G(\mu))] =: g(\ell).
\]

As a result, for any \( \epsilon > 0 \),

\[
\tilde{P} \left( \sup_{\beta \in \mathbb{R}^p} \hat{\gamma}(\beta) > 1 - \frac{1}{2} g(\ell) + \epsilon \right) \leq \tilde{P} \left( 1 - \frac{1}{2} \min_{1 \leq i \leq q} v_i > 1 - \frac{g(\ell)}{2} + \epsilon \right) \leq \tilde{P} \left( \min_{1 \leq i \leq q} v_i < \min_{1 \leq i \leq q} \left[ G(\tilde{x}_i^T \beta_*) \wedge (1 - G(\tilde{x}_i^T \beta_*)) \right] - 2\epsilon \right) \to 0.
\]

Thus, the conclusion holds with \( \rho = 1 - \frac{g(\ell)}{2} \).

**Acknowledgment.** The second author was supported by NSF Grant DMS-15-11945.

## Appendix

### A Drift and Minorization for the Autoregressive Process

In this section we study the performance of the TV bounds given in Roberts and Tweedie (1999) and Baxendale (2005) in high dimensional settings using the toy example in Section [2]. Both these bounds are constructed via drift and minorization, similar to Rosenthal’s (1995) bound in Proposition [1].

The first bound we consider is from Section 5 of Roberts and Tweedie (1999). The following is a simplified version.
Proposition 28. (Roberts and Tweedie, 1999) Suppose that for some measurable \( V : X \to [1, \infty) \), there exist \( \lambda < 1 \), \( b < \infty \), and \( d \geq b/[2(1 - \lambda)] - 1 \) such that

\[
\int_X V(x')K(x, x') \, dx' \leq \lambda V(x) + b 1_C(x)
\]  

(8)

for all \( x \in X \), where \( C = \{ x \in X : V(x) \leq d \} \). Suppose further that there exist \( \gamma < 1 \) and a probability measure \( \nu : B \to [0, 1] \) such that for all \( x \in C \),

\[
K(x, \cdot) \geq (1 - \gamma)\nu(\cdot).
\]

Then \( \rho^TV \leq \hat{\rho}_{RT} \), where \( \hat{\rho}_{RT} \) is some number between \( \gamma \) and 1 (which we do not specify for the sake of simplicity).

We now show that \( \hat{\rho}_{RT} \) is not stable for the autoregressive process when the dimension is high. More specifically, we have the following.

Proposition 29. Let \( K(x, \cdot), x \in \mathbb{R}^p \), be the probability measure associated with the \( N(x/2, 3I_p/4) \) distribution. For any sequence of d&m conditions, \( \hat{\rho}_{RT} \to 1 \) at an exponential rate as \( p \to \infty \).

Proof. Let \( \lambda, b, d, C, \) and \( \gamma \) be as in Proposition 28. It suffices to show that \( \gamma \to 1 \) at an exponential rate as \( p \to \infty \).

We begin by showing a result that is analogous to Lemma 3. By Proposition 4.24 in Hairer (2006), \( \Pi V < \infty \). Then the drift condition (8) implies that

\[
b \Pi(C) \geq (1 - \lambda)\Pi V \geq 1 - \lambda.
\]  

(9)

Suppose that \( b \leq 4(1 - \lambda) \), then (9) yields \( \Pi(C) \geq 1/4 \). Suppose now that \( b > 4(1 - \lambda) \). Then \( d > 1 \). It’s easy to verify that \( \Pi V \geq \Pi(C) + d(1 - \Pi(C)) \). Combining this with (9) yields

\[
b \Pi(C) \geq (1 - \lambda)[\Pi(C) + d(1 - \Pi(C))]\]

It follows that

\[
\Pi(C) \geq \frac{d}{b/(1 - \lambda) + d - 1}.
\]

Since \( d \geq b/[2(1 - \lambda)] - 1 \), and \( d > 1 \),

\[
\Pi(C) \geq \frac{d}{3d + 1} \geq \frac{1}{4}.
\]

Following the argument in the proof of Proposition 4, we see that

\[
\gamma \geq 1 - 2\Phi \left(-\sqrt{\frac{m_p'}{3}}\right),
\]

27
where \( m'_p \) is the \( 1/4 \)th quantile of a \( \chi^2_p \) distribution. Since \( \chi^2_p/p \to 1 \) in distribution, \( m'_p/p \to 1 \). Therefore, \( \gamma \to 1 \) exponentially fast as \( p \to \infty \), and the result follows.

Remark 30. It seems to us that there is a typo in Roberts and Tweedie (1999). To be specific, in Proposition 28, \( d \) should be greater than \( b/(1 - \lambda) - 1 \), rather than just \( b/[2(1 - \lambda)] - 1 \). If we assume this, then \( \Pi(C) \geq 1/2 \), and Proposition 29 can be proved in exactly the same way as Proposition 4.

We say that a Markov chain with \( Mtf(K(\cdot, \cdot)) \) is positive if for any complex function \( g \) square integrable with respect to \( \Pi \),

\[
\int_{\mathbb{R}^p} \left[ \int_{\mathbb{R}^p} g(x') K(x, dx') \right] \overline{g(x)} \Pi(dx) \\
\propto \int_{\mathbb{R}^p} \left[ \int_{\mathbb{R}^p} g(x') \exp \left( -\frac{2}{3} \left\| x' - \frac{x}{2} \right\|^2 \right) dx' \right] \overline{g(x)} \exp \left( -\frac{\|x\|^2}{2} \right) dx \\
\propto \int_{\mathbb{R}^p} \left[ \int_{\mathbb{R}^p} g(x') \exp \left( -\left\| x' - \frac{y}{2} \right\|^2 \right) dx' \right] \left[ \int_{\mathbb{R}^p} \overline{g(x)} \exp \left( -\| x - \frac{y}{2} \|^2 \right) dx \right] \exp \left( -\frac{\|y\|^2}{4} \right) dy \\
\geq 0 .
\]

The second result we investigate is due to Baxendale (2005), which goes as follows.

Proposition 31. (Baxendale, 2005) Let \( K(\cdot, \cdot) \) define a reversible and positive Markov chain. Suppose that there exist measurable \( V : X \to [1, \infty) \), \( \lambda < 1 \), \( A < \infty \), and \( C \in \mathcal{B} \) such that

\[
\int_X V(x') K(x, dx') \leq \begin{cases} 
\lambda V(x) , & x \in X \setminus C , \\
A , & x \in C .
\end{cases}
\]

Suppose further that there exist \( \gamma < 1 \), and a probability measure \( \nu : \mathcal{B} \to [0, 1] \) such that

\[
K(x, \cdot) \geq (1 - \gamma) \nu(\cdot)
\]

for all \( x \in C \). Then \( \rho^{TV}_x \leq \hat{\rho}_{\text{Bax}} \), where

\[
\hat{\rho}_{\text{Bax}} = \begin{cases} 
\max \{ \lambda, \gamma^{1/\alpha} \} , & \gamma > 0 , \\
\lambda , & \gamma = 0 ,
\end{cases}
\]

and

\[
\alpha = 1 + (\log \lambda^{-1})^{-1} \left( \log \frac{A - (1 - \gamma)}{\gamma} \right) .
\]
Note that in the above proposition, $V \geq 1$. Thus, $\lambda > 0$, $A \geq 1$, which implies that $\hat{\rho}_{\text{Bax}}$ is always well-defined. Contrary to the approach of Rosenthal (1995) and Roberts and Tweedie (1999), Baxendale’s (2005) method does not impose a constant positive lower bound on $\Pi(C)$. However, we still have the following result.

**Proposition 32.** Let $K(x, \cdot)$, $x \in \mathbb{R}^p$, be the probability measure associated with the $N(x/2, 3I_p/4)$ distribution. For any sequence of d&m conditions, $\hat{\rho}_{\text{Bax}} \to 1$ as $p \to \infty$.

To prove this result, we make use of the following lemma.

**Lemma 33.** Let $\lambda$, $A$, and $C$ be as in Proposition 31, then $\Pi(C) \geq \log \frac{\lambda^{-1}}{\log \lambda^{-1} + \log A} > 0$.

**Proof.** Recall that $A \geq 1$ and $\lambda^{-1} \in (1, \infty)$. Thus, $\log \frac{\lambda^{-1}}{\log \lambda^{-1} + \log A}$ is well-defined and positive.

Let $\{X_m\}_{m=0}^{\infty}$ be a chain generated by $K(\cdot, \cdot)$ such that $X_0 \in C$ is fixed. For an integer $m \geq 1$, let $N_m = \sum_{i=0}^{m-1} 1_C(X_i)$. By arguments similar to the proofs of Lemma 3 and 4 in Rosenthal (1995), for any integer $i \geq 1$,

$$P(N_m < i) \leq \lambda^m (\lambda^{-1} A)^{i-1}.$$

Let $i_0 = \lfloor m \log \frac{\lambda^{-1}}{\log \lambda^{-1} + \log A} \rfloor$, where $\lfloor x \rfloor$, $x \in \mathbb{R}$, returns the largest integer that does not exceed $x$. Then $(\lambda^{-1} A)^{i_0} \leq \lambda^{-m}$. It follows that

$$\mathbb{E}N_m \geq \sum_{i=1}^{i_0} P(N_m \geq i) \geq i_0 - \lambda^m (\lambda^{-1} A)^{i_0} - 1 \geq i_0 - \frac{1 - \lambda^m}{\lambda^{-1} A - 1}. \quad (10)$$

Now, by ergodicity, $N_m/m \to \Pi(C)$ as $m \to \infty$, almost surely. By dominated convergence, $\mathbb{E}N_m/m \to \Pi(C)$ as $m \to \infty$. Dividing the left and right sides of (10) by $m$ and letting $m \to \infty$ shows that

$$\Pi(C) \geq \frac{\log \lambda^{-1}}{\log \lambda^{-1} + \log A}.$$

**Proof of Proposition 32.** Let $\lambda$, $A$, $C$, $\gamma$, and $\alpha$ be as in Proposition 31. It’s easy to see that $\gamma = 0$ only if $C$ is of Lebesgue measure 0. This contradicts Lemma 33. Thus, $\gamma > 0$, and $\hat{\rho}_{\text{Bax}} \geq \gamma^{1/\alpha}$.

Since $A \geq 1$, $\alpha \geq (\log \lambda^{-1} + \log A)/\log \lambda^{-1}$. By Lemma 33, $\hat{\rho}_{\text{Bax}} \geq \gamma^{\Pi(C)}$. Now, let $2r(C)$ be the diameter of $C$. By the same reasoning as in the proof of Proposition 29,

$$\gamma \geq 1 - 2\Phi \left( -\sqrt{\frac{1}{3}} r(C) \right). \quad (11)$$

29
It follows that as \( p \) grows, any subsequence of \( \gamma^{\Pi(C)} \) tends to 1 if in that subsequence, \( r(C) \to \infty \). Hence, to prove the result, it suffices to show that for any subsequences such that \( r(C) \) is bounded, \( \gamma^{\Pi(C)} \) still converges to 1.

For the remainder of this proof, assume that \( r(C) \) is bounded. It’s easy to verify that \( \Pi(C) \leq \Pi(\{x : x^T x \leq r(C)^2\}) = \int_0^{r(C)^2} f_p(s) \, ds \),

(12)

where

\[
 f_p(s) = \frac{1}{2^{p/2} \Gamma(p/2)} s^{p/2} e^{-s/2}, \quad s > 0,
\]

is the pdf of the \( \chi^2_p \) distribution. It’s not difficult to see that \( \Pi(C) \to 0 \) as \( p \to \infty \). Therefore, \( \gamma^{\Pi(C)} \to 1 \) unless for some subsequence, \( \gamma \to 0 \), which can happen only if for that subsequence, \( r(C) \to 0 \).

Assume for the rest of the proof that \( r(C) \to 0 \) as \( p \to \infty \). We only have to show that \( \gamma^{\Pi(C)} \to 1 \) in this case. Note that

\[
 \int_0^{r(C)^2} f_p(s) \, ds \leq \frac{1}{(p/2 + 1)2^{p/2} \Gamma(p/2)} r(C)^{p+2},
\]

while

\[
 1 - 2\Phi \left(-\sqrt{\frac{1}{3}r(C)}\right) \sim \sqrt{\frac{2}{3\pi}} r(C).
\]

It follows from (11) and (12) that \( \Pi(C) \) goes to 0 at a faster rate than \( \gamma \). Hence, \( \gamma^{\Pi(C)} \to 1 \).

\[\square\]

B Proofs

B.1 Proposition 7

Proof. Fix \( x, y \in X \). Let \( \mathcal{F}_m \) be the \( \sigma \)-algebra generated by \( \{(F_i(x), F_i(y))\}_{i=0}^m \), and let \( \{\psi_m\}_{m=0}^\infty = \{\psi(F_m(x), F_m(y))\}_{m=0}^\infty \). Let

\[
 t_1 = \inf_{m \geq 0} \{(F_m(x), F_m(y)) \in C \times C\},
\]

and for \( i > 1 \), let

\[
 t_i = \inf_{m \geq t_{i-1}+1} \{(F_m(x), F_m(y)) \in C \times C\}.
\]

By Lemma 3 and 4 in Rosenthal (1995), \( t_i \) is finite almost surely for all \( i \) under \((A1)\). It follows from condition \((A2')\) and \((A3)\) that for every \( i > 1 \),

\[
 \mathbb{E}(\psi_{t_{i+1}}|\mathcal{F}_0) = \sum_{j=0}^{\infty} \mathbb{E}[1_{t_i=j}\mathbb{E}(\psi_{j+1}|\mathcal{F}_j)|\mathcal{F}_0] \leq \gamma \mathbb{E}(\psi_{t_i}|\mathcal{F}_0) \leq \gamma \mathbb{E}(\psi_{t_{i-1}+1}|\mathcal{F}_0).
\]

30
Similarly, \( \mathbb{E}(\psi_{t+1}|\mathcal{F}_0) \leq \gamma \psi_0 \). Hence,

\[
\mathbb{E}(\psi_{t+1}|\mathcal{F}_0) \leq \psi_0 \gamma^t .
\]  

(13)

Now, for any \( m \geq 0 \) and \( i \geq 1 \), repeated uses of condition (A3) yield

\[
\mathbb{E}(\psi_m|\mathcal{F}_0) \leq \mathbb{E}(\psi_{t+1}\mathbb{1}_{[t_i< m]} + \psi_1\mathbb{1}_{[t_i \geq m]}|\mathcal{F}_0) \leq \mathbb{E}(\psi_{t+1}|\mathcal{F}_0) + \psi_0 \mathbb{P}(t_i \geq m|\mathcal{F}_0) .
\]

Thus, by (13),

\[
\mathbb{E}(\psi(F_m(x), F_m(y))|\mathcal{F}_0) \leq \psi(x, y) \gamma^i + \psi(x, y) \mathbb{P}(t_i \geq m|\mathcal{F}_0) .
\]  

(14)

By Lemma 3 and 4 in Rosenthal (1995),

\[
\mathbb{P}(t_i \geq m|\mathcal{F}_0) \leq (V(x) + V(y) + 1) \left( \frac{1 + 2L + \lambda d}{1 + \lambda d} \right)^{m-i+1} [1 + 2(\lambda d + L)]^{i-1} .
\]

If follows from (14) that

\[
\mathbb{E}(\psi(F_m(x), F_m(y))|\mathcal{F}_0)
\]

\[
\leq \psi(x, y) \gamma^i + \psi(x, y)(V(x) + V(y) + 1) \left( \frac{1 + 2L + \lambda d}{1 + \lambda d} \right)^{m-i+1} [1 + 2(\lambda d + L)]^{i-1} .
\]

Let \( a \in (0, 1) \). Taking \( i \) to be an integer such that \( i - 1 \leq an \leq i \) yields

\[
\mathbb{E}(\psi(F_m(x), F_m(y))|\mathcal{F}_0)
\]

\[
\leq \psi(x, y) \gamma^m + \psi(x, y)(V(x) + V(y) + 1) \left\{ \frac{1 + 2L + \lambda d}{1 + \lambda d} \right\}^{1-a} [1 + 2(\lambda d + L)]^a .
\]

Suppose now that \( y \) is random, and \( y \sim \Pi \), independently of \( F_m \), then

\[
W_\psi(K^m_x, \Pi) \leq \mathbb{E}(\psi(F_m(x), F_m(y))|\mathcal{F}_0).
\]

where the expectation is taken over \( F_m \) and \( y \). The result then follows. \( \square \)

**B.2 Lemma [10]**

**Proof.** Let \( x, y \in C \), and assume that \( x \neq y \). Fix \( f \) so that \( h(t) := \hat{D}_{x+t(y-x)} f , t \in [0, 1] \), is Riemann integrable. Let \( \epsilon > 0 \) be arbitrary. Then there exists \( \delta > 0 \) such that for any partition of \( [0, 1] \) whose norm is less than \( \delta \), a Riemann sum of \( h \) with respect to this partition is within \( \epsilon \) of \( \int_0^1 h(t) \, dt \).

By definition, for every \( t \in [0, 1] \), there exists \( \delta(t) > 0 \) such that

\[
\| f(x + t'(y-x)) - f(x + t(y-x)) \| \leq (D_{x+t(y-x)} f + \epsilon)|t' - t||y - x|
\]

\[
\leq (h(t) + \epsilon)|t' - t||y - x|
\]

31
whenever \( t' \in (t - \delta(t), t + \delta(t)) \). Without loss of generality, we can assume that \( \delta(t) \leq \delta/2 \) for all \( t \in [0, 1] \).

By the compactness of \([0, 1] \), there exists a sequence of points \( 0 = t_0 < t_1 < \cdots < t_{m+1} = 1 \), such that \([0, 1] \subset \bigcup_{i=0}^{m+1} (t_i - \delta(t_i), t_i + \delta(t_i)) \). As a result, there exists a finite partition of \([0, 1] \), denoted by

\[
0 = t_0 < s_0 < t_1 < s_1 < \cdots < t_m < s_m < t_{m+1} = 1 ,
\]
such that \( t_{i+1} - s_i < \delta(t_{i+1}), s_i - t_i < \delta(t_i) \) for \( i = 0, 1, \ldots, m \). Let \( s_{-1} = 0 \) and \( s_{m+1} = 1 \). Since \( \delta(t) \leq \delta/2 \) for all \( t, s_i - s_{i-1} < \delta \) for \( i \geq 0 \). Thus,

\[
\begin{align*}
\|f(y) - f(x)\| & \leq \sum_{i=0}^{m} (\|f(x + s_i(y - x)) - f(x + t_i(y - x))\| + \|f(x + t_{i+1}(y - x)) - f(x + s_i(y - x))\|) \\
& \leq \sum_{i=0}^{m+1} (h(t_i) + \epsilon) (s_i - s_{i-1})\|y - x\| \\
& \leq \left( \int_{0}^{1} h(t) \, dt + 2\epsilon \right) \|y - x\| .
\end{align*}
\]

Since \( \epsilon \) is arbitrary, \( \|f(y) - f(x)\| \leq \int_{0}^{1} h(t) \, dt \|y - x\| \). Now, let \( f \) be random, and take expectations on both sides. This yields

\[
\mathbb{E}\|f(y) - f(x)\| \leq \int_{0}^{1} \mathbb{E}\tilde{D}_{x+t(y-x)} f \, dt \|y - x\| \leq \sup_{x' \in C} \mathbb{E}\tilde{D}_{x'} f \|y - x\| .
\]

\[\square\]

### B.3 Proposition [14]

We prove the result using Proposition [13].

**Proof.** Recall that the density function for \( K(\beta, \cdot) \), \( \beta \in \mathbb{R}^p \) is given by

\[
k(\beta, \beta') = \int_{\mathbb{R}_+^p} \pi_1(\beta' \mid z, X, y) \pi_2(z \mid \beta, X, y) \, dz , \beta' \in \mathbb{R}^p ,
\]

where \( \pi_1 \) and \( \pi_2 \) are introduced in Subsection [4.1]. For \( \beta \in \mathbb{R}^p \), let

\[
Z(\beta, U) = \left( H(U_1, x_1^T \beta, y_1) \ H(U_2, x_2^T \beta, y_2) \ \cdots \ H(U_n, x_n^T \beta, y_n) \right)^T ,
\]

where \( U \) and \( H \) are defined in Subsection [4.1]. Then \( Z(\beta, U) \sim \pi_2(\cdot \mid \beta, X, y) \), and

\[
k(\beta, \beta') = \mathbb{E}\pi_1(\beta' \mid Z(\beta, U), X, y) , \beta, \beta' \in \mathbb{R}^p .
\]

32
As a result, for $\alpha, \beta \in \mathbb{R}^p$,
\[
\int_{\mathbb{R}^p} |k(\beta, \beta') - k(\alpha, \beta')| \, d\beta' \\
= \int_{\mathbb{R}^p} |\mathbb{E}\pi_1(\beta' | Z(\beta, U), X, y) - \mathbb{E}\pi_1(\beta' | Z(\alpha, U), X, y)| \, d\beta' \\
\leq \mathbb{E} \int_{\mathbb{R}^p} |\pi_1(\beta' | Z(\beta, U), X, y) - \pi_1(\beta' | Z(\alpha, U), X, y)| \, d\beta' .
\tag{15}
\]

Now, for $z \in \mathbb{R}^n$, $\pi_1(\cdot | z, X, y)$ is the pdf of $N(\Sigma^{-1} (X^T z + Qv), \Sigma^{-1})$. One can then verify that the right hand side of (15) is equal to
\[
2 - 4\mathbb{E}\Phi \left( -\frac{1}{2} \|\Sigma^{-1/2} X^T [Z(\beta, U) - Z(\alpha, U)]\|_2 \right) \leq \frac{2}{\sqrt{2\pi}} \mathbb{E}\|\Sigma^{-1} X^T [Z(\beta, U) - Z(\alpha, U)]\| .
\]

By definition of the random mapping $f$ defined in Subsection 4.1,
\[
\Sigma^{-1} X^T [Z(\beta, U) - Z(\alpha, U)] = f(\beta) - f(\alpha) .
\]

By Lemma 9 and 17 $\|f(\beta) - f(\alpha)\| \leq \|\beta - \alpha\|$. Thus,
\[
\int_{\mathbb{R}^p} |k(\beta, \beta') - k(\alpha, \beta')| \, d\beta' \leq \frac{2}{\sqrt{2\pi}} \|\beta - \alpha\| ,
\]
and the result follows from Proposition 13. $$\square$$

B.4 Lemma 16

Proof. We first prove that $s(u, \mu) \in (0, 1)$, where $u \in (0, 1)$ and $\mu \in \mathbb{R}$. It suffices to show that
\[
\frac{u\phi(\mu)}{\phi[\Phi^{-1}(\Phi(\mu)u)]} < 1 . \tag{16}
\]

To this end, let $x = H(1 - u, \mu, 1) > 0$, that is,
\[
u = \frac{1 - \Phi(x - \mu)}{\Phi(\mu)} .
\]

Then (16) is equivalent to
\[
\log \phi(\mu) - \log \Phi(\mu) < \log \phi(\mu - x) - \log \Phi(\mu - x) .
\]

This inequality holds if for $x' > 0$,
\[
\frac{d}{dx'} \left[ \log \phi(\mu - x') - \log \Phi(\mu - x') \right] > 0 ,
\]
which is equivalent to
\[
\mu - x' + \frac{\phi(\mu - x')}{\Phi(\mu - x')} > 0 .
\]
By a well-known result on Mill’s ratio \citep{Gordon1941},
\[
x' + \frac{\phi(x')}{\Phi(x')} > 0
\]
for each \(x' \in \mathbb{R}\). Hence, \((\ref{eq:Phi})\) holds and the result follows.

Onto the second assertion. Let \(\mu < 0\) and \(u \in (0, 1)\). It suffices to show that
\[
\frac{\phi(\mu)}{\phi(\Phi^{-1}(\Phi(\mu)u))} \geq 1.
\]
It’s clear that \(\Phi^{-1}(\Phi(\mu)u) \leq \mu\). Then \((\ref{eq:Phi2})\) holds because \(\phi(\cdot)\) is an increasing function on \((-\infty, 0)\). \(\square\)

### B.5 Proposition 18

To prove this result we invoke the matrix Hoeffding inequalities from \cite{Mackey2014}, which is stated in the following lemma.

**Lemma 34.** \cite{Mackey2014} \(\mathbb{H}^p\) be the set of \(p \times p\) complex Hermitian matrices. Let \(m\) be a positive integer. Consider a sequence of independent random matrices in \(\mathbb{H}^p\), \(\{M_i\}_{i=1}^m\), and a sequence of deterministic matrices in \(\mathbb{H}^p\), \(\{A_i\}_{i=1}^m\). Suppose that \(\mathbb{E}M_i = 0\) and \(M_i^2 \preceq A_i^2\) almost surely for each \(i\). Then
\[
\mathbb{E}\lambda_{\max}(\sum_{i=1}^m M_i) \leq \sigma' \sqrt{2 \log p},
\]
where \(\sigma' = \lambda_{\max}(\sum_{i=1}^m A_i^2)\).

**Proof of Proposition 18** For \(i = 1, 2, \ldots, n\), let
\[
\Delta_i(\beta, U) = \frac{\partial H(U, \mu, y_i)}{\partial y_i} \bigg|_{\mu = x_i^T \beta} - \mathbb{E}\frac{\partial H(U, \mu, y_i)}{\partial y_i} \bigg|_{\mu = x_i^T \beta}, \quad \beta \in \mathbb{R}^p.
\]
Then for \(\beta \in \mathbb{R}^p\),
\[
\Sigma^{-1/2}X^TS(\beta, U)X\Sigma^{-1/2} - \Sigma^{-1/2}X^T\mathbb{E}S(\beta, U)X\Sigma^{-1/2} = \sum_{i=1}^n \Sigma^{-1/2}x_i \Delta_i(\beta, U) x_i^T \Sigma^{-1/2}.
\]  \(\text{(18)}\)

For each \(i\), define \(M_i\) to be
\[
\Sigma^{-1/2}x_i \Delta_i(\beta, U) x_i^T \Sigma^{-1/2},
\]
and let \(A_i = \Sigma^{-1/2}x_i x_i^T \Sigma^{-1/2}\). Then \(\mathbb{E}M_i = 0\) for all \(i\). We now show that \(M_i^2 \preceq A_i^2\) for \(i = 1, 2, \ldots, n\).

We know from Lemma \(16\) that \(\Delta_i(\beta, U) \in [-1, 1]\) for all \(i\) and \(\beta\), which implies that
\[
\left(\Sigma^{-1/2}x_i \Delta_i(\beta, U) x_i^T \Sigma^{-1/2}\right)^2 = (x_i^T \Sigma^{-1}x_i) \left(\Sigma^{-1/2}x_i \Delta_i^2(\beta, U) x_i^T \Sigma^{-1/2}\right) \preceq (x_i^T \Sigma^{-1}x_i) \left(\Sigma^{-1/2}x_i x_i^T \Sigma^{-1/2}\right)
\]
\[
= \left(\Sigma^{-1/2}x_i x_i^T \Sigma^{-1/2}\right)^2.
\]
Then by Lemma 34,

\[
\mathbb{E}\lambda_{\max}\left(\sum_{i=1}^{n} \Delta_i(\beta, U) x_i^T \Sigma^{-1/2}\right) \leq \sigma \sqrt{2 \log p}.
\]

The result then follows immediately from (18) and Weyl’s inequality.

B.6 A lemma related to Proposition 24

Lemma 35. For \( \delta \in [0, 1] \) and \( \alpha \in \mathbb{R}^p \setminus \{0\} \), let \( J(\alpha, \delta) = \{ \beta \in \mathbb{R}^p : \alpha^T \beta \geq \delta \|\alpha\|_2 \|\beta\|_2 \} \). Suppose that \( \alpha \in J(\alpha_0, \sqrt{1 - \delta^2}) \) for some \( \alpha_0 \) and \( \delta \), and that \( \beta \in J(\alpha_0, \delta) \). Then \( \alpha^T \beta \geq 0 \).

Proof. For two vectors \( \alpha_1 \) and \( \alpha_2 \) in \( \mathbb{R}^p \), let \( \alpha_1^T \alpha_2 \) be their inner-product. Let \( P_0 \) be the orthogonal projection onto the subspace spanned by \( \alpha_0 \). Then \( P_0 \alpha = (\alpha_0^T \alpha / \|\alpha_0\|_2^2) \alpha_0 \), and \( P_0 \beta = (\alpha_0^T \beta / \|\alpha_0\|_2^2) \alpha_0 \). Note that

\[
\|P_0 \alpha\|_2 = \frac{\alpha_0^T \alpha}{\|\alpha_0\|_2^2} \geq \sqrt{1 - \delta^2} \|\alpha\|_2,
\]

\[
\|P_0 \beta\|_2 = \frac{\alpha_0^T \beta}{\|\alpha_0\|_2} \geq \delta \|\beta\|_2.
\]

As a result,

\[
\alpha^T \beta = (P_0 \alpha)^T (P_0 \beta) + [(1 - P_0) \alpha]^T [(1 - P_0) \beta] \\
\geq \frac{(\alpha_0^T \alpha)(\alpha_0^T \beta)}{\|\alpha_0\|_2^2} - \|\alpha_0\|_2^2 \|1 - P_0\|_2 \|1 - P_0\|_2 \\
\geq \delta \sqrt{1 - \delta^2} \|\alpha\|_2 \|\beta\|_2 - \sqrt{\|\alpha\|_2^2 - \|P_0 \alpha\|_2^2} \sqrt{\|\beta\|_2^2 - \|P_0 \beta\|_2^2} \\
\geq 0.
\]

\[\square\]

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.

BARAGATTI, M. and POMMERET, D. (2012). A study of variable selection using g-prior distribution with ridge parameter. Computational Statistics & Data Analysis 56 1920–1934.

BAXENDALE, P. H. (2005). Renewal theory and computable convergence rates for geometrically ergodic Markov chains. Annals of Applied Probability 15 700–738.
BUTKOVSKY, O. (2014). Subgeometric rates of convergence of Markov processes in the Wasserstein metric. *Annals of Applied Probability* **24** 526–552.

CHAKRABORTY, S. and KHARE, K. (2017). Convergence properties of Gibbs samplers for Bayesian probit regression with proper priors. *Electronic Journal of Statistics* **11** 177–210.

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.

DURMUS, A. and MOULINES, É. (2015). Quantitative bounds of convergence for geometrically ergodic Markov chain in the Wasserstein distance with application to the Metropolis adjusted Langevin algorithm. *Statistics and Computing* **25** 5–19.

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

GORDON, R. D. (1941). Values of Mills’ ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *Annals of Mathematical Statistics* **12** 364–366.

GUPTA, M. and IBRAHIM, J. G. (2007). Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *Journal of the American Statistical Association* **102** 867–880.

HAIRER, M. (2006). Ergodic properties of Markov processes. Lecture notes.

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*. Springer.

HAIRER, M., MATTINGLY, J. C. and SCHUETZOW, M. (2011). Asymptotic coupling and a general form of Harris theorem with applications to stochastic delay equations. *Probability Theory and Related Fields* **149** 223–259.

JARNER, S. and TWEEDIE, R. (2001). Locally contracting iterated functions and stability of Markov chains. *Journal of Applied Probability* **38** 494–507.

JOHNDROW, J. E., SMITH, A., PILLAI, N. and DUNSON, D. B. (2018). MCMC for imbalanced categorical data. *Journal of the American Statistical Association*, to appear.

MACKEY, L., JORDAN, M. I., CHEN, R. Y., FARRELL, B. and TROPP, J. A. (2014). Matrix concentration inequalities via the method of exchangeable pairs. *Annals of Probability* **42** 906–945.
MADRAS, N. and SEZER, D. (2010). Quantitative bounds for Markov chain convergence: Wasserstein and total variation distances. *Bernoulli* **16** 882–908.

MANGOUBI, O. and SMITH, A. (2017). Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. *arXiv:1708.07114*.

MEYN, S. P. and TWEEDIE, R. L. (2012). *Markov Chains and Stochastic Stability*. 2nd ed. Springer Science & Business Media.

OLLIVIER, Y. (2009). Ricci curvature of Markov chains on metric spaces. *Journal of Functional Analysis* **256** 810–864.

QIN, Q. and HOBERT, J. P. (2019+). Convergence complexity analysis of Albert and Chib’s algorithm for Bayesian probit regression. *Annals of Statistics*, to appear.

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 TWEEDIE, R. L. (1999). Bounds on regeneration times and convergence rates for Markov chains. *Stochastic Processes and their Applications* **80** 211–229.

ROSENTHAL, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association* **90** 558–566.

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.

STEINSALTZ, D. (1999). Locally contractive iterated function systems. *Annals of Probability* 1952–1979.

TRILLOS, N. G., KAPLAN, Z., SAMAKHOANA, T. and SANZ-ALONSO, D. (2017). On the consistency of graph-based Bayesian learning and the scalability of sampling algorithms. *arXiv:1710.07702*.

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

YANG, J. and ROSENTHAL, J. S. (2017). Complexity results for MCMC derived from quantitative bounds. *arXiv:1708.00829*.
YANG, Y., WAINWRIGHT, M. J. and JORDAN, M. I. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *Annals of Statistics* **44** 2497–2532.