Some Convergence Results for Metropolis-Hastings Algorithms

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Austin Brown

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Galin L. Jones, Adviser

August 2022

Acknowledgements

I would like to thank my advisor, Galin Jones, for his guidance in writing this thesis. Thanks to Professor Charles Geyer and Professor Qian Qin for some interesting conversations on Markov chain Monte Carlo. Thanks to Professor Wei-Kuo Chen for helpful comments in improving this thesis. My mother, Georgia, has always shown me love and support and I am forever thankful. Finally, the many conversations with my grandfather, Arthur, have been an inspiration for me throughout my academic pursuit.

Abstract

This thesis is concerned with the computational effort required by a Metropolis-Hastings algorithm to converge to the target distribution in total variation and Wasserstein distances. First, under mild assumptions, we show the sharp convergence rate in total variation is also sharp in weaker Wasserstein distances for the Metropolis-Hastings independence sampler on $\mathbb{R}^d$. We derive exact convergence expressions for general Wasserstein distances when initialization is at a specific point. Using optimization, we construct a novel centered independent proposal to develop exact convergence rates in Bayesian quantile regression and many generalized linear model settings. We show the exact convergence rate can be upper bounded in Bayesian binary response regression (e.g. logistic and probit) when the sample size and dimension grow together.

Next, practitioners are often left tuning Metropolis-Hastings algorithms by trial and error or using optimal scaling guidelines to avoid poor empirical performance. We develop general lower bounds on the convergence rates of geometrically ergodic Metropolis-Hastings algorithms on $\mathbb{R}^d$ to study their computational complexity. If the target density concentrates with a parameter n (e.g. Bayesian posterior concentration, Laplace approximations), we show the convergence rate can tend to 1 exponentially fast if the tuning parameters do not depend carefully on d and n.

Finally we look at a different topic and consider Bayesian error-in-variable (EIV) linear regression accounting for additional additive Gaussian error in the features (covariates) and response. We construct a 3-variable deterministic scan Gibbs sampler using independent normal and inverse-gamma priors. We prove this Gibbs sampler is geometrically ergodic which ensures a central limit theorem.

Contents

List of Figures **v**

1 Introduction **1**

1.1 Markov chain Monte Carlo algorithms 1

1.2 Metropolis-Hastings 3

1.3 Convergence of Metropolis-Hastings 4

1.4 Contributions 7

2 Exact Convergence Analysis for Metropolis-Hastings Independence Samplers in Wasserstein Distances **11**

2.1 Overview 11

2.2 MHI and Wasserstein distance 13

2.3 Exact convergence rates for MHI samplers 15

2.3.1 Application: Bayesian quantile regression 17

2.4 MHI with centered Gaussian proposals 19

2.4.1 Application: Bayesian generalized linear models 21

2.4.2 Convergence complexity analysis of MHI in binary response regression 23

2.5 Final remarks 24

3 Lower Bounds on the Rate of Convergence for Metropolis-Hastings

Algorithms **26**

3.1 Overview 26

3.2 Lower bounds for Metropolis-Hastings 29

3.2.1 Wasserstein lower bounds 34

3.3 Lower bounds for RWM with log-concave targets 40

3.3.1 Application: Bayesian logistic regression with Zellner's g-prior 43

3.4 Lower bounds under concentration 46

3.4.1 Application: generalized flat prior Bayesian logistic regression 52

3.5 Comparison with conductance methods 57

3.6 Discussion and further research directions 61

3.7 Supplementary material and code availability 61

4 Geometric Ergodicity of a Gibbs Sampler for Bayesian Error-in-Variable Regression **62**

4.1 Overview 62

4.2 General Gibbs Sampler for EIV regression 65

4.2.1 Bayesian EIV regression with errors in the features 74

4.2.2 Bayesian EIV regression with errors in the response and features 76

4.3 Conclusion and Future Directions 79

References **80**

A Exact Convergence Analysis for Metropolis-Hastings Independence Samplers in Wasserstein Distances **95**

A.1 Proof of Theorem 1 95

A.2 Proof of Theorem 2 102

A.3 Proof of Theorem 3 104

List of Figures

1.1 A visual diagram of the optimal transportation problem 6

2.1 Limiting decrease in the Wasserstein distance 25

3.1 Lower bounds on the RWM algorithm for logistic regression with Zellner'g g-prior 45

3.2 Lower bounds on the RWM algorithm for flat prior logistic regression 58

Chapter 1

Introduction

1.1 Markov chain Monte Carlo algorithms

Statistics and many areas of science including epidemiology, statistical physics, and neuroscience involve computing integrals with respect to some probability measure. For example, the main subjects of study in Bayesian statistics are posterior integrals. Let Π be a probability measure on $\mathbb{R}^d$ where d is a positive integer. The Monte Carlo method approximates integrals by generating samples $\theta_0, \ldots, \theta_T$ representative of Π so that for integrable, real-valued functions f, in some sense,

$$\frac{1}{T}\sum_{t=0}^{T-1} f(\theta_t) \approx \int_{\mathbb{R}^d} f d\Pi.$$

Independent sampling methods such as inverting the cumulative distribution function and rejection sampling [94] are often challenging or even infeasible to implement in modern high-dimensional problems. Markov chain Monte Carlo (MCMC) algorithms instead generate correlated samples from a Markov chain which can be readily simulated on a computer. Two widely utilized algorithms include Metropolis-Hastings [47, 75] and Gibbs sampling [38]. The goal is then to simulate an MCMC algorithm for sufficiently long so the simulated samples are representative of Π and can be used

in Monte Carlo approximations [57].

MCMC algorithms generate a discrete-time Markov chain $(\theta_t)_{t=0}^{\infty}$ initialized from a probability measure ν. We will be interested in time-homogeneous Markov chains where the probability distribution of $\theta_1|\theta$ is specified by a Markov transition kernel P defined for $\theta \in \mathbb{R}^d$ and Borel measurable sets $B \subseteq \mathbb{R}^d$ by $P(\theta, B)$. For positive integers $t \geq 2$ and defining $P^1 \equiv P$, the distribution of $\theta_t|\theta$ is defined recursively by the t-step Markov kernel

$$P^t(\theta, B) = \int_{\mathbb{R}^d} P^{t-1}(\cdot, B) dP(\theta, \cdot).$$

The marginal distribution of θ_t starting from any initial probability distribution ν on $\mathbb{R}^d$ is defined

$$\nu P^t = \int_{\mathbb{R}^d} P^t(\theta', \cdot) d\nu(\theta').$$

The Markov kernel P^t defines a linear operator on Borel measurable functions $f : \mathbb{R}^d \to \mathbb{R}$ by

$$P^t f(\theta) = \int_{\mathbb{R}^d} f dP^t(\theta, \cdot)$$

and its dual operation on probability measures by $P^{t*}\nu = \nu P^t$.

A Markov chain has invariant distribution Π if it preserves the marginal distribution of the Markov chain in the sense that $\Pi P^t = \Pi$. Markov chains such as Metropolis-Hastings satisfy a stronger *reversibility* condition. A Markov kernel P is *reversible* for Π if it satisfies the symmetric condition defined for Borel measurable sets $A, B \subseteq \mathbb{R}^d$ by

$$\int_A P(\cdot, B) d\Pi = \int_B P(\cdot, A) d\Pi.$$

Reversibility implies Π is invariant for the Markov kernel and a reversible Markov operator is a self-adjoint linear operator in $L^2(\Pi)$.

1.2 Metropolis-Hastings

One of the most widely used and important MCMC algorithms is the Metropolis-Hastings algorithm. In statistics especially, it is often the case the normalizing constant is unknown for Bayesian posteriors or the posterior conditionals cannot be sampled directly and Metropolis-Hastings requires neither. Estimating the normalizing constant for probability densities is a nontrivial problem required in methods such as importance sampling. Alternative MCMC algorithms such as Gibbs samplers require sampling the conditionals of the target distribution directly which can only be achieved in specific situations.

The original Metropolis-Rosenbluth algorithm was introduced using a symmetric proposal [75] and later extended to more general proposals which we introduce here [47, 113, 114]. Suppose Π has probability density π with respect to a Borel measure μ on $\mathbb{R}^d$. For each $\theta \in \mathbb{R}^d$, a proposal is a Markov kernel $Q(\theta, \cdot)$ on $\mathbb{R}^d$ with Markov transition density $q(\theta, \cdot)$ with respect to μ. The proposal is used to generate samples and these samples are accepted or rejected using an acceptance function $a(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ [113, 114]. When initialized with θ_0, for positive integers t, the Markov chain generates independently $\theta'_t|\theta_{t-1}$ from $Q(\theta_{t-1}, \cdot)$ and a standard uniform random variable U_t producing the random variable

$$\theta_t = \begin{cases} \theta'_t, & \text{if } U_t \leq a\left(\theta_{t-1}, \theta'_t\right) \\ \theta_{t-1}, & \text{if } U_t > a\left(\theta_{t-1}, \theta'_t\right) \end{cases}.$$

Denote the Dirac probability measure at the point $x \in \mathbb{R}^d$ by δ_x. The Metropolis-Hastings Markov kernel P is defined for $\theta \in \mathbb{R}^d$ and Borel measurable sets $B \subseteq \mathbb{R}^d$ by

$$P(\theta, B) = \int_B a(\theta, \theta') dQ(\theta, \theta') + \delta_\theta(B) \int_{\mathbb{R}^d} \left(1 - a(\theta, \theta')\right) dQ(\theta, \theta').$$

For Π to be invariant, the acceptance function must satisfy for Borel measurable sets $A, B \subseteq \mathbb{R}^d$,

$$\int_A \int_B a(\theta, \theta') dQ(\theta, \theta') d\Pi(\theta) = \int_B \int_A a(\theta', \theta) dQ(\theta', \theta) d\Pi(\theta') \tag{1.1}$$

and in this case, the Markov kernel is reversible for Π [114]. The standard Metropolis-Hastings acceptance function is defined for each $\theta, \theta' \in \mathbb{R}^d$ by

$$a(\theta, \theta') = \begin{cases} \min\left\{\frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')}, 1\right\}, & \text{if } \pi(\theta)q(\theta, \theta') > 0 \\ 1, & \text{if } \pi(\theta)q(\theta, \theta') = 0 \end{cases}.$$

Using this acceptance function, Metropolis-Hastings satisfies (1.1), and does not require the normalizing constant of the target density π. Acceptance functions are not unique [114] and alternatives exist such as Barker's acceptance [5]. However, the standard Metropolis-Hastings acceptance is closely related to rejection sampling and minimizes the asymptotic variance in the Markov chain central limit theorem among alternative choices [83, 114].

1.3 Convergence of Metropolis-Hastings

Under relatively weak conditions, generating samples from an MCMC algorithm such as Metropolis-Hastings will asymptotically draw samples from the target distribution [94]. A fundamental question in MCMC is determining the *finite* number of iterations required to generate a representative sample from the target distribution [58, 59, 102]. The topic of convergence analysis compares the marginal distribution of the Markov chain denoted by νP^t to the target distribution Π using a discrepancy measure. Some examples of discrepancy measures include the total variation and Wasserstein distances [122, 123]. A popular approach is to show geometric ergodicity

and provide an exponential convergence rate to the target distribution from any point of initialization in total variation [45, 59, 76, 102, 113]. Let $\mathcal{C}(\nu P^t, \Pi)$ be the set of all joint probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals distributions νP^t and Π respectively. The total variation distance between νP^t and Π is defined

$$\left\|\nu P^t - \Pi\right\|_{\mathrm{TV}} = \inf_{\xi \in \mathcal{C}(\nu P^t, \Pi)} \int_{\mathbb{R}^d \times \mathbb{R}^d} I_{\theta \neq \omega} d\xi(\theta, \omega).$$

A Metropolis-Hastings kernel P is *geometrically ergodic* [113] if there is a convergence rate $\rho \in (0, 1)$ and a function $\theta \mapsto M(\theta)$ where for every positive integer t and every point initialization $\theta \in \mathbb{R}^d$ with $\pi(\theta) > 0$, $\|P^t(\theta, \cdot) - \Pi\|_{\mathrm{TV}} \leq M(\theta)\rho^t$. Here we write the marginal distribution of the Markov chain $\delta_\theta P^t$ as the t-step kernel $P^t(\theta, \cdot)$.

Convergence analyses of general state space Markov chains have traditionally focused on studying their convergence rates in the total variation distance [76, 102, 113]. These convergence rates have received significant attention, at least in part, because they provide a key sufficient condition for the existence of central limit theorems [57] and the validity of methods for assessing the reliability of the simulation effort [100, 119]. However, convergence analysis of Metropolis-Hastings Markov chains is well-understood to be difficult and typically results in establishing a *qualitative* convergence rate [46, 53, 56, 74, 99]. Explicit convergence rates have been rare with the exception of some Metropolis-Hastings independence samplers [113]. For example, a widely known result gives general conditions for the random-walk algorithm to be geometrically ergodic [53], but the scaling properties of the convergence rate with the problem size (e.g. the dimension or posterior sample size) remain unknown.

There has been significant recent interest in the geometric convergence properties of Monte Carlo Markov chains on general state spaces in high-dimensional settings [27, 30, 46, 55, 88, 92, 130] and traditional approaches can have limitations in this regime [89]. This has led to an interest in considering the convergence rates of Monte

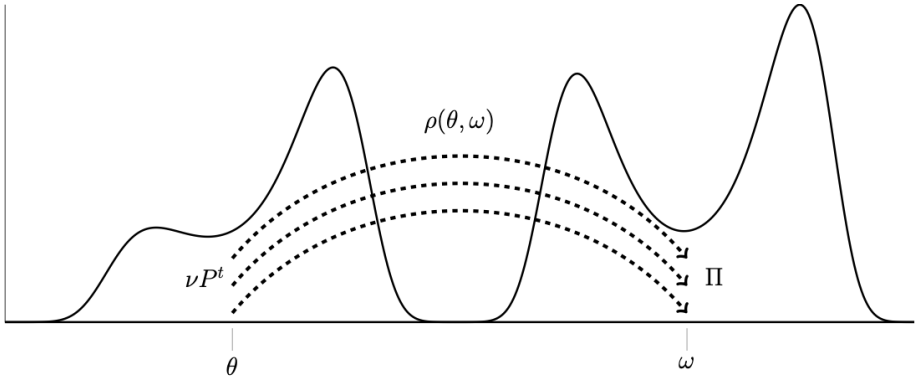


Figure 1.1: A diagram visually showing the optimal transport problem of moving point masses from the distribution νP^t to the distribution Π.

Carlo Markov chains using Wasserstein distances [39, 46, 54, 72, 90, 91] which may scale to large problem sizes where other approaches have had difficulties [26, 46, 91]. Convergence analyses in Wasserstein distances also result in benefits similar to those obtained using total variation such as central limit theorems and concentration inequalities for time averages of the Markov chain [46, 54, 60, 66].

The L^1-Wasserstein distance [61, 122, 123], which we will call the Wasserstein distance, is defined for two probability measures νP^t and Π as a measure of the optimal way to transport all of the mass from the probability measure to the other. A lower semicontinuous metric function ρ specifies the cost to transport each point mass. This intuitive definition is shown visually in Figure 1.1.

The Wasserstein distance is defined as the optimal average transport cost

$$\mathcal{W}_\rho\left(\nu P^t, \Pi\right) = \inf_{\xi \in \mathcal{C}(\nu P^t, \Pi)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \rho(\theta, \omega) d\xi(\theta, \omega).$$

The Wasserstein distance may be viewed as an infinite-dimensional linear program and admits a dual infinite-dimensional linear program [122]. This is expressed in the

famous Kantorovich-Rubinstein theorem ([122], Theorem 1.14) which states that

$$\mathcal{W}_\rho(\nu P^t, \Pi) = \sup_{\substack{\varphi \in M_b(\mathbb{R}^d) \\ \|\varphi\|_{\mathrm{Lip}(\rho)} \leq 1}} \left[\int_{\mathbb{R}^d} \varphi d\nu P^t - \int_{\mathbb{R}^d} \varphi d\Pi \right]$$

where $M_b(\mathbb{R}^d)$ is the set of bounded, Borel measurable, real-valued functions on $\mathbb{R}^d$ and

$$\mathrm{Lip}(\rho)(\varphi) = \sup_{x,y \in \mathbb{R}^d, x \neq y} |\varphi(y) - \varphi(x)| / \rho(x, y)$$

is the Lipschitz norm measuring the maximum Lipschitz constant for a real-valued function $\varphi : \mathbb{R}^d \to \mathbb{R}$. From a statistical viewpoint, this dual representation says the Wasserstein distance measures the worst-case bias between these two probability measures over these functions. When the metric ρ is the Hamming metric, then the Wasserstein distance is the total variation distance. A more interesting example is $\rho(\theta, \omega) = \min\{\|\theta - \omega\|, 1\}$ which is always less than the Hamming metric used in total variation yielding a *weaker* Wasserstein distance discrepancy measure.

1.4 Contributions

The main part of this thesis is concerned with the computational effort required by a Metropolis-Hastings algorithm to converge when the target distribution Π has Lebesgue density defined on $\mathbb{R}^d$. As mentioned in the previous section, the majority of convergence analysis for Metropolis-Hastings algorithms focuses on upper bounds to the convergence rates which are often non-explicit (qualitative) or too crude to be computed and employed in applications. In comparison, lower bounds on the convergence rate can provide practitioners with further insight into the *optimal* convergence behavior of the Metropolis-Hastings algorithm. Similar to establishing an explicit upper bound on a convergence rate, obtaining useful lower bounds, which

may also match the upper bound, is regarded as a challenging problem even for toy examples.

In Chapter 2, we consider the Metropolis-Hastings independence (MHI) sampler, which uses any proposal independent of previous iterations [113], for target distributions on $\mathbb{R}^d$. We look to the question: *can the geometric convergence rate be improved using Wasserstein distances alternative to total variation for the MHI sampler?* Under mild assumptions, we show the sharp convergence rate in total variation obtained using a global minorization condition [124] is also the sharp convergence rate for Wasserstein distances weaker than total variation. We derive exact convergence expressions in more general Wasserstein distances when initialization is at a specific point.

Due to modern high-dimensional statistical applications, there has been significant recent interest in the convergence complexity of MCMC algorithms with respect to the dimension d and sample size n [6, 12, 30, 55, 88, 92, 129, 130, 132]. We then study if the exact convergence rate obtained for the MHI sampler can yield informative convergence rates for practitioners which can scale to large problem sizes since the convergence analysis is exact. Using optimization, we construct a novel centered independent proposal to develop exact convergence rates in Bayesian quantile regression and many generalized linear model settings. We then consider scaling properties of the exact convergence rate to large dimensions and sample sizes in high-dimensional Bayesian binary response regression (e.g. logistic and probit regression) with Gaussian priors. We show the exact convergence rate can be upper bounded in Bayesian binary response regression (e.g. logistic and probit) when the sample size and dimension grow together.

In Chapter 3, we look to aid practitioners in answering the question: *how does one properly tune general geometrically ergodic Metropolis-Hastings algorithms to avoid poor empirical performance?* Unlike in Gibbs samplers, Metropolis-Hastings often

require choosing tuning parameters for the proposal which can be an arduous task in practice. For example, the widely used random-walk and Metropolis-adjusted Langevin proposals [98, 113] require carefully choosing a variance parameter. There are some qualitative results on the geometric ergodicity of some Metropolis-Hastings algorithms, but the computational complexity of the convergence and dependence on the tuning parameters are still largely unknown [53, 56, 98, 103, 119]. Consequently, practitioners are often left tuning these algorithms by trial and error or using optimal scaling guidelines [95, 96] to avoid poor empirical performance.

We develop general lower bounds on the convergence rates of geometrically ergodic Metropolis-Hastings algorithms for target distributions on $\mathbb{R}^d$ to study their computational complexity. General lower bounds on the mixing times are also developed when the algorithms are not necessarily geometrically ergodic and similar lower bounds are given in Wasserstein distances. If the target density concentrates with a parameter n (e.g. Bayesian posterior concentration, Laplace approximations), we show the convergence rate can tend to 1 exponentially fast if the tuning parameters do not depend carefully on d and n. Examples are applied to Bayesian logistic regression with Zellner's g-prior and flat prior Bayesian logistic regression.

Finally, in Chapter 4, we look at a different topic in the geometric ergodicity of Gibbs samplers. Many problems in epidemiology [1, 13, 14, 20] and astrophysics [48] among other areas of science [43, 86, 111] involve errors in the variables (EIV) which classical linear regression does not take into account. EIV can occur in many situations such as measurement error in data collection [48], a discrepancy between the data distribution and the model [13, 14], or purposeful adversarial attacks against the data [42, 110]. Not surprisingly, multiple critical issues arise in parameter estimation and statistical inference when ignoring additional errors in the data such as poor predictive performance [42], statistical bias [22, 67, 121], and estimators fail to be consistent [77].

We consider Bayesian error-in-variable (EIV) linear regression accounting for additional additive Gaussian error in the features (covariates) and response. We construct a 3-variable deterministic scan Gibbs sampler for Bayesian error-in-variables regression using independent normal and inverse-gamma priors. We prove this Gibbs sampler is geometrically ergodic which ensures a central limit theorem for many time averages from the Markov chain.

Chapter 2

Exact Convergence Analysis for Metropolis-Hastings Independence Samplers in Wasserstein Distances

2.1 Overview

In this chapter, we study exact convergence rates of Metropolis-Hastings independence sampler (MHI) Markov chains in Wasserstein distances alternative to total variation. The exact convergence behavior of Metropolis-Hastings algorithms across various Wasserstein distances has not been previously studied. Only in specific Wasserstein distances, have upper bounds been developed [26, 46]. Our main contribution develops exact convergence rates which are universal across Wasserstein distances for the MHI sampler. Surprisingly, we show the sharp convergence rate in total variation [74, 113, 124] is also sharp for Wasserstein distances weaker than total variation under mild assumptions. When the algorithm is started at a specific point, we give exact convergence expressions across more general Wasserstein distances. We also provide a practically relevant application of our theoretical results by developing exact convergence expressions using normal-inverse-gamma proposals in the Bayesian quantile regression setting.

Motivated by the general theoretical work, we consider using a centered Gaussian proposal and derive exact convergence expressions in Wasserstein distances for a large class of target distributions. The centered Gaussian proposal matches the maximal point of the proposal density with that of the target density. By centering an independent proposal, we directly imbue the Markov chain with a strong attraction to a set where the target distribution has high probability. This centered Gaussian proposal is similar to using a Laplace approximation [85, 106, 113], but differs in its covariance matrix. We study this MHI in several Bayesian generalized linear models and derive exact convergence expressions in general Wasserstein distances.

Our techniques are based on a condition [74, 113, 124] which is well-known but has previously been difficult to scale in high-dimensional settings. The novelty in our analysis is a carefully constructed proposal to develop exact convergence rates across Wasserstein distances. We then consider scaling properties of the exact convergence rate to large dimensions and sample sizes in high-dimensional Bayesian binary response regression (e.g. logistic and probit regression) with Gaussian priors. Data augmentation algorithms have been developed for these models [2, 87], but the required matrix inversions at each iteration can be computationally intensive. We derive an explicit asymptotic upper bound on the convergence rate of our MHI for general Wasserstein distances when the sample size and dimension increase in such a way that the ratio $d/n \to \gamma \in (0, +\infty)$. In this case, we show a well-known condition yields informative convergence rates for practitioners which can scale to large problem sizes when the convergence analysis is exact.

To the best of our knowledge, this work is the first to successfully address the convergence complexity of Metropolis-Hastings in general Wasserstein distances when both the sample size and the dimension increase. Previously under the conditions of a central limit theorem, the convergence complexity in total variation of random walk Metropolis (RWM) on a compact set was studied [6]. In contrast, our convergence

complexity results do not rely on the underlying space being compact. The dimension dependence of the mixing time has been studied in specific Wasserstein distances and total variation for Metropolis-Hastings algorithms such as Metropolis-adjusted Langevin (MALA) and RWM for certain log-concave target distributions [28, 29]. We take into account the sample size and upper bound the convergence rate which provides further theoretical guarantees for time averages of the Markov chain [57, 60]. Related results have investigated the convergence properties of some high-dimensional Gibbs samplers [81, 82] or studied the cases when the dimension or the sample size increase individually [30, 55, 91, 92].

The remainder is organized as follows. In Section 2.2, we define the Metropolis-Hastings independence sampler and the Wasserstein distance. In Section 2.3, we develop exact convergence rates in the Wasserstein distance for the MHI sampler and apply this theory to Bayesian quantile regression. In Section 2.4, we study a centered Gaussian proposal to obtain exact convergence expressions and apply this to many popular Bayesian generalized linear models used in statistics. We also develop high-dimensional convergence complexity results for Bayesian binary response regression in the large dimension and large sample size regime. Section 2.5 contains some final remarks. Some technical details and proofs are deferred to the appendices.

2.2 MHI and Wasserstein distance

As they will be considered here, MHI algorithms simulate a Markov chain with invariant distribution Π supported on a nonempty set $\Theta \subseteq \mathbb{R}^d$ using a proposal distribution Q which, to avoid trivialities, is assumed throughout to be different than Π. We also assume throughout that Π has Lebesgue density π with support Θ and Q has Lebesgue

density q with support Θ. Define

$$a(\theta, \theta') = \begin{cases} \min\left\{\frac{\pi(\theta')q(\theta)}{\pi(\theta)q(\theta')}, 1\right\}, & \text{if } \pi(\theta)q(\theta') > 0 \\ 1, & \text{if } \pi(\theta)q(\theta') = 0 \end{cases}.$$

We will consider MHI algorithms initialized at a point $\theta_0 \in \Theta$. MHI proceeds as follows: for $t \in \{1, 2, \ldots\}$, given θ_{t-1}, draw $\theta'_t \sim Q(\cdot)$ and $U_t \sim \text{Unif}(0, 1)$ independently so that

$$\theta_t = \begin{cases} \theta'_t, & \text{if } U_t \leq a\left(\theta_{t-1}, \theta'_t\right) \\ \theta_{t-1}, & \text{otherwise} \end{cases}.$$

If δ_θ denotes the Dirac measure at the point θ, the MHI Markov kernel P is defined for $\theta \in \mathbb{R}^d$ and $B \subseteq \mathbb{R}^d$ by

$$P(\theta, B) = \int_B a(\theta, \theta')q(\theta')d\theta' + \delta_\theta(B)\left(1 - \int a(\theta, \theta')q(\theta')d\theta'\right).$$

With $P \equiv P^1$, define the Markov kernel at iteration time $t \geq 2$ recursively by

$$P^t(\theta, B) = \int P(\theta, d\theta')P^{t-1}(\theta', B).$$

Let $\mathcal{C}(P^t(\theta, \cdot), \Pi)$ be the set of all joint probability measures with marginals $P^t(\theta, \cdot)$ and Π and ρ be a lower semicontinuous metric. The L_1-Wasserstein distance [61, 122, 123], which we will call simply the Wasserstein distance, is

$$\mathcal{W}_\rho\left(P^t(\theta, \cdot), \Pi\right) = \inf_{\xi \in \mathcal{C}(P^t(\theta, \cdot), \Pi)} \int \rho(\theta, \omega)d\xi(\theta, \omega).$$

2.3 Exact convergence rates for MHI samplers

When the ratio of the proposal and target densities is bounded below by a positive number, that is,

$$\epsilon^* = \inf_{\theta \in \Theta} \{q(\theta)/\pi(\theta)\} > 0,$$

the MHI sampler is uniformly ergodic in total variation with convergence rate upper bounded by $1 - \epsilon^*$ [113, Corollary 4]. Unlike in accept-reject sampling, ϵ^* does not need to be known explicitly or computed in order to implement MHI. However, this requirement was shown to be necessary for uniform ergodicity in total variation [74, Theorem 2.1]. More recently, it was shown the convergence rate cannot be improved [124, Theorem 1]. We show this is the case even in weaker Wasserstein distances.

Theorem 1. *Suppose* $\rho(\cdot, \cdot) \leq 1$. *Then*

$$\sup_{\theta \in \Theta} \mathcal{W}_\rho(P^t(\theta, \cdot), \Pi) \leq (1 - \epsilon^*)^t \sup_{\theta \in \Theta} \int \rho(\cdot, \theta) d\Pi.$$

If in addition, q *is lower semicontinuous and* π *is upper semicontinuous on* Θ*, and there is a sequence of compact sets increasing in diameter to* Θ*, then*

$$(1 - \epsilon^*)^t \inf_{\theta \in \Theta} \int \rho(\cdot, \theta) d\Pi \leq \sup_{\theta \in \Theta} \mathcal{W}_\rho(P^t(\theta, \cdot), \Pi) \leq (1 - \epsilon^*)^t \sup_{\theta \in \Theta} \int \rho(\cdot, \theta) d\Pi.$$

Proof. The proof is provided in Appendix Appendix A.1. □

The additional assumptions are not required when working with the total variation norm [124, Theorem 1], but are mild assumptions that hold in many practical applications. The upper bound constant can improve upon upper bounds in total variation [113] if for example, ρ is continuous and Θ is compact. If $\epsilon^* = 0$, Theorem 1

also gives the lower bound

$$\inf_{\theta\in\Theta}\int \rho(\cdot,\theta)d\Pi \leq \sup_{\theta\in\Theta}\mathcal{W}_\rho(P^t(\theta,\cdot),\Pi),$$

which shows MHI cannot converge uniformly from any starting point for many Wasserstein distances. Thus, under mild assumptions, Theorem 1 gives a complete characterization of the worst-case convergence of the MHI algorithm in many Wasserstein distances.

Exact convergence expressions are available when the Markov chain is initialized at $\theta^* = \operatorname{argmin}\{q(\theta)/\pi(\theta) : \theta \in \Theta\}$.

Proposition 1. *Suppose there exists a solution*

$$\theta^* = \operatorname{argmin}\{q(\theta)/\pi(\theta) : \theta \in \Theta\}.$$

If $\epsilon_{\theta^*} = q(\theta^*)/\pi(\theta^*)$*, then*

$$\mathcal{W}_\rho\left(P^t(\theta^*,\cdot),\Pi\right) = (1-\epsilon_{\theta^*})^t\int \rho(\theta,\theta^*)d\Pi(\theta).$$

Proof. Under our assumptions $P^t(\theta^*,\cdot)$ can be represented as a convex combination of the target distribution and the Dirac measure at the point θ^* [124, Remark 1, Theorem 2], that is,

$$P^t(\theta^*,\cdot) = \left(1-(1-\epsilon_{\theta^*})^t\right)\Pi + (1-\epsilon_{\theta^*})^t\,\delta_{\theta^*}. \tag{2.1}$$

Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be a function such that $\int_\Theta |\psi| d\Pi < \infty$. We have the identity,

$$\int_\Theta \psi dP^t(\theta^*,\cdot) = \left(1-(1-\epsilon_{\theta^*})^t\right)\int_\Theta \psi d\Pi(B) + (1-\epsilon_{\theta^*})^t\,\psi(\theta^*). \tag{2.2}$$

Since the only coupling between Π^* and the Dirac measure δ_{θ^*} is the independent coupling [40], the Wasserstein distance takes the simple form $\mathcal{W}_\rho\left(\delta_{\theta^*}, \Pi\right) = \int \rho(\theta, \theta^*) d\Pi(\theta)$.

Since q is not exactly π, then $\epsilon_{\theta^*} \in (0,1)$. Let $M_b(\mathbb{R}^d)$ be the set of bounded measurable functions on $\mathbb{R}^d$ and for real-valued functions φ, let

$$\|\varphi\|_{\mathrm{Lip}(\rho)} = \sup_{x,y,x\neq y} \{|\varphi(x) - \varphi(y)|/\rho(x,y)\}$$

denote the Lipschitz norm with respect to the distance ρ. Applying the Kantorovich-Rubinstein theorem [122, Theorem 1.14],

$$\begin{aligned}
\mathcal{W}_\rho\left(P^t(\theta^*,\cdot), \Pi\right) &= \sup_{\substack{\varphi\in M_b(\mathbb{R}^d) \\ \|\varphi\|_{\mathrm{Lip}(\rho)}\leq 1}} \int_\Theta \varphi d\left(P^t(\theta^*,\cdot) - \Pi\right) \\
&= \sup_{\substack{\varphi\in M_b(\mathbb{R}^d) \\ \|\varphi\|_{\mathrm{Lip}(\rho)}\leq 1}} \left\{(1-\epsilon_{\theta^*})^t \int_\Theta \varphi d\left(\delta_{\theta^*} - \Pi\right)\right\} \\
&= (1-\epsilon_{\theta^*})^t \sup_{\substack{\varphi\in M_b(\mathbb{R}^d) \\ \|\varphi\|_{\mathrm{Lip}(\rho)}\leq 1}} \int_\Theta \varphi d\left(\delta_{\theta^*} - \Pi\right) \\
&= (1-\epsilon_{\theta^*})^t \mathcal{W}_\rho\left(\delta_{\theta^*}, \Pi\right) \\
&= (1-\epsilon_{\theta^*})^t \int_\Theta \rho(\theta,\theta^*) d\Pi(\theta).
\end{aligned}$$

□

.

2.3.1 Application: Bayesian quantile regression

Fix $r \in (0,1)$ and suppose, for $i = 1,\ldots,n$, ϵ_i are independent and identically distributed (i.i.d.) with density

$$p_r(\epsilon) = r(1-r)\left(\exp((1-r)\epsilon) I_{\epsilon<0} + \exp(-r\epsilon) I_{\epsilon\geq 0}\right).$$

Let $v_0, s_0 \in (0, \infty)$ and $C \in \mathbb{R}^{d \times d}$ be symmetric positive-definite. We paramaterize the inverse gamma distribution so that if $\sigma \sim \mathrm{IG}(v, s)$ for some $v, s \in (0, \infty)$, then it has a density proportional to $\sigma^{-v-1} \exp(-s/\sigma)$. Assume the Bayesian quantile regression model for $i \in 1, \ldots, n$ where $X_i \in \mathbb{R}^d$ is fixed and

$$\begin{aligned} \sigma &\sim \mathrm{IG}(v_0, s_0) \\ \beta|\sigma &\sim \mathrm{N}_d(0, \sigma C) \\ Y_i &= \beta^T X_i + \sigma \epsilon_i. \end{aligned}$$

With $Y = (Y_1, \ldots, Y_n)^T$ and $X = (X_1, \ldots, X_n)^T$, let $\Pi(\cdot|X, Y)$ denote the posterior and $\pi(\cdot|X, Y)$ denote the density for this Bayesian model with normalizing constant $Z_{\Pi(\cdot|X,Y)}$.

Upper bounds on the convergence rate were previously investigated for Gibbs samplers [65] in this setting. We will study the MHI algorithm with a normal-inverse-gamma proposal constructed as follows. Define $\ell_r(u) = u(r - I_{u<0})$ and $s_{n,r} : \mathbb{R}^d \to \mathbb{R}$ by $s_{n,r}(\beta) = \sum_{i=1}^n \ell_r(Y_i - \beta^T X_i) + \beta^T C^{-1} \beta/2$. Since $s_{n,r}$ is strongly convex, let $\beta^* \in \mathbb{R}^d$ be the unique minimum of the function $s_{n,r}$. Now the MHI proposal is given by

$$\begin{aligned} \sigma &\sim \mathrm{IG}(n + v_0, s_0 + s_{n,r}(\beta^*)) \\ \beta|\sigma &\sim \mathrm{N}_d(\beta^*, \sigma C). \end{aligned}$$

Let $\Gamma : \mathbb{R} \to \mathbb{R}$ be the usual Gamma function and define

$$\epsilon_{\beta^*} = Z_{\Pi(\cdot|X,Y)} (2\pi)^{-\frac{d}{2}} \det(C)^{-1/2} (s_0 + s_{n,r}(\beta^*))^{n+v_0} \Gamma(n + v_0)^{-1}.$$

The following gives an exact convergence rate of this algorithm which completely characterizes its convergence from a specific initialization.

Theorem 2. *For any* $\sigma_0 \in (0, \infty)$

$$\mathcal{W}_\rho \left(P^t((\beta^*, \sigma_0), \cdot), \Pi(\cdot|X, Y) \right) = (1 - \epsilon_{\beta^*})^t \int \rho((\beta, \sigma), (\beta^*, \sigma_0)) d\Pi(\beta, \sigma|X, Y).$$

Proof. The proof is provided in Appendix Appendix A.2. □

Note that ϵ_{β^*} is difficult to compute since it depends on the normalizing constant, but we give an example later where upper bounding the convergence rate is possible in Bayesian logistic regression.

2.4 MHI with centered Gaussian proposals

Recently, centered drift functions have been used to improve convergence analyses of some Monte Carlo Markov chains [30, 88, 91]. Our focus is instead on centering the proposal distribution, that is, matching the optimal points of the proposal and target densities similar to Laplace approximations.

We shall see in the next section that by centering a Gaussian proposal, we may satisfy the assumptions of Proposition 1 for a general class of target distributions with θ^* being the optimum of the target's density. While we focus on Gaussian proposals, the technique of centering proposals is in fact more general.

We will assume the target distribution Π is a probability distribution supported on $\mathbb{R}^d$. With $f : \mathbb{R}^d \to \mathbb{R}$ and normalizing constant Z_Π, define the density π by $\pi(\theta) = Z_\Pi^{-1} \exp(-f(\theta))$. Let θ^* be the unique maximum of π, $\alpha \in (0, +\infty)$, and $C \in \mathbb{R}^{d \times d}$ be a symmetric, positive-definite matrix. Let the proposal distribution Q with density q correspond to a d-dimensional Gaussian distribution, $\mathrm{N}_d(\theta^*, \alpha^{-1}C)$. In this case, the ratio of the proposal density and target density is $\epsilon_{\theta^*} = (2\pi)^{-d/2} \alpha^{d/2} \det(C)^{-1/2} Z_\Pi \exp(f(\theta^*))$.

Proposition 2. *If θ^* exists and, for any $\theta \in \mathbb{R}^d$,*

$$f(\theta) \geq f(\theta^*) + \alpha \left(\theta - \theta^*\right)^T C^{-1} \left(\theta - \theta^*\right)/2,$$

then

$$\mathcal{W}_\rho \left(P^t(\theta^*, \cdot), \Pi\right) = (1 - \epsilon_{\theta^*})^t \int \rho(\theta, \theta^*) d\Pi(\theta).$$

Proof. Since the proposal density has been centered at the point θ^*, it then satisfies $q(\theta^*) = (2\pi)^{-d/2} \alpha^{d/2} \det(C)^{-1/2}$. For every $\theta \in \mathbb{R}^d$, we have the lower bound

$$\begin{aligned}
\frac{q(\theta)}{\pi(\theta)} &= (2\pi)^{-d/2} \alpha^{d/2} \det(C)^{-1/2} Z_\Pi \exp\left(f(\theta) - \frac{\alpha}{2} \left(\theta - \theta^*\right)^T C^{-1} \left(\theta - \theta^*\right)\right) \\
&\geq (2\pi)^{-d/2} \alpha^{d/2} \det(C)^{-1/2} Z_\Pi \exp\left(f(\theta^*)\right) \\
&= \frac{q(\theta^*)}{\pi(\theta^*)}.
\end{aligned}$$

Since both densities are positive and the proposal is independent of the previous iteration, we have shown that the conditions for Proposition 1 are satisfied and an application of Proposition 1 with the proposal and target distribution Q and Π as we have defined them in this section completes the proof. □

The point θ^* is guaranteed to exist if the function f satisfies a convexity property. A function $h : \mathbb{R}^d \to \mathbb{R}$ is strongly convex with parameter μ if there is an $\mu \in (0, +\infty)$ so that $h(\cdot) - \mu \|\cdot\|^2/2$ is convex [49, 80]. The norm in this definition is often taken to be the Euclidean norm, but we will use the norm induced by the matrix C^{-1}. We consider using a Gaussian proposal centered at a point θ_0 which is not necessarily the optimum of the target density. Let $g_{f(\theta_0)} \in \mathbb{R}^d$ be a subgradient of f at θ_0 [80]. For a point $\theta_0 \in \mathbb{R}^d$, we consider the proposal corresponding to a d-dimensional Gaussian distribution, $\mathrm{N}_d(\theta_0 - \alpha^{-1} C g_{f(\theta_0)}, \alpha^{-1} C)$. When f is differentiable, this construction of the proposal uses the gradient of f in a similar way as MALA. The ratio of the proposal

and target density evaluated at θ_0 is $\epsilon_{\theta_0} = (2\pi)^{-d/2} \det(\alpha^{-1}C)^{-1/2} Z_\Pi \exp(f(\theta_0) - g_{f(\theta_0)}{}^T C g_{f(\theta_0)}/(2\alpha))$. Choosing $\theta_0 \equiv \theta^*$ yields the centered Gaussian proposal, but we also have an exact convergence expression in other cases.

Proposition 3. *If the function $\theta \mapsto f(\theta) - \alpha\theta^T C^{-1}\theta/2$ is convex for all points on $\mathbb{R}^d$, then*

$$\mathcal{W}_\rho\left(P^t(\theta_0, \cdot), \Pi\right) = (1 - \epsilon_{\theta_0})^t \int \rho(\theta, \theta_0) d\Pi(\theta).$$

Proof. Since the function $f(\theta) - \alpha\theta^T C^{-1}\theta/2$ is convex for all points on $\mathbb{R}^d$, then for each $\theta \in \mathbb{R}^d$,

$$\begin{aligned} f(\theta) &\geq f(\theta_0) + g_{f(\theta_0)}{}^T(\theta - \theta_0) + \frac{\alpha}{2}(\theta - \theta_0)^T C^{-1}(\theta - \theta_0) \\ &= f(\theta_0) - \frac{1}{2\alpha}(C g_{f(\theta_0)})^T g_{f(\theta_0)} \\ &\quad + \frac{\alpha}{2}\left(\theta - \theta_0 + \alpha^{-1} C g_{f(\theta_0)}\right)^T C^{-1}\left(\theta - \theta_0 + \alpha^{-1} C g_{f(\theta_0)}\right). \end{aligned}$$

This implies for every $\theta \in \mathbb{R}^d$, the ratio of the proposal density q corresponding to the distribution $\mathrm{N}_d(\theta_0 - \alpha^{-1} C g_{f(\theta_0)}, \alpha^{-1} C)$ and target density π satisfies

$$\frac{q(\theta)}{\pi(\theta)} \geq \frac{q(\theta_0)}{\pi(\theta_0)} = \epsilon_{\theta_0}.$$

An application of Proposition 1 completes the proof. □

2.4.1 Application: Bayesian generalized linear models

We consider Bayesian Poisson and negative-binomial regression for count response regression and Bayesian logistic and probit regression for binary response regression. Suppose there are n discrete-valued responses Y_i with features $X_i \in \mathbb{R}^d$, and a parameter $\beta \in \mathbb{R}^d$. For Poisson regression, assume the Y_i's are conditionally independent

with

$$Y_i \big| X_i, \beta \sim \text{Poisson}\left(\exp\left(\beta^T X_i\right)\right).$$

Similarly, for negative-binomial regression, if $\xi \in (0, +\infty)$, assume

$$Y_i \big| X_i, \beta \sim \text{Negative-Binomial}\left(\xi, \left(1 + \exp\left(-\beta^T X_i\right)\right)^{-1}\right).$$

For binary response regression, if $S : \mathbb{R} \to (0, 1)$, assume

$$Y_i | X_i, \beta \sim \text{Bernoulli}\left(S\left(\beta^T X_i\right)\right).$$

For logistic regression, we will consider $S(x) = (1 + \exp(x))^{-1}$ and for probit regression, we will consider $S(x)$ to be the cumulative distribution function of a standard Gaussian random variable. Suppose $\beta \sim \text{N}_d(0, \alpha^{-1}C)$ where $\alpha \in (0, +\infty)$ and $C \in \mathbb{R}^{d \times d}$ is a symmetric, positive-definite matrix. Both α and C are assumed to be known. Define the vector $Y = (Y_1, \ldots, Y_n)^T$ and the matrix $X = (X_1, \ldots, X_n)^T$. Let $\Pi(\cdot|X, Y)$ denote the posterior with density $\pi(\cdot|X, Y)$. If ℓ_n denotes the negative log-likelihood for each model, the posterior density is characterized by

$$\pi(\beta|X, Y) = Z^{-1}_{\Pi(\cdot|X,Y))} \exp\left(-\ell_n(\beta) - \frac{\alpha}{2}\beta^T C^{-1}\beta\right).$$

Observe that the function ℓ_n is convex in all four models we consider. Let β^* denote the unique maximum of $\pi(\cdot|X, Y)$. For the MHI algorithm, we use a $\text{N}_d(\beta^*, \alpha^{-1}C)$ proposal distribution, and Proposition 3 immediately yields the following for each posterior.

Corollary 1. *We have*

$$\mathcal{W}_\rho\left(P^t(\beta^*, \cdot), \Pi(\cdot|X, Y)\right) = (1 - \epsilon_{\beta^*})^t \int \rho(\beta, \beta^*) d\Pi\left(\beta|X, Y\right)$$

where $\epsilon_{\beta^*} = \exp(\ell_n(\beta^*) + \frac{\alpha}{2}\beta^{*T}C^{-1}\beta^*)Z_{\Pi(\cdot|X,Y)}\left((2\pi)^{d/2}\det\left(\alpha^{-1}C\right)^{1/2}\right)^{-1}$.

2.4.2 Convergence complexity analysis of MHI in binary response regression

Our goal now is to obtain an upper bound on the rate of convergence established in Corollary 1 in high dimensions for binary response regression. In this context, it is natural to treat the $(Y_i, X_i)_{i=1}^n$ as stochastic; each time the sample size increases, the additional observation is randomly generated. Specifically, we will assume that $(Y_i, X_i)_{i=1}^n$ are independent with $Y_i|X_i, \beta \sim \text{Bernoulli}\left(S\left(\beta^T X_i\right)\right)$ and $X_i \sim N_d(0, \sigma^2 n^{-1} I_d)$ with $\sigma^2 \in (0, +\infty)$. Similar scaling assumptions on the data are used for high-dimensional maximum-likelihood theory in logistic regression [109]. We will also assume the limit of the trace of the covariance matrix used in our prior is finite, that is, $tr(C) \to s_0 \in (0, +\infty)$ as $d \to +\infty$. Note that this is a necessary condition that the trace of the covariance is finite for Gaussian distributions to exist in an infinite-dimensional Hilbert space [9].

Theorem 3. *Suppose that the following conditions hold.*

1. *The negative log-likelihood ℓ_n is a twice continuously differentiable convex function.*

2. *There is a universal constant $r_0 \in (0, +\infty)$ such that the largest eigenvalue of the Hessian of the negative log-likelihood H_{ℓ_n} satisfies for every $\beta \in \mathbb{R}^d$,*

$$\lambda_{max}\left(H_{\ell_n}(\beta)\right) \leq r_0 \lambda_{max}\left(X^T X\right).$$

Let $a_0 = r_0(1+\gamma^{1/2})^2\sigma^2 s_0/(2\alpha)$. If $d, n \to +\infty$ in such a way that $d/n \to \gamma \in (0, +\infty)$,

then, almost surely

$$\limsup_{d/n\to\gamma} \mathcal{W}_\rho \left(P^t(\beta^*,\cdot), \Pi(\cdot|X,Y)\right) \leq M_0(1-\exp\left(-a_0\right))^t$$

where $M_0 = \limsup_{d/n\to\gamma} \int \rho(\beta,\beta^*) d\Pi(\beta|X,Y)$.

Proof. The proof is provided in Appendix Appendix A.3. □

Theorem 3 applies to both Bayesian logistic and probit regression. For logistic regression, ℓ_n is a twice continuously differentiable convex function and we may choose $r_0 = 4^{-1}$. Similarly for probit regression, ℓ_n is also a twice continuously differentiable convex function and we may choose $r_0 = 1$ [24].

In Figure 2.1, we plot $(1-\exp\left(-a_0\right))^t$, the limiting decrease in the Wasserstein distance according to our upper bound, with varying values of the limiting ratio γ with the other remaining values in a_0 fixed. We observe that as this ratio increases, the number of iterations needed to approximately converge may still increase rather rapidly.

2.5 Final remarks

We studied the exact convergence behavior of the MHI sampler across general Wasserstein distances. We showed upper and lower bounds on the worst-case convergence rate for Wasserstein distances weaker than the total variation distance. When starting at a certain point, we gave exact convergence expressions. It remains an open question if the convergence rate is the same for every initial starting point as it is for total variation [124]. By centering an independent proposal, we directly imbue the Markov chain with a strong attraction to a set where the target distribution has high probability. We showed this technique can provide uniform control over acceptance probability yielding exact convergence rates in Bayesian quantile regression.

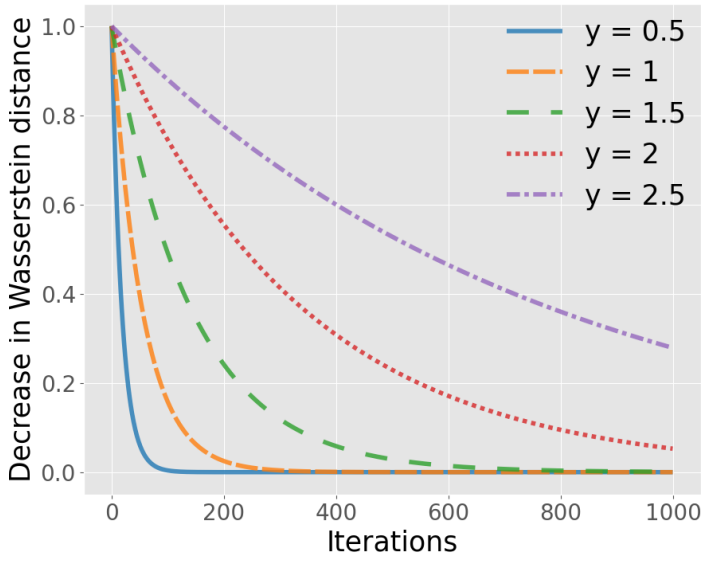


Figure 2.1: The limiting decrease in the Wasserstein distance using different values of γ, the limiting ratio of the dimension and sample size, versus the number of iterations.

The centered MHI algorithm turns out to have many applications for posteriors that arise in Bayesian generalized linear models where more sophisticated proposals are often used. With additional assumptions on the data and prior, we also showed this exact convergence rate may be upper bounded when sampling high-dimensional posteriors in Bayesian binary response regression. Explicit convergence rates for Metropolis-Hastings algorithms such as MALA and RWM are difficult to obtain especially in high dimensions. Although connections between the MHI algorithm and rejection sampling are well-known [71, 113, 124], the centered MHI algorithm may provide insight into the convergence behavior of more popular Metropolis-Hastings algorithms such as the MALA and RWM algorithms.

Chapter 3

Lower Bounds on the Rate of Convergence for Metropolis-Hastings Algorithms

3.1 Overview

In this chapter, we develop new lower bounds on the convergence rates for geometrically ergodic Metropolis-Hastings algorithms on $\mathbb{R}^d$ in order to study their computational complexity. In particular, we are concerned with conditions on the tuning parameters which imply the convergence rate rapidly tends to 1 with the dimension d and sample size n. When these algorithms are not necessarily geometrically ergodic, we also develop new lower bounds on the mixing times. Analogous lower bounds are developed for the geometric convergence rate and mixing times in certain Wasserstein distances. Our goal is to aid practitioners in understanding when geometrically converging Metropolis-Hastings algorithms may *fail* to produce a representative sample within an available number of iterations.

Few results are available on lower bounding the convergence rates for geometrically ergodic Metropolis-Hastings algorithms [97, 124]. While drift and minorization conditions are often used to establish geometric ergodicity [113, 53, 98], the lower

bounds we develop are independent of the method used where specific techniques have limitations [89]. Lower bounds on the convergence rates which tend to 1 rapidly also suggest estimators from the Markov chain may be unreliable in these settings. For example, convergence rates are used to upper bound the asymptotic variance in the Markov chain central limit theorem [35, 57] and mean squared error [104, 68] as well as in concentration inequalities [60].

We apply our lower bounds to the random walk Metropolis-Hastings (RWM) algorithm for certain log-concave target densities and an application to Bayesian logistic with Zellner's g-prior is studied as $d, n \to \infty$. In this example, we show the convergence rate can tend to 1 exponentially fast if the tuning parameters do not depend carefully on d and n. We then focus on a class of Metropolis-Hastings algorithms using a general Gaussian proposal which in special cases is the proposal used in random-walk Metropolis-Hastings (RWM) [113], Metropolis-adjusted Langevin algorithm (MALA) [96, 98], and other variants such as Riemannian manifold MALA [41]. We consider more general target densities under conditions they concentrate towards their maximum point with n (e.g. Bayesian posterior concentration, Laplace approximations) as $d, n \to \infty$. We show if the tuning parameters do not depend on n, the convergence rate can tend to 1 rapidly as $d, n \to \infty$. Flat prior Bayesian logistic as only $n \to \infty$ is studied as an application.

The approach here is inspired by previous techniques used to obtain exact convergence rates in the Metropolis-Hastings independence sampler [124]. Others have instead focused on the conductance and spectral theory [10, 46, 55, 69, 126]. The conductance has been used to upper bound the spectral gap in parallel and simulated tempering [126] and for more general Metropolis-Hastings where a Gaussian target was studied with RWM [46, Proposition 2.16]. An upper bound on the spectral gap lower bounds the convergence rate for reversible Markov chains starting from appropriate "warm-start" distributions [35, 69]. We focus on convergence rates for

Metropolis-Hastings starting from an arbitrary point which is often easier to implement in practice.

Our work is also related to previous research focused on lower bounding the mixing time for Metropolis-Hastings algorithms [105, 126, 127]. An upper bound on the spectral gap for parallel tempering [126] was extended to lower bound the mixing time for parallel tempering and adaptive variants using the conductance [105]. A general lower bound on the mixing time for Metropolis-Hastings was shown using the conductance [10, Theorem 3.3] where a standard Gaussian target was studied with RWM. A general lower bound on the uniform mixing time for reversible, uniformly ergodic Markov chains uses spectral theory [127]. In Section 3.5, we show the lower bounds in this chapter can be obtained under stronger assumptions using the conductance.

Recently, a significant research interest has emerged regarding the dimension dependence on the mixing time for the MALA algorithm [18, 28, 64, 70]. Lower bounds on the total variation mixing times were shown for MALA in some specific examples [18, 64, 70]. A more general lower bound on the mixing time in the chi-square divergence was also developed for reversible Markov chains [64].

The remainder is organized as follows. In Section 3.2, we develop general lower bounds for the convergence rates and mixing times of Metropolis-Hastings algorithms in total variation and Wasserstein distances. In Section 3.3, the RWM algorithm is studied for certain log-concave target densities and an application to Bayesian logistic regression with Zellner's g-prior is studied. In Section 3.4, we study a class of Metropolis-Hastings algorithms with a general Gaussian proposal under concentration and develop conditions when the convergence rate rapidly tends to 1 as $d, n \to \infty$. An example with flat prior Bayesian logistic regression is studied in detail as $n \to \infty$. In Section 3.5, we compare our techniques to previous techniques using the conductance under stronger assumptions. In Section 3.6, we discuss our findings and future research directions.

3.2 Lower bounds for Metropolis-Hastings

Let $\mathbb{Z}_+$ denote the set of positive integers and for $d \in \mathbb{Z}_+$, let $\mathbb{R}^d$ be the $d-$dimensional Euclidean space. For probability measures, the Borel sigma algebra will be assumed throughout. Sets and functions will automatically be assumed to be Borel measurable without reference unless otherwise stated. Let $\text{supp}(f)$ denote the support of a function f, that is, the set such that $f \neq 0$. Let δ_x denote the Dirac measure at the point x. We will denote the p-norms on $\mathbb{R}^d$ by $\|\cdot\|_p$. In some cases, we denote $\wedge, \vee$ for the minimum and maximum respectively.

We will consider Metropolis-Hastings algorithms with target distribution Π^* on $\mathbb{R}^d$ with Lebesgue density π^*. Starting from θ_0, the Metropolis-Hastings algorithm generates a Markov chain $\theta_0, \theta_1, \ldots, \theta_t$. For each $\theta \in \mathbb{R}^d$, define a proposal by the Markov kernel $Q(\theta, \cdot)$ on $\mathbb{R}^d$ with Lebesgue density $q(\theta, \cdot)$. The proposal is used to generate proposed samples and these samples are accepted or rejected using an acceptance ratio

$$a(\theta, \theta') = \begin{cases} \frac{\pi^*(\theta')q(\theta', \theta)}{\pi^*(\theta)q(\theta, \theta')} \wedge 1, & \text{if } \pi^*(\theta)q(\theta, \theta') > 0 \\ 1, & \text{if } \pi^*(\theta)q(\theta, \theta') = 0 \end{cases}.$$

For $t \in \mathbb{Z}_+$, the algorithm generates $\theta'_t|\theta_{t-1} \sim Q(\theta_{t-1}, \cdot)$ and $U_t \sim \text{Uniform}(0, 1)$ independently producing

$$\theta_t = \begin{cases} \theta'_t, & \text{if } U_t \leq a\left(\theta_{t-1}, \theta'_t\right) \\ \theta_{t-1}, & \text{otherwise} \end{cases}.$$

Define the acceptance probability by

$$A(\theta) = \int_{\mathbb{R}^d} a(\theta, \theta')q(\theta, \theta')d\theta' \in (0, 1].$$

The Metropolis-Hastings Markov kernel P with proposal density defined by $\theta \mapsto q(\theta, \cdot)$ and target density π^* is defined for $\theta \in \mathbb{R}^d$ and sets $B \subseteq \mathbb{R}^d$ by

$$P(\theta, B) = \int_B a(\theta, \theta') q(\theta, \theta') d\theta' + \delta_\theta(B) \left(1 - A(\theta)\right).$$

The Markov kernel for iteration time $t \in \mathbb{Z}_+$ with $t \geq 2$ is defined recursively by

$$P^t(\theta, \cdot) = \int_{\mathbb{R}^d} P^{t-1}(\cdot, \cdot) dP(\theta, \cdot)$$

where $P \equiv P^1$.

We will define the total variation distance between probability measures μ, ν on $\mathbb{R}^d$ by

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{f:\mathbb{R}^d \to [0,1]} \left| \int f d\mu - \int f d\nu \right|.$$

The total variation lower bound for Metropolis-Hastings independence samplers on $\mathbb{R}^d$ [124, Lemma 1] can be extended more generally using a similar technique. The following lower bound intuitively says a Metropolis-Hastings algorithm cannot converge faster than its first acceptance from the point at which it starts.

Lemma 1. *For every $t \in \mathbb{Z}_+$ and every $\theta \in \mathbb{R}^d$,*

$$\left\| P^t(\theta, \cdot) - \Pi^* \right\|_{TV} \geq \left(1 - A(\theta)\right)^t.$$

Proof. Fix $\theta \in \mathbb{R}^d$ and define the bounded function $\varphi(\cdot) = I_{\{\theta\}}(\cdot)$. Using that $\varphi(\theta) =$

1,

$$\begin{aligned}\int_{\mathbb{R}^d} \varphi dP^t(\theta,\cdot) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \varphi dP^{t-1}(\theta',\cdot)dP(\theta,\theta') \\ &\geq (1-A(\theta))\int_{\mathbb{R}^d} \varphi dP^{t-1}(\theta,\cdot) \\ &\geq (1-A(\theta))^t\varphi(\theta) \\ &= (1-A(\theta))^t.\end{aligned}$$

Since π^* is a Lebesgue density on $\mathbb{R}^d$, $\int_{\mathbb{R}^d} \varphi d\Pi^* = 0$. Let Ψ be the set of functions $\psi : \mathbb{R}^d \to [0,1]$. Combining these, we have

$$\begin{aligned}\left\|P^t(\theta,\cdot) - \Pi^*\right\|_{\mathrm{TV}} &= \sup_{\psi\in\Psi}\left|\int_{\mathbb{R}^d}\psi dP^t(\theta,\cdot) - \int_{\mathbb{R}^d}\psi d\Pi^*\right| \\ &\geq \int_{\mathbb{R}^d}\varphi dP^t(\theta,\cdot) - \int_{\mathbb{R}^d}\varphi d\Pi^* \\ &= \int_{\mathbb{R}^d}\varphi dP^t(\theta,\cdot) \\ &\geq (1-A(\theta))^t.\end{aligned}$$

□

Lemma 1 will yield a lower bound on the rate of convergence for geometrically ergodic Metropolis-Hastings algorithms which we now define [113]. A Markov kernel P is (ρ, M)-geometrically ergodic if there is a $\rho \in (0,1)$ and a function $\theta \mapsto M(\theta) \in (0,\infty)$ such that for every $\theta \in \mathrm{supp}(\pi^*)$ and every $t \in \mathbb{Z}_+$,

$$\left\|P^t(\theta,\cdot) - \Pi^*\right\|_{\mathrm{TV}} \leq M(\theta)\rho^t.$$

The most popular tools for establishing this geometric ergodicity are drift and minorization conditions [45, 53, 58, 76, 102]. Few results for Metropolis-Hastings give

explicit values for ρ, but sufficient conditions for the existence of ρ and M are available for the RWM algorithm [53] and also the MALA algorithm [98]. Independent of the method used, we have the following lower bound in terms of the rejection probability.

Lemma 2. *If P is (ρ, M)-geometrically ergodic, then for any $\theta \in \mathbb{R}^d$,*

$$\rho \geq 1 - A(\theta).$$

Proof. The lower bound is trivial for $\theta \notin \text{supp}(\pi^*)$ since $A(\theta) = 1$ in this case. Fix $\theta \in \text{supp}(\pi^*)$. We apply Lemma 1 and use the assumed geometric ergodicity to obtain for every $t \in \mathbb{Z}_+$,

$$M(\theta)\rho^t \geq \left\| P^t(\theta, \cdot) - \Pi^* \right\|_{\text{TV}} \geq (1 - A(\theta))^t$$

and thus

$$M(\theta)^{1/t}\rho \geq 1 - A(\theta).$$

Since this holds for every $t \in \mathbb{Z}_+$, this implies

$$\rho = \lim_{t \to \infty} M(\theta)^{1/t}\rho \geq 1 - A(\theta).$$

□

Suppose then for some point θ_0, we can find a rate function $r : \mathbb{Z}_+ \to (0, \infty)$ such that $\lim_{d\to\infty} r(d) = 0$ and $A(\theta_0) \leq r(\cdot)$. If in this case the Metropolis-Hastings algorithm is (ρ, M)-geometrically ergodic, then $r(\cdot) \geq 1 - \rho$. In other words, if the acceptance probability is tending to 0 for just one starting point, then the geometric rate of convergence tends to 1. We will be interested when the rate function tends to 0 at a rapid rate with its input being the dimension d or the sample size n.

Lemma 1 can also be used to lower bound the mixing time. For $\epsilon \in (0, 1)$, and a

nonempty set $K \subset \mathbb{R}^d$, define the (ϵ, K)-mixing time for P by

$$\tau_{\epsilon,K} = \inf \left\{ t \in \mathbb{Z}_+ : \sup_{\theta \in K} \left\| P^t(\theta, \cdot) - \Pi^* \right\|_{\mathrm{TV}} \leq \epsilon \right\}.$$

where it is understood that $\inf \varnothing = \infty$. Alternative definitions of the mixing time instead initialize the algorithm from a certain class of distributions [18, 28, 64, 105, 70, 126, 127, 128].

Lemma 3. *For any $\theta \in \mathbb{R}^d$, if $\theta \in K$, the (ϵ, K)-mixing for P satisfies*

$$\tau_{\epsilon,K} \geq \log(\epsilon^{-1}) \left(\frac{1}{A(\theta)} - 1 \right).$$

Proof. If $\tau_{\epsilon,K} = \infty$, there is nothing to do, so assume $\tau_{\epsilon,K} < \infty$. We apply Lemma 1 to obtain

$$\begin{aligned} \epsilon &\geq \sup_{\theta' \in K} \| P^{\tau_{\epsilon,K}}(\theta', \cdot) - \Pi^* \|_{\mathrm{TV}} \\ &\geq \| P^{\tau_{\epsilon,K}}(\theta, \cdot) - \Pi^* \|_{\mathrm{TV}} \\ &\geq (1 - A(\theta))^{\tau_{\epsilon,K}}. \end{aligned}$$

For $x \in (0,1)$, $-\log(1-x) \leq x/(1-x)$. Since $\epsilon \in (0,1)$, then for $A(\theta) \in (0,1)$,

$$\tau_{\epsilon,K} \geq \frac{\log(\epsilon^{-1})}{-\log\left(1 - A(\theta)\right)} \geq \log(\epsilon^{-1}) \left(\frac{1}{A(\theta)} - 1 \right).$$

If $A(\theta) = 1$, the lower bound is trivial. □

If the acceptance probability is exponentially decreasing with the dimension, Lemma 3 implies the mixing time is at least exponential in the dimension or *slow mixing* which relates to previous work [105]. Lemma 3 also gives necessary conditions for a finite uniform mixing time.

Lemma 4. *If* $\inf_{\theta\in\mathbb{R}^d} A(\theta) = 0$*, the* $(\epsilon, \mathbb{R}^d)$*-mixing time for* P *satisfies* $\tau_{\epsilon,\mathbb{R}^d} = \infty$*.*

Proof. Since $\inf_{\theta\in\mathbb{R}^d} A(\theta) = 0$, choose a sequence $(\theta_n) \subset \mathbb{R}^d$ so that $\lim_{n\to\infty} A(\theta_n) = 0$. For every $\theta_n \in \mathbb{R}^d$, by Lemma 3,

$$\tau_{\epsilon,\mathbb{R}^d} \geq \log(\epsilon^{-1}) \left(\frac{1}{A(\theta_n)} - 1 \right)$$

Since $\epsilon \in (0, 1)$, taking the limit $n \to \infty$ completes the proof. □

3.2.1 Wasserstein lower bounds

Recently, the Wasserstein distances [61, 122, 123] from optimal transportation distances have become popular in the convergence analysis of high-dimensional MCMC algorithms [21, 26, 46, 88, 92]. Mainly this is due to improved scaling properties in high dimensions compared to the total variation distance. We develop analogous lower bounds developed previously in Section 3.2 for many Wasserstein distances.

For two probability measures μ, ν, let $\mathcal{C}(\mu, \nu)$ be the set of all joint probability measures with marginals μ, ν. With a metric cost function $c(\cdot, \cdot)$, the c-Wasserstein distance or Kantorovich distance is defined

$$\mathcal{W}_c(\mu, \nu) = \inf_{\Gamma\in\mathcal{C}(\mu,\nu)} \int c(\theta, \omega) d\Gamma(\theta, \omega).$$

We will consider Wasserstein distances satisfying the following condition:

Assumption 1. *Suppose* $c(\cdot, \cdot) \geq \|\cdot - \cdot\|_1$.

Since norms are equivalent on $\mathbb{R}^d$, these cost metrics include $c(\cdot, \cdot) = c_0 \|\cdot - \cdot\|$ for any norm $\|\cdot\|$ on $\mathbb{R}^d$ and some constant $c_0 \in (0, \infty)$. We also consider the following mild assumption on π^*.

Assumption 2. *Suppose* $s = \sup_{\theta\in\mathbb{R}^d} \pi^*(\theta) < \infty$.

We begin with a lower bound on the Wasserstein distance for general Metropolis-Hastings Markov chains.

Theorem 4. *Let assumptions 1 and 2 hold. Then with* $C_{d,\pi^*} = d\left(2s^{\frac{1}{d}}(1+d)^{1+\frac{1}{d}}\right)^{-1}$, *we have the lower bound for every* $t \in \mathbb{Z}_+$ *and every* $\theta \in \mathbb{R}^d$*:*

$$\mathcal{W}_c(P^t(\theta,\cdot),\Pi^*) \geq C_{d,\pi^*}\left(1 - A(\theta)\right)^{t\left(1+\frac{1}{d}\right)}.$$

Proof. We will first construct a suitable Lipschitz function. Fix $\theta \in \mathbb{R}^d$, and fix $\alpha \in (0,\infty)$. Define the function $\varphi_{\alpha,\theta} : \mathbb{R}^d \to \mathbb{R}$ by $\varphi_{\alpha,\theta}(\omega) = \exp\left(-\alpha \left\|\omega - \theta\right\|_1\right)$. We have for every $\omega, \omega' \in \mathbb{R}^d$,

$$\begin{aligned}
|\varphi_{\alpha,\theta}(\omega) - \varphi_{\alpha,\theta}(\omega')| &= |\exp\left(-\alpha\left\|\omega-\theta\right\|_1\right) - \exp\left(-\alpha\left\|\omega'-\theta\right\|_1\right)| \\
&\leq \alpha|\left\|\omega-\theta\right\|_1 - \left\|\omega'-\theta\right\|_1| \\
&\leq \alpha\left\|\omega-\omega'\right\|_1.
\end{aligned}$$

Therefore, $\alpha^{-1}\varphi_{\alpha,\theta}$ is a bounded Lipschitz function with respect to the distance $\|\cdot\|_1$ and the Lipschitz constant is 1. By assumption there exists s such that $\pi \leq s$. Then, using the fact that $\int_{\mathbb{R}^d} \exp\left(-\alpha\left\|\theta'-\theta\right\|_1\right) d\theta' = 2^d\alpha^{-d}$, we obtain

$$\begin{aligned}
\int \varphi_{\alpha,\theta} d\Pi^* &= \int \varphi_{\alpha,\theta}(\theta')\pi^*(\theta')d\theta' \\
&\leq s\int_{\mathbb{R}^d} \exp\left(-\alpha\left\|\theta'-\theta\right\|_1\right) d\theta' \\
&= s2^d\alpha^{-d}. \qquad (3.1)
\end{aligned}$$

Fix a positive integer t. Then, for each $s \in \{0,\ldots,t-1\}$ and each $\omega \in \mathbb{R}^d$, we obtain

the lower bound

$$\begin{aligned}
&\int \varphi_{\alpha,\theta}(\theta') \left(1 - A(\theta')\right)^s P(\omega, d\theta') \\
&= \int \varphi_{\alpha,\theta}(\theta') \left(1 - A(\theta')\right)^s a(\omega, \theta') dQ(\omega, \theta') + \varphi_{\alpha,\theta}(\omega) \left(1 - A(\omega)\right)^{s+1} \\
&\geq \varphi_{\alpha,\theta}(\omega) \left(1 - A(\omega)\right)^{s+1}.
\end{aligned}$$

We now apply this lower bound multiple times:

$$\begin{aligned}
\int \varphi_{\alpha,\theta}(\theta_t) P^t(\theta, d\theta_t) &= \int \left\{ \int \varphi_{\alpha,\theta}(\theta_t) P(\theta_{t-1}, d\theta_t) \right\} P^{t-1}(\theta, d\theta_{t-1}) \\
&\geq \int \varphi_{\alpha,\theta}(\theta_{t-1}) \left(1 - A(\theta_{t-1})\right) P^{t-1}(\theta, d\theta_{t-1}) \\
&\vdots \\
&\geq \varphi_{\alpha,\theta}(\theta) \left(1 - A(\theta)\right)^t \\
&= \left(1 - A(\theta)\right)^t. \qquad (3.2)
\end{aligned}$$

The final step follows from the fact that $\varphi_{\alpha,\theta}(\theta) = 1$. Combining (3.1) and (3.2), we then have the lower bound,

$$\int \left(\frac{1}{\alpha} \varphi_{\alpha,\theta}\right) dP^t(\theta, \cdot) - \int \left(\frac{1}{\alpha} \varphi_{\alpha,\theta}\right) d\Pi^* \geq \frac{\left(1 - A(\theta)\right)^t - s 2^d \alpha^{-d}}{\alpha}. \qquad (3.3)$$

The case where $\mathcal{W}_{\|\cdot\|_1}(P^t(\theta, \cdot), \Pi^*) = +\infty$ is trivial so we assume this is finite. We then have by the Kantorovich-Rubinstein theorem [122, Theorem 1.14] and the lower

bound in (3.3),

$$\begin{aligned}\mathcal{W}_{\|\cdot\|_1}(P^t(\theta,\cdot),\Pi) &\geq \sup_{\|\varphi\|_{\mathrm{Lip}(\|\cdot\|_1)}\leq 1}\left[\int \varphi dP^t(\theta,\cdot) - \int \varphi d\Pi^*\right] \\ &\geq \int \left(\frac{1}{\alpha}\varphi_{\alpha,\theta}\right) dP^t(\theta,\cdot) - \int \left(\frac{1}{\alpha}\varphi_{\alpha,\theta}\right) d\Pi^* \\ &\geq \frac{(1-A(\theta))^t - s2^d\alpha^{-d}}{\alpha}.\end{aligned}$$

If $A(\theta) = 1$, then taking the limit of $\alpha \to \infty$, completes the proof. Suppose then that $A(\theta) < 1$. Maximizing this lower bound with respect to α yields $\alpha = 2s^{\frac{1}{d}}(1+d)^{\frac{1}{d}}(1-A(\theta))^{-\frac{t}{d}}$. We then have

$$\frac{(1-A(\theta))^t - s2^d\alpha^{-d}}{\alpha} = \frac{\left(1-\frac{1}{1+d}\right)}{2s^{\frac{1}{d}}(1+d)^{\frac{1}{d}}}(1-A(\theta))^{t\left(1+\frac{1}{d}\right)}.$$

This completes the proof for the norm $\|\cdot\|_1$ and by assumption, $\mathcal{W}_c(\cdot,\cdot) \geq \mathcal{W}_{\|\cdot\|_1}(\cdot,\cdot)$ so the proof is complete. □

Similar to the total variation distance, this lower bound will yield a lower bound on the rate of convergence for Metropolis-Hastings algorithms which exponentially converge in the Wasserstein distance. We will say P is (c,ρ,M)-geometrically ergodic if there is a $\rho \in (0,1)$ and a function $\theta \mapsto M(\theta) \in (0,\infty)$ such that for every $\theta \in \mathrm{supp}(\pi^*)$ and every $t \in \mathbb{Z}_+$,

$$\mathcal{W}_c\left(P^t(\theta,\cdot),\Pi^*\right) \leq M(\theta)\rho^t.$$

Compared to the previous definition of geometric ergodicity in total variation, here we specify a metric cost function c for the Wasserstein distance. We have the following lower bound on this geometric rate of convergence.

Lemma 5. *Let assumptions 1 and 2 hold. If P is (c,ρ,M)-geometrically ergodic,*

then for any $\theta \in \mathbb{R}^d$,

$$\rho^{\frac{d}{d+1}} \geq 1 - A(\theta).$$

Proof. The lower bound is trivial for $\theta \notin \text{supp}(\pi^*)$. Fix $\theta \in \text{supp}(\pi^*)$. We apply Theorem 4 and using C_{d,π^*} as in Theorem 4, we obtain for every $t \in \mathbb{Z}_+$,

$$M(\theta)\rho^t \geq \mathcal{W}_c\left(P^t(\theta,\cdot), \Pi^*\right) \geq C_{d,\pi^*}(1 - A(\theta))^{(1+d^{-1})t}.$$

Thus,

$$M(\theta)^{1/t}\rho \geq C_{d,\pi^*}^{1/t}(1 - A(\theta))^{(1+d^{-1})}.$$

Since this holds for every $t \in \mathbb{Z}_+$,

$$\rho = \lim_{t\to\infty} M(\theta)^{1/t}\rho \geq \lim_{t\to\infty} C_{d,\pi^*}^{1/t}\left(1 - A(\theta)\right)^{(1+d^{-1})} \geq \left(1 - A(\theta)\right)^{(1+d^{-1})}.$$

□

Recall that if the rejection probability cannot be bounded below one, that is, if the acceptance probability satisfies $\inf_{\theta\in\mathbb{R}^d} A(\theta) = 0$, then a Metropolis-Hastings algorithm fails to be geometrically ergodic [99, Proposition 5.1]. Lemma 5 immediately yields the same in these Wasserstein distances under assumptions 1 and 2.

Corollary 2. *Let assumptions 1 and 2 hold. If* $\inf_{\theta\in\mathbb{R}^d} A(\theta) = 0$, *P cannot be* (c, ρ, M)*-geometrically ergodic.*

Proof. If it were (c, ρ, M)-geometrically ergodic, then by Lemma 5, $\rho^{\frac{d}{d+1}} \geq 1$ which is a contradiction. □

See [99, Proposition 5.2] and [74, Example 1] for some examples where this can happen. We also define the mixing time in the Wasserstein distance. For $\epsilon \in (0, 1)$, a metric cost function c, and a nonempty set $K \subseteq \mathbb{R}^d$, define the (c, ϵ, K)-mixing time

for P by

$$\tau_{c,\epsilon,K} = \inf\left\{t \in \mathbb{Z}_+ : \sup_{\theta\in K} \mathcal{W}_c\left(P^t(\theta,\cdot),\Pi^*\right) \leq \epsilon\right\}.$$

We have the following lower bound on this mixing time.

Lemma 6. *Let assumptions 1 and 2 hold. For any $\theta \in \mathbb{R}^d$, if $\theta \in K$, the (c,ϵ,K)-mixing time for P satisfies*

$$\tau_{c,\epsilon,K} \geq \frac{d}{d+1}\log\left(\frac{1}{8s^{1/d}\epsilon}\right)\left(\frac{1}{A(\theta)}-1\right).$$

Proof. The result is trivial if $\tau_{c,\epsilon,K} = \infty$, so we assume $\tau_{c,\epsilon,K} < \infty$. The function $d\,(1+d)^{-1-\frac{1}{d}}$ is increasing when $d \geq 1$, and at $d = 1$, this is $1/4$. We apply Theorem 4 to obtain

$$\begin{aligned}
\epsilon &\geq \sup_{\theta'\in K} \mathcal{W}_d\left(P^{\tau_{c,\epsilon,K}}(\theta',\cdot),\Pi^*\right)\\
&\geq \mathcal{W}_d\left(P^{\tau_{c,\epsilon,K}}(\theta,\cdot),\Pi^*\right)\\
&\geq \frac{d}{2s^{\frac{1}{d}}\,(1+d)^{1+\frac{1}{d}}}(1-A(\theta))^{(1+d^{-1})\tau_{c,\epsilon,K}}\\
&\geq \frac{1}{8s^{\frac{1}{d}}}(1-A(\theta))^{(1+d^{-1})\tau_{c,\epsilon,K}}.
\end{aligned}$$

Suppose first $8\epsilon s^{1/d} \leq 1$. Since $-\log(1-x) \leq x/(1-x)$ for $x \in (0,1)$, then for $A(\theta) \in (0,1)$,

$$\tau_{c,\epsilon,K} \geq \frac{\frac{d}{d+1}\log\left(\frac{1}{8s^{1/d}\epsilon}\right)}{-\log\left(1-A(\theta)\right)} \geq \frac{d}{d+1}\log\left(\frac{1}{8s^{1/d}\epsilon}\right)\left(\frac{1}{A(\theta)}-1\right).$$

If $A(\theta) = 1$ or if $8\epsilon s^{1/d} > 1$, the bound is trivial. □

Unlike in the total variation lower bound, ϵ needs to be sufficiently small for this to be meaningful. Since $d/d+1 \approx 1$ for large d, then roughly speaking, for small ϵ, we can expect similar mixing time complexity lower bounds in these Wasserstein

distances as that of the total variation distance. We also have necessary conditions for a finite mixing time.

Lemma 7. *Let assumptions 1 and 2 hold. Suppose* $\inf_{\theta\in\mathbb{R}^d} A(\theta) = 0$. *For any* $\epsilon \in (0, 1/(8s^{1/d}))$, *the* (c, ϵ, K)*-mixing time for* P *satisfies* $\tau_{c,\epsilon,\mathbb{R}^d} = \infty$.

Proof. Choose a sequence $(\theta_n) \subset \mathbb{R}^d$ so that $\lim_{n\to\infty} A(\theta_n) = 0$. By Lemma 6,

$$\tau_{c,\epsilon,\mathbb{R}^d} \geq \frac{d}{d+1} \log\left(\frac{1}{8\epsilon s^{1/d}}\right)\left(\frac{1}{A(\theta_n)} - 1\right).$$

Since $8\epsilon s^{1/d} < 1$, taking the limit $n \to \infty$ completes the proof. □

3.3 Lower bounds for RWM with log-concave targets

Let us now apply these general lower bounds to the RWM algorithm which uses a Gaussian proposal centered at the previous point with variance parameter $h \in (0, \infty)$. Let $f : \text{supp}(\pi^*) \to \mathbb{R}$, and we will write $\pi^* = Z^{-1}\exp(-f)$ where Z is the normalizing constant. A density π^* is called log-concave if f is convex on its convex support. We will consider densities where f satisfies a strong-convexity requirement [80]. For $\mu \in (0, \infty)$, we will say a function h is μ-strongly convex on a convex set S if $x \mapsto h(x) - \mu \|x\|_2^2/2$ is convex on S.

We will need to choose a point where the acceptance probability of the RWM algorithm is small. The RWM algorithm will always accept from the proposal if the proposed sample increases the target density function π^*. Intuitively, if the point θ^* maximizes π^*, then RWM should have a small acceptance probability here as no proposed move can increase π^*. We upper bound the acceptance probability at this maximum point in the following Proposition.

Proposition 4. *For $h' \in (0, \infty)$, suppose supp(π^*) is convex and f is h'^{-1}-strongly convex on supp(π^*). Let θ^* be the maximum of the target density. Then the acceptance probability for RWM satisfies*

$$A(\theta^*) \leq (h/h' + 1)^{-d/2}.$$

Proof. Using the convexity assumption, by the subgradient inequality [80, Corollary 3.2.3], for every $\theta \in \text{supp}(\pi^*)$,

$$f(\theta) - f(\theta^*) \geq \frac{1}{2h'} \|\theta - \theta^*\|_2^2 .$$

Thus,

$$\begin{aligned} A(\theta^*) &\leq \frac{1}{(2\pi h)^{d/2}} \int \exp\left(f(\theta^*) - f(\theta)\right) \exp\left(-\frac{1}{2h} \|\theta - \theta^*\|_2^2\right) d\theta \\ &\leq \frac{1}{(2\pi h)^{d/2}} \int \exp\left(-\frac{1}{2}(h'^{-1} + h^{-1}) \|\theta - \theta^*\|_2^2\right) d\theta \\ &\leq \frac{1}{(h/h' + 1)^{d/2}}. \end{aligned}$$

□

Under Proposition 4, if the RWM algorithm is (ρ, M) geometrically ergodic, then

$$(h/h' + 1)^{-d/2} \geq 1 - \rho$$

and if $\theta^* \in K$, the (e^{-1}, K)-mixing time satisfies $\tau_{e^{-1},K} \geq (h/h' + 1)^{d/2} - 1$. The results of Section 3.2.1 can also be applied for Wasserstein distances. In particular, it must be the case that $h \propto h'/d$ or else the convergence rate can tend to 1 with the dimension which corresponds to optimal scaling guidelines where $h = 2.38^2/d$ [95].

An adversarial example using a Gaussian target showed the spectral gap for the

RWM algorithm tends to 0 polynomially fast with the dimension [46]. Using Proposition 4, we show an adversarial example in a non-toy example where the convergence rate can tend to 1 exponentially in the dimension.

Example 1. *(RWM for Bayesian logistic regression) Consider the Bayesian logistic regression model where for* $i \in 1, \ldots, n$, $X_i \in \mathbb{R}^d$ *and*

$$\begin{aligned} &\beta \sim N(0, \sigma^2_{prior} I_d) \\ &Y_i | X_i, \beta \sim Bern\left(\left(1 + \exp\left(-\beta^T X_i\right)\right)^{-1}\right). \end{aligned}$$

In this case, the posterior density is the target density. It was shown in [119, Theorem 3] that for any proposal variance parameter h*, the RWM algorithm is* (ρ, M)*-geometrically ergodic for some unknown* (ρ, M)*. By Proposition 4, the convergence rate for the RWM algorithm satisfies*

$$(h/\sigma^2_{prior} + 1)^{-d/2} \geq 1 - \rho.$$

We must then choose $h \propto \sigma^2_{prior}/d$ *or the convergence rate can tend to* 1 *as the dimension increases.*

In [10, Theroem 3.3, Section 3.2], the mixing time was shown to be exponential for a standard Gaussian target distribution if h does not scale like $1/d$. Using Proposition 4 and the results of Section 3.2, we can show an exponential mixing time for more general log-concave target distributions such as Example 1 as well as exponential mixing times in certain Wasserstein distances.

3.3.1 Application: Bayesian logistic regression with Zellner's g-prior

We look at Bayesian logistic regression with Zellner's g-prior [131] where the data-generating mechanism need not be correct. Let $(Y_i, X_i)_{i=1}^n$ with Y_i taking values in $\{0,1\}$ and X_i taking values in $\mathbb{R}^d$. Let $g \in (0,\infty)$ be a fixed constant and $X = (X_1,\ldots,X_n)^T$. Let $\sigma(u) = (1+\exp(-u))^{-1}$ denote the sigmoid function which will model the probability for the response. If the values of X^TX are positive-definite, consider the random Bayesian logistic regression density with Zellner's g-prior [131]

$$\pi_n^*(\beta) \propto \prod_{i=1}^n \sigma\left(\beta^T X_i\right)^{Y_i} \left(1-\sigma\left(\beta^T X_i\right)\right)^{1-Y_i} \exp\left(-\frac{1}{2g}\beta^T X^T X \beta\right)$$

Let β_n^* denote the random posterior maximum and the target density for RWM in this example is π_n^*. Using Proposition 4, we have the following.

Corollary 3. *Assume $(X_{i,j})_{i,j}$ are i.i.d. random variables with zero mean, unit variance, and a finite fourth moment. If $n \to \infty$ in such a way that $d_n/n \to \gamma \in (0,1)$, then with probability 1, for all sufficiently large n, the acceptance probability for RWM satisfies*

$$A(\beta_n^*) \leq \left(\frac{hn(1-\sqrt{\gamma})^2}{2g}+1\right)^{-d_n/2}.$$

Proof. We will show with probability 1, for sufficiently large n, the target density is strongly convex to apply Proposition 4. With f_n defined by

$$f_n(\beta) = \frac{1}{n}\sum_{i=1}^n \left[\log\left(1+\exp\left(\beta^T X_i\right)\right) - Y_i \beta^T X_i\right] + \frac{1}{2gn}\beta^T X^T X \beta,$$

write the posterior density $\pi_n^* \propto \exp(-nf_n)$. By [4, Theorem 2], with probability 1,

with $d_n/n \to \gamma \in (0,1)$, then

$$\lambda_{min}\left(\frac{1}{n}X^TX\right) \geq \frac{1}{2}(1-\sqrt{\gamma})^2$$

for all n sufficiently large. Let $\nabla^2 f_n$ denote the Hessian of f_n. We have for every $v, u \in \mathbb{R}^d$,

$$u^T\nabla^2 f_n(v)u \geq \frac{1}{gn}u^TX^TXu \geq \frac{(1-\sqrt{\gamma})^2}{2g}\|u\|_2^2.$$

By [80, Theorem 2.1.11], then $nf_n(\cdot)$ is strongly convex with convexity parameter $\frac{n(1-\sqrt{\gamma})^2}{2g}$ on $\mathbb{R}^d$ with probability 1. □

Under Corollary 3, with probability 1, for sufficiently large n, then if the RWM algorithm is (ρ_n, M_n)-geometrically ergodic for each n,

$$\left(\frac{hn(1-\sqrt{\gamma})^2}{2g}+1\right)^{-d_n/2} \geq 1-\rho_n.$$

and if $\beta_n^* \in K$, the (e^{-1}, K)-mixing time satisfies

$$\tau_{e^{-1},K} \geq \left(\frac{hn(1-\sqrt{\gamma})^2}{2g}+1\right)^{d_n/2} - 1.$$

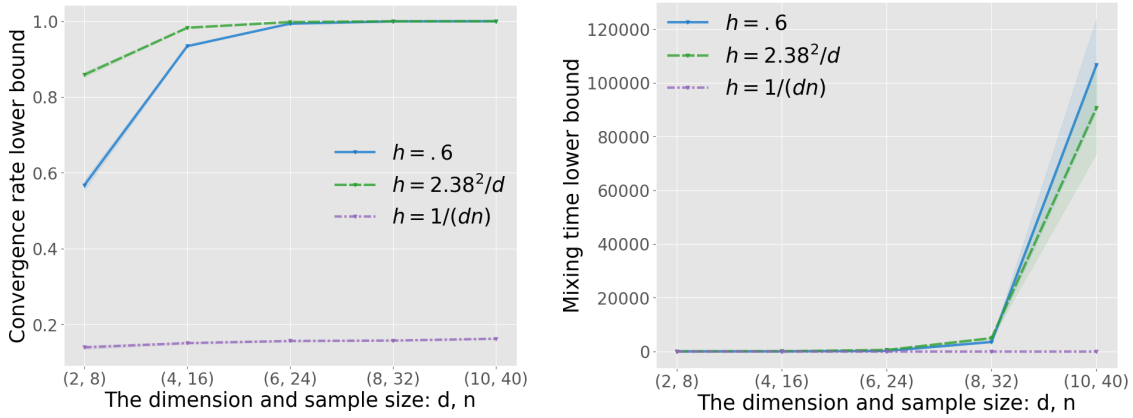
We must then choose $h \propto 2g/((1-\sqrt{\gamma})^2 nd)$ or the convergence rate can tend to 1 with d_n.

Let us now empirically investigate the convergence of the RWM algorithm in this example for different tuning parameters. We use a standard Monte Carlo estimate using 10^3 samples to the acceptance probability at the posterior maximum β_n^*. Assuming this RWM algorithm is geometrically ergodic and using Lemma 1, we will estimate the lower bound to the convergence rate. Similarly, we will use Lemma 3 to estimate a lower bound to the (e^{-1}, K)-mixing time. We analyze artificial data $(Y_i, X_i)_i$ where $X_i \sim \text{Unif}(-2, 2)$ and choose $g = 10$. We generate this data with

increasing dimensions d and sample sizes $n = 4d$, specifically,

$$(d, n) \in \{(2, 8), (4, 16), (4, 24), (8, 32), (10, 40)\}.$$

As a guideline, Corollary 3 tells us we should look to choose $h \leq \frac{2 \cdot 10}{nd(1-\sqrt{4^{-1}})^2} = 20/d^2$ at least when n is large enough. This example does not satisfy the required theoretical assumptions for optimal scaling guidelines [95] and we anticipate this choice to perform poorly here. We compare optimal scaling with $h = 2.38^2/d$, a fixed variance parameter $h = .6$, and scaling with $h = 1/(dn)$ according to Corollary 3. Repeating the simulation 50 times with randomly generated data, Figure 3.1 shows our estimates of these lower bounds using the average within 1 standard error. According to our theory, optimal scaling and fixed parameter choices should behave *poorly* as the dimension increases which corresponds with the simulation results shown in Figure 3.1.



(a) (b)

Figure 3.1: Lower bounds to the rate of convergence and (e^{-1}, K)-mixing time of the RWM algorithm for the logistic regression model with Zellner's g-prior. The shaded regions represent 1 standard deviation from the average after repeated simulations.

3.4 Lower bounds under concentration

In Example 3.3.1, we looked at the lower bounds on RWM under the concentration of a certain strongly log-concave posterior. If the target density is concentrating to its maximal point with n, then it seems reasonable the tuning parameters of the Metropolis-Hastings algorithm should also depend on the parameter n. In infinitely unbalanced Bayesian logistic regression, proposals which depend on the sample size have been shown to exhibit more appealing convergence complexity when compared to data-augmented Gibbs samplers [55]. Our goal is to formalize this concept in more generality with the lower bounds in this section.

We now look at more general target distributions indexed by a parameter $n \in \mathbb{Z}_+$ such as Bayesian posteriors and generalized Bayesian posteriors [78] where n is the sample size of the data. Let $f_n : \mathbb{R}^d \to \mathbb{R}$ and $Z_n = \int_{\mathbb{R}^d} \exp(-nf_n(\theta))d\theta$. Define the target density by $\pi_n^*(\theta) = Z_n^{-1} \exp(-nf_n(\theta))$.

We will define a general Gaussian proposal for Metropolis-Hastings that in special cases is the MALA and RWM proposal among others. Let $\gamma_{m,V}$ denote the Gaussian density with mean $m \in \mathbb{R}^d$ and positive-definite, symmetric covariance matrix $V \in \mathbb{R}^{d\times d}$. The proposal requires tuning parameters $h \in (0, \infty)$ and a positive-definite, symmetric covariance matrix $C \in \mathbb{R}^{d\times d}$. Define the function $g : \mathbb{R}^d \to \mathbb{R}^d$ depending on h and C (e.g. $g \equiv g(\cdot|h, C)$. We then define the Metropolis-Hastings kernel $P_{h,C,g}$ using the proposal density defined by $\theta \mapsto \gamma_{\theta-g(\theta),hC}$ and target density π_n^*. For specific choices of g and C, this is the proposal used in many popular Metropolis-Hastings algorithms.

Example 2. *(RWM proposals) When the covariance $C = I$ and $g \equiv 0$, then this is the proposal used in the RWM algorithm [113]. For more general covariance C, the proposal can be viewed as a Riemannian manifold RWM algorithm (see [41] for Riemannian manifold MALA).*

When f_n is differentiable, we can define proposals motivated through discretizing the continuous-time Langevin dynamics [98].

Example 3. *(MALA proposals) When the covariance $C = I$ and*

$$g(\theta) = h\nabla n f_n(\theta)/2$$

then this is the proposal used in MALA [98]. For more general covariance C, the function $g(\theta) = hC\nabla n f_n(\theta)/2$ defines the proposal used in Riemannian manifold MALA [41].

We will make the following assumption on the target density π_n^* throughout this section:

Assumption 3. *There exists at least one $\theta_n^* \in supp(\pi_n^*)$ which maximizes π_n^* for each n.*

The following result lower bounds the convergence rate independently of the function g. In particular, this holds for both the RWM and the MALA algorithms when g, h, and C are chosen appropriately.

Proposition 5. *The acceptance probability $A_{g,h,C}(\cdot)$ for the kernel $P_{g,h,C}$ satisfies*

$$A_{g,h,C}(\theta_n^*) \leq \frac{1}{\pi_n^*(\theta_n^*)(2\pi h)^{d/2}\det(C)^{1/2}}.$$

Proof. For any $\theta \in \mathbb{R}^d$, we have

$$\gamma_{\theta - g(\theta), hC}(\theta^*) \leq (2\pi h)^{-d/2}\det(C)^{-1/2}.$$

This implies that

$$\begin{aligned}
&A_{h,C,g}(\theta_n^*) \\
&= \int_{\mathrm{supp}(\pi_n^*)} \min\left\{ \frac{\gamma_{\theta_n^* - g(\theta_n^*), hC}(\theta)}{\pi_n^*(\theta)}, \frac{\gamma_{\theta - g(\theta), hC}(\theta_n^*)}{\pi_n^*(\theta_n^*)} \right\} \pi_n^*(\theta) d\theta \\
&\leq \int_{\mathrm{supp}(\pi_n^*)} \frac{\gamma_{\theta - g(\theta), hC}(\theta_n^*)}{\pi_n^*(\theta_n^*)} \pi_n^*(\theta) d\theta \\
&\leq \frac{1}{\pi_n^*(\theta_n^*)(2\pi h)^{d/2} \det(C)^{1/2}}. \qquad (3.4)
\end{aligned}$$

□

Propositon 5 says if Metropolis-Hastings with kernel $P_{h,C,g}$ is (ρ_n, M_n)-geometrically ergodic for each n, then

$$\frac{1}{\pi_n^*(\theta_n^*)(2\pi h)^{d/2} \det(C)^{1/2}} \geq 1 - \rho_n$$

and if $\theta_n^* \in K$, the (e^{-1}, K)-mixing time satisfies

$$\tau_{e^{-1},K} \geq \pi_n^*(\theta_n^*)(2\pi h)^{d/2} \det(C)^{1/2} - 1.$$

In particular, if $h \geq \frac{1}{\pi \det(C)^{1/d} \pi_n^*(\theta_n^*)^{2/d}}$,

$$2^{-d/2} \geq 1 - \rho \qquad \text{and} \qquad \tau_{e^{-1},K} \geq 2^{d/2} - 1.$$

Similarly, we could apply our results in Wasserstein distances. The bounds from Proposition 5 are quite "loose" compared to looking at the acceptance probability using the previous section. However, Proposition 5 says we can study "global" convergence properties of the Metropolis-Hastings Markov chain by only looking "locally" at the concentration of $\pi_n^*(\theta_n^*)$.

Even though the bounds in Proposition 5 are crude, this puts stringent restrictions

on the magnitude of the tuning parameter h for RWM and MALA type algorithms. Exploring complex, multi-modal target densities would likely require the tuning parameter h to be large enough to move between the various modes of the target density. However, Proposition 5 then forces h to be small in order to be computationally efficient.

We use a Laplace approximation to lower bound the target density at its maximum point. The following result is inspired by previous results on Laplace approximations [62, 107, 112, 115, 116]. Compared to high-dimensional Laplace approximations [107, 112], we do not require f_n to be differentiable. The merit of Proposition 6 will be in studying Metropolis-Hastings algorithms in combination with Proposition 5.

Proposition 6. *Assume for some $\kappa \in (0,1)$, the dimension $d \equiv d_n$ satisfies $d_n \leq n^\kappa$. Assume the following holds for some constants $\delta_0, \lambda_0, f^*, I_0 \in (0,\infty)$ and for all sufficiently large (n, d_n):*

1. f_n is λ_0^{-1}-strongly convex for all $\|\theta - \theta_n^*\|_2 \leq \delta_0$.

2. The optimal point satisfies the strict optimality condition:

$$\inf_{\|v\|>\delta_0} f_n(\theta_n^* + v) \geq f_n(\theta_n^*) + f^*$$

3. The integral is controlled away from the optimum:

$$\int_{\|v\|_2>\delta_0} \exp\left(-\left(f_n(\theta_n^* + v) - f_n(\theta_n^*)\right)\right) dv \leq (I_0 n)^{d_n}.$$

Then for any $c \in (0,1]$, for all sufficiently large (n, d_n), the density π_n^* concentrates at θ_n^* with

$$\pi_n^*(\theta_n^*) \geq \frac{1}{1+c}\left(\frac{n}{2\pi\lambda_0}\right)^{d_n/2}.$$

Proof. Take n sufficiently large so that we may use each of the assumptions. After

changing the variables, we have the decomposition

$$\begin{aligned}\frac{1}{\pi^*(\theta_n^*)} &= \int_{\mathbb{R}^d} \exp(-n(f_n(\theta) - f_n(\theta_n^*)))d\theta \\ &= \frac{1}{n^{d/2}} \int_{\left\|un^{-1/2}\right\|_2 \leq \delta_0} \exp(-n(f_n(\theta_n^* + un^{-1/2}) - f_n(\theta_n^*)))du \qquad (3.5) \\ &\quad + \int_{\|v\|_2 > \delta_0} \exp(-n(f_n(\theta_n^* + v) - f_n(\theta_n^*)))dv. \qquad (3.6)\end{aligned}$$

We control the first integral (3.5). Since the closed ball $\overline{B_{\delta_0}(\theta_n^*)}$ is convex, for all $\left\|un^{-1/2}\right\|_2 \leq \delta_0$, by the subgradient inequality [80, Lemma 3.2.3],

$$nf_n(\theta_n^* + un^{-1/2}) - nf_n(\theta_n^*) \geq \frac{1}{2\lambda_0} \|u\|_2^2 .$$

This implies

$$\begin{aligned}&\frac{1}{n^{d_n/2}} \int_{\left\|n^{-1/2}u\right\|_2 \leq \delta_0} \exp(nf_n(\theta_n^*) - nf_n(\theta_n^* + un^{-1/2}))du \\ &\leq \frac{1}{n^{d_n/2}} \int_{\left\|un^{-1/2}\right\|_2 \leq \delta_0} \exp\left(-\frac{1}{2\lambda_0} \|u\|_2^2\right) du \\ &\leq \left(\frac{2\pi\lambda_0}{n}\right)^{d_n/2} .\end{aligned}$$

We now control the second integral (3.6). We have

$$\begin{aligned}&\int_{\|v\|_2 > \delta_0} \exp\left(-n\left(f_n(\theta_n^* + v) - f_n(\theta_n^*)\right)\right) dv \\ &= \int_{\|v\|_2 > \delta_0} \exp\left(-(n-1)\left(f_n(\theta_n^* + v) - f_n(\theta_n^*)\right) - \left(f_n(\theta_n^* + v) - f_n(\theta_n^*)\right)\right) dv \\ &\leq e^{-(n-1)f^*} \int_{\|v\|_2 > \delta_0} \exp\left(-\left(f_n(\theta_n^* + v) - f_n(\theta_n^*)\right)\right) dv \\ &\leq e^{-(n-1)f^*} (I_0 n)^{d_n}.\end{aligned}$$

Combining these results,

$$\begin{aligned}&\int_{\mathbb{R}^d} \exp(-n(f_n(\theta) - f_n(\theta_n^*)))d\theta \\ &\leq \left(\frac{2\pi\lambda_0}{n}\right)^{d_n/2} \left[1 + \left(\frac{n}{2\pi\lambda_0}\right)^{d_n/2} e^{-(n-1)f^*}(I_0 n)^{d_n}\right].\end{aligned}$$

Since $d_n \leq n^\kappa$,

$$\limsup_{n\to\infty} \left(\frac{n}{2\pi\lambda_0}\right)^{d_n/2} e^{-(n-1)f^*}(I_0 n)^{d_n} = 0.$$

The desired result follows at once. □

In Proposition 6, we require the dimension does not grow too fast with n. The first assumption is a locally strong convex assumption which ensures sufficient curvature of f_n locally near the maximum point of f_n. Since we only require a lower bound on the density, we avoid the need to control higher order derivatives used in high-dimensional Laplace approximations [107, 112]. The second and third assumptions ensure sufficient decay of f_n away from θ_n^* similar to assumptions made previously [62, 112]. Similar assumptions are also used for generalized Bayesian posterior densities with proper priors when the dimension is fixed [78, Theorem 4].

Combining Proposition 5 with Proposition 6 immediately yields the following.

Corollary 4. *Under the conditions of Proposition 6, for any $c \in (0, 1]$ and for all sufficiently large (n, d_n), the acceptance probability $A_{g,h,C}(\cdot)$ for the kernel $P_{g,h,C}$ satisfies*

$$A_{g,h,C}(\theta_n^*) \leq \left(\frac{\lambda_0}{nh}\right)^{d_n/2} \frac{1+c}{\det(C)^{1/2}}.$$

Corollary 4 can be particularly important when tuning Metropolis-Hastings algorithms used in Bayesian statistics. Under the conditions of Corollary 4, if Metropolis-

Hastings with kernel $P_{h,C,g}$ is (ρ_n, M_n)-geometrically ergodic for each n,

$$\left(\frac{\lambda_0}{nh}\right)^{d_n/2} \frac{1+c}{\det(C)^{1/2}} \geq 1 - \rho_n$$

and similarly as before, we have bounds for the mixing time and Wasserstein distances. In particular, if the target density is concentrating at its maximum point and the tuning parameters h and C do not depend carefully on n, then Corollary 4 says it can happen that $\lim_{(n,d_n)\to\infty} \rho_n = 1$ rapidly.

3.4.1 Application: generalized flat prior Bayesian logistic regression

We will consider flat prior Bayesian logistic regression without assuming the correctness of the data generation. Let $(Y_i, X_i)_{i=1}^n$ be i.i.d. with Y_i taking values in $\{0, 1\}$ and X_i taking values in $\mathbb{R}^d$. With the sigmoid function σ, consider the following random Bayesian logistic regression density

$$\pi_n^*(\beta) \propto \prod_{i=1}^n \sigma\left(\beta^T X_i\right)^{Y_i} \left(1 - \sigma\left(\beta^T X_i\right)\right)^{1-Y_i}$$

Conditions for the propriety of the posterior and also the maximum likelihood estimate have been previously investigated [17].

Assumption 4. *Let $z_i = 1$ if $Y_i = 0$ and $z_i = -1$ if $Y_i = 1$ and define X^* to be the matrix with rows $z_i X_i^T$. Suppose:*

1. *$X = (X_1, \ldots, X_n)^T$ is full column rank.*
2. *There exists a vector $a \in \mathbb{R}^n$ with all components positive such that ${X^*}^T a = 0$.*

If Assumption 4 holds with probability 1 for all sufficiently large n then under these conditions, π_n^* is a random probability density and the MLE β_n^* exists ([17, Theorem

2.1, 3.1]). The Pólya-Gamma Gibbs sampler is geometrically ergodic for this model [125], but to the best of our knowledge, it remains an open question if Metropolis-Hastings is geometrically ergodic. Since the prior is not proper here, we cannot look to apply previous results on generalized posterior concentration [78]. Under certain conditions, we show this generalized posterior density can indeed concentrate so that Corollary 4 holds.

Theorem 5. *Assume the following:*

1. *With probability* 1*, Assumption 4 holds for all sufficiently large* n*.*
2. *The MLE* β_n^* *is almost surely consistent to some* $\beta_0 \in \mathbb{R}^d$*.*
3. $\|X_1\|_2 \leq 1$ *with probability* 1*.*
4. *For* $u \in \mathbb{R}^d$*, if* $u \neq 0$*, then* $X_1^T u \neq 0$ *with probability* 1*.*

Then with probability 1*, for all sufficiently large* n*, there is a* $\lambda_0 \in (0, \infty)$ *so that the acceptance probability* $A_{g,h,C}(\cdot)$ *for the kernel* $P_{g,h,C}$ *satisfies*

$$A_{g,h,C}(\beta_n^*) \leq \left(\frac{\lambda_0}{nh}\right)^{d/2} \frac{2}{\det(C)^{1/2}}.$$

Proof. We will show the conditions of Proposition 6 hold with probability 1 for large enough n. Using Assumption 1, with probability 1, we will assume n is sufficiently large so that the posterior density π_n^* and the MLE β_n^* exist. Define the Bayesian logistic regression loss function

$$f_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left[\log\left(1 + \exp\left(\beta^T X_i\right)\right) - Y_i \beta^T X_i\right]$$

and write $\pi_n^* = Z_n^{-1} \exp(-n f_n)$ where $Z_n = \int \exp(-n f_n(\theta)) d\theta$. We first develop sufficient curvature of the target density at β_n^*. Denote the k-th derivative matrix or

tensor of the function f_n by $\nabla^k f_n$. For every $v \in \mathbb{R}^d$, we have

$$v^T \nabla^2 f_n(\beta_n^*) v = \frac{1}{n} \sum_{i=1}^n v^T X_i X_i^T v \sigma\left(X_i^T \beta_n^*\right)\left(1 - \sigma\left(X_i^T \beta_n^*\right)\right)$$

and

$$\nabla^3 f_n(\beta_n^*)(v,v,v) = \frac{1}{n} \sum_{i=1}^n (X_i^T v)^3 \sigma\left(X_i^T \beta_n^*\right)\left(1 - \sigma\left(X_i^T \beta_n^*\right)\right)\left(1 - 2\sigma\left(X_i^T \beta_n^*\right)\right).$$

Since $\|X_i\|_2 \leq 1$ with probability 1, by the strong law of large numbers [34, Theorem 10.13], almost surely,

$$\begin{aligned}\nabla^2 f_n(\beta_0) &= \frac{1}{n} \sum_{i=1}^n X_i X_i^T \sigma\left(X_i^T \beta_0\right)\left(1 - \sigma\left(X_i^T \beta_0\right)\right) \\ &\to \mathbb{E}\left(X_1 X_1^T \sigma\left(X_1^T \beta_0\right)\left(1 - \sigma\left(X_1^T \beta_0\right)\right)\right).\end{aligned}$$

By Assumption 4, for any $u \in \mathbb{R}^d$, $u \neq 0$, with probability 1,

$$u^T X_1 X_1^T u \sigma\left(X_1^T \beta_0\right)\left(1 - \sigma\left(X_1^T \beta_0\right)\right) > 0.$$

Since expectations preserve strict inequalities, we can find a sufficiently small $\epsilon_0 \in (0,1)$ so that for any $u \in \mathbb{R}^d$, $u \neq 0$,

$$\left(\frac{u}{\|u\|_2}\right)^T \mathbb{E}\left(X_1 X_1^T \sigma\left(X_1^T \beta_0\right)\left(1 - \sigma\left(X_1^T \beta_0\right)\right)\right)\left(\frac{u}{\|u\|_2}\right) \geq \epsilon_0.$$

Combining these results, for all $u \in \mathbb{R}^d$, with probability 1,

$$u^T \nabla^2 f_n(\beta_0) u \geq \frac{\epsilon_0}{2} \|u\|_2^2$$

for all sufficiently large n. By Assumption 3, $\|X_i\|_2 \leq 1$ with probability 1 and for

$x \in (0,1)$, $|x(1-x)(1-2x)| \leq 1/4$, so by the mean value theorem, almost surely,

$$\left\|\nabla^2 f_n(\beta_n^*) - \nabla^2 f_n(\beta_0)\right\|_2 \leq 4^{-1} \left\|\beta_n^* - \beta_0\right\|_2 \to 0.$$

For all $u \in \mathbb{R}^d$, with probability 1,

$$\begin{aligned} u^T \nabla^2 f_n(\beta_n^*) u &\geq u^T \nabla^2 f_n(\beta_0) u - \frac{\epsilon_0}{4} \|u\|_2^2 \\ &\geq \frac{\epsilon_0}{4} \|u\|_2^2 \end{aligned} \tag{3.7}$$

for all n sufficiently large.

For the remainder of this argument, we will assume n is sufficiently large so that (3.7) holds with probability 1, and the remainder of the proof is taken to hold with probability 1 without reference. Since $\|X_i\|_2 \leq 1$ and as in [3, Section 3], for all $u \in \mathbb{R}^d$ and $t \in (0, \infty)$,

$$|\nabla^3 f_n(\beta_n^* + tu)(u,u,u)| \leq \|u\|_2 \, u^T \nabla^2 f_n(\beta_n^* + tu) u$$

and we can apply the self-concordant analysis of [3, Proposition 1]. Using [3, Proposition 1 (6)], for $\|v\|_2 \leq 2$ and all $u, v \in \mathbb{R}^d$,

$$u^T \nabla^2 f_n(\beta_n^* + v) u \geq u^T \nabla^2 f_n(\beta_n^*) u \exp(-\|v\|_2) \geq \frac{\epsilon_0}{4} \exp(-2) \|u\|_2^2 .$$

Since closed balls are convex, then by [80, Theorem 2.1.11], f_n is strongly convex on the closed ball $\overline{B_2(\beta_n^*)}$. Thus, we have satisfied the local strong convexity condition (1) in Proposition 6. Using [3, Proposition 1 (3)], since $e^{-x} \geq 1 - x$, and $\nabla f_n(\beta_n^*) \equiv 0$,

then for all $v \in \mathbb{R}^d$ with $v \neq 0$,

$$\begin{aligned} f_n(\beta_n^* + v) - f_n(\beta_n^*) &\geq \frac{v^T \nabla^2 f_n(\beta_n^*) v}{\|v\|_2^2} \left(\exp(-\|v\|_2) + \|v\|_2 - 1\right) \\ &\geq \frac{\epsilon_0}{4} \left(\exp(-\|v\|_2) + \|v\|_2 - 1\right) \\ &\geq \frac{\epsilon_0}{4} \left(\|v\|_2 - 1\right). \end{aligned}$$

Thus, for $\|v\|_2 > 2$, we have satisfied the strict optimality condition (2) in Proposition 6. We also have the required control of the integral (3) in Proposition 6:

$$\begin{aligned} &\int_{\|v\|_2 > 2} \exp\left(-\left(f_n(\beta_n^* + v) - f_n(\beta_n^*)\right)\right) dv \\ &\leq \exp(\frac{\epsilon_0}{4}) \int_{\|v\|_2 > 2} \exp\left(-\frac{\epsilon_0}{4} \|v\|_2\right) dv < \infty. \end{aligned}$$

□

Assumption 1 ensures existence of the posterior density and MLE using [17, Theorem 2.1, 3.1]. Consistency of the MLE in Assumption 2 is a well-studied problem and conditions are available when the model is correctly specified [31] or using M-estimation [118, Example 5.40]. Assumption 4 was used previously in [78, Theorem 13] and is used to ensure identifiability in GLM models [118]. Assumption 2 requires standardization of the features which is often done in practice for numerical stability. Under the conditions of Theorem 5, with probability 1, if Metropolis-Hastings with kernel $P_{g,C,h}$ is (ρ_n, M_n)-geometrically ergodic for each n, then

$$\left(\frac{\lambda_0}{nh}\right)^{d/2} \frac{2}{\det(C)^{1/2}} \geq 1 - \rho_n$$

for sufficiently large n. Similarly, we have results for the mixing time and Wasserstein distances.

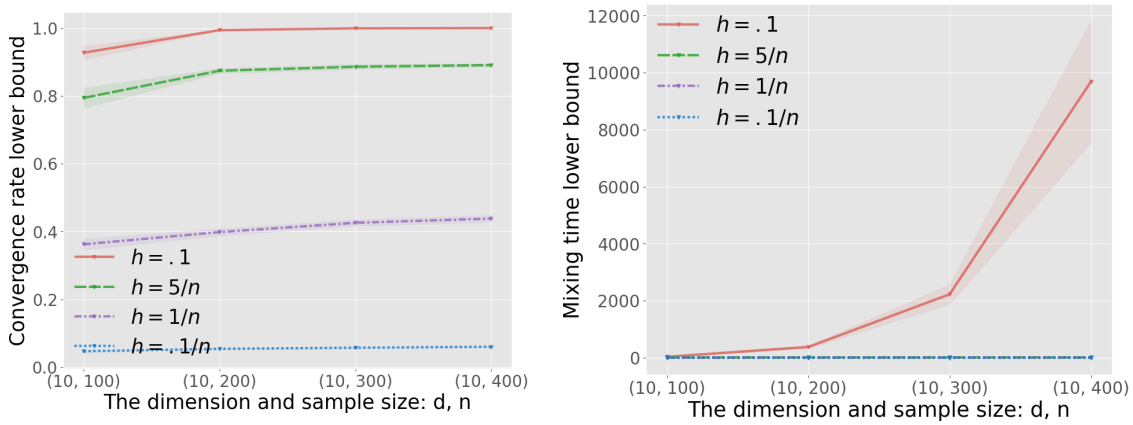
In this example, we do not have an explicit value for λ_0, so we will investigate the robustness of scaling with the sample size n empirically. We use a standard Monte Carlo estimate using 10^3 samples to the acceptance probability in the RWM algorithm at the MLE β_n^*. We compare a fixed variance parameter $h = .1$, and scaling with $h = 5/n$, $h = 1/n$, and $h = .1/n$. As before, we use Lemma 3 to estimate a lower bound to the (e^{-1}, K)-mixing time and if this RWM algorithm is geometrically ergodic, then using Lemma 1, we will estimate the lower bound to the convergence rate. We analyze artificial data $(Y_i, X_i)_i$ where $X_i \sim \text{Unif}(-2, 2)$ and generate this data in fixed dimension $d = 10$ and increasing sample sizes $n \in \{100, 200, 300, 400\}$. We repeat the simulation 50 times with randomly generated data and Figure 3.2 shows our estimates to these lower bounds using the average within 1 estimated standard error. We can see that $h = 5/n$ scales worse than $h = 1/n, .1/n$, but the scaling is not nearly as problematic as the fixed variance parameter which is commonly used in practice.

If the RWM is not geometrically ergodic, our simulation results on the convergence rate can be modified accordingly. We can replace the flat prior with a uniform prior over a large compact set, so then RWM is geometrically ergodic by a global minorization condition. We can then expect a similar simulation result if the compact set is chosen large enough.

3.5 Comparison with conductance methods

In this section, we show how the conductance can be used to lower bound the convergence rate and compare it to the techniques used in this chapter. If a Markov kernel P satisfies

$$\sup_{\substack{f \in L^2(\Pi^*) \\ \left\|f - \int f d\Pi^*\right\|_{L^2(\Pi^*)} \neq 0}} \frac{\left\|Pf - \int f d\Pi^*\right\|_{L^2(\Pi^*)}}{\left\|f - \int f d\Pi^*\right\|_{L^2(\Pi^*)}} = \beta \in (0, 1)$$



(a) (b)

Figure 3.2: Lower bounds to the rate of convergence and (e^{-1}, K)-mixing time of the RWM algorithm for flat prior logistic regression. The shaded regions represent 1 standard deviation from the average after repeated simulations.

then it is said to have a $1-\beta$ spectral gap. For sets B with $0 < \Pi^*(B) < 1$, define

$$k_P(B) = \frac{\int_B P(\cdot, B^c) d\Pi^*}{\Pi^*(B)\Pi^*(B^c)}$$

and define the conductance [69] $k_P = \inf\{k_P(B) : 0 < \Pi^*(B) < 1\}$. The conductance can be used to upper bound the spectral gap [69, Theorem 2.1]:

$$1 - \beta \leq k_P.$$

If P is a reversible Markov kernel,[35, Theorem 2.1] gives an equivalence between a spectral gap and geometric convergence in total variation. Specifically, there is a $\rho \in (0,1)$ such that for every probability measure μ with $d\mu/d\Pi^* \in L^2(\Pi^*)$, there is a constant $C_\mu \in (0, \infty)$ such that

$$\left\|\mu P^t - \Pi^*\right\|_{\mathrm{TV}} \leq C_\mu \rho^t$$

if and only if there is a spectral gap with $\beta \leq \rho$. Under these conditions, the convergence rate can be lower bounded by the conductance so that

$$1 - \rho \leq k_P.$$

Under further conditions on the function M, the conductance will also lower bound the convergence rate if P is $(\rho, M)-$geometrically ergodic [35, Proposition 2.1 (ii), Theorem 2.1].

As described in [46], for sets B with $0 < \Pi^*(B) \leq 1/2$, to move from B to B^c, the Metropolis-Hastings kernel P must accept and [46, Proposition 2.16] gives

$$k_P \leq 2 \sup_{x \in B} A(x).$$

For Metropolis-Hastings with target distributions having a Lebesgue density on $\mathbb{R}^d$, if the acceptance is upper semicontinuous, then we show the following.

Proposition 7. *Let P be a Metropolis-Hastings kernel with an upper semicontinuous acceptance probability $A(\cdot)$. Then for any $\theta \in supp(\pi^*)$,*

$$k_P \leq A(\theta).$$

Proof. Fix $\theta_0 \in \text{supp}(\pi^*)$ and fix $r \in (0, \infty)$ and let $B_r = \{x \in \mathbb{R}^d : \|x - \theta_0\|_2 \leq r\}$ which is the closed ball of radius r around θ_0. Since Π^* has Lebesgue density on $\mathbb{R}^d$, $\lim_{r \downarrow 0} \Pi^*(B_r) = 0$. Now we can choose r small enough so that $\Pi^*(B_r) < 1$. Since $\theta_0 \in \text{supp}(\pi^*)$, we have $\Pi^*(B_r) > 0$. For $x \in B_r$, $\delta_x(B_r^c) = 0$ and so for all small

enough r,

$$k_P \leq \frac{\int_{B_r} P(\cdot, B_r^c) d\Pi^*}{\Pi^*(B_r)\Pi^*(B_r^c)} = \frac{\int_{B_r}\int_{B_r^c} a(\cdot,\cdot) dQ(\cdot,\cdot) d\Pi^*}{\Pi^*(B_r)\Pi^*(B_r^c)} \leq \frac{\int_{B_r} A(\cdot) d\Pi^*}{\Pi^*(B_r)\Pi^*(B_r^c)} \leq \frac{\sup_{x\in B_r} A(x)}{\Pi^*(B_r^c)}.$$

Since this holds for all small r, taking the limit $k_P \leq \lim_{r\downarrow 0} \sup_{x\in B_r} A(x)$. Since $A(\cdot)$ is upper semicontinuous, we can choose $(x_r^*) \subseteq B_r$ such that for each r,

$$\sup_{x\in B_r} A(x) = A(x_r^*).$$

Since $\|x_r^* - \theta_0\|_2 \leq r$, then $\lim_{r\downarrow 0} x_r^* = \theta_0$. By upper semicontinuity,

$$A(\theta_0) \geq \lim_{r\downarrow 0} A(x_r^*) \geq \lim_{r\downarrow 0} \sup_{x\in B_r} A(x).$$

□

Therefore, under stronger conditions on M and the acceptance probability, we can combine [35, Proposition 2.1 (ii), Theorem 2.1] and Proposition 7 to show Lemma 2. If the acceptance $A(\cdot)$ is upper semicontinuous, we can also obtain a comparable result to Lemma 3 using the conductance. From [10, Theorem 2.34], for every $\Pi^*(A) \in (0, 1/2]$, the (ϵ, A)-mixing time for the Markov kernel P satisfies

$$\tau_{1/4,A} \geq \frac{1}{4k_P(A)}.$$

If the acceptance is upper semicontinuous, we can use Proposition 7 to show that also

$$\tau_{1/4,A} \geq \frac{1}{4A(\theta)}.$$

In our analysis, we do not make the additional assumptions used in this section for lower bounds with respect to total variation. It is an interesting open question if stronger assumptions combined with the conductance can be used to develop similar Wasserstein lower bounds as in this section since this appears previously unstudied.

3.6 Discussion and further research directions

We hope the lower bounds studied in this chapter can further guide practitioners in tuning some Metropolis-Hastings algorithms. We showed the computational complexity of geometrically ergodic Metropolis-Hastings algorithms can grow rapidly in d and n if the tuning parameters do not carefully take into account the parameter n. We showed RWM and MALA type algorithms have strong restrictions on the size of the variance parameter h when the target density concentrates. Small values of h dissuade the algorithm from exploring the target distribution and large values cause the convergence rate to tend to 1 exponentially fast. We showed empirically in Bayesian logistic regression that scaling the variance of the proposal in the RWM algorithm with the sample size is crucial in avoiding poor performance when the target distribution concentrates. We also showed the lower bounds in many Wasserstein distances are similar to that of total variation and we can expect similar behavior here. A future research direction is to analyze the lower bounds for other algorithms such as Hamiltonian Monte Carlo as well as many other adaptive algorithms.

3.7 Supplementary material and code availability

The code used for the simulations and plots is made available for download at `https://github.com/austindavidbrown/lower-bounds-for-Metropolis-Hastings`.

Chapter 4

Geometric Ergodicity of a Gibbs Sampler for Bayesian Error-in-Variable Regression

4.1 Overview

Bayesian approaches develop a strategy for additional error in the variables by constructing a new model incorporating additional error. We consider Bayesian EIV linear regression [16, 23, 32, 51, 73, 79, 93, 101, 117, 120] accounting for additive Gaussian error in the features (covariates) and response. We assume the variability of the additive Gaussian error is known beforehand which arises often in astrophysics in the presence of known instrumentation error [48, 63]. Alternative approaches to EIV models attempt to correct existing parameter estimation methods such as least squares or method of moments with weighting and other techniques [36, 108]. Several other strategies for EIV modeling are discussed in more comprehensive treatments on the topic [13, 14, 36].

We write $x \sim N_d(m, C)$ to mean the d-dimensional normal distribution with mean m and symmetric, positive-definite (SPD) covariance matrix C. We also write $x \sim$ Inverse-gamma(x, y) to be the inverse-gamma distribution with shape and rate

parameters $x, y \in (0, \infty)$. Let $(Y_i, X_i, Z_i)_{i=1}^n$ be independent and identically distributed (i.i.d.) where the response Y_i is real-valued along with features X_i taking values in $\mathbb{R}^p$ and fixed, known features $Z_i \in \mathbb{R}^q$ where n, p, q positive integers. Let $\alpha \in \mathbb{R}^q$, $\beta \in \mathbb{R}^p$, and $\sigma^2 \in (0, \infty)$ be unknown regression and variance parameters respectively. We introduce new parameters $\mathcal{A} = (\mathcal{A}_1, \ldots, \mathcal{A}_n)^T$ with $\mathcal{A}_i \in \mathbb{R}^p$ to model additional error in X_i using classical or Berkson errors [7]. The classical error model assumes $X_i|\mathcal{A}_i$ and the Berkson error model [7] assumes instead a data-dependent prior $\mathcal{A}_i|X_i$. When there is additional error in X_i, the EIV linear regression model for $i \in 1, \ldots, n$ is i.i.d. with

$$Y_i|\mathcal{A}_i, \alpha, \beta, \sigma^2 \sim N_1(Z_i^T\alpha + \mathcal{A}_i^T\beta, \sigma^2) \tag{4.1a}$$

$$X_i|\mathcal{A}_i \sim N_p(\mathcal{A}_i, V_i) \text{ (Classical)} \quad \text{or} \quad \mathcal{A}_i|X_i \sim N_p(X_i, V_i) \text{ (Berkson)} \tag{4.1b}$$

where the SPD matrices $V_i \in \mathbb{R}^{p\times p}$ are known. When there is also additional error in the responses Y_i, we assume an i.i.d. hierarchical regression model with

$$Y_i|\nu_i \sim N_1(\nu_i, u_i^2) \tag{4.2a}$$

$$\nu_i|\mathcal{A}_i, \alpha, \beta, \sigma^2 \sim N_1(Z_i^T\alpha + \mathcal{A}_i^T\beta, \sigma) \tag{4.2b}$$

$$X_i|\mathcal{A}_i \sim N_p(\mathcal{A}_i, V_i) \text{ (Classical)} \quad \text{or} \quad \mathcal{A}_i|X_i \sim N_p(X_i, V_i) \text{ (Berkson)} \tag{4.2c}$$

where $u_i^2 \in (0, \infty)$ are known.

We will be interested in the posterior for both models (4.1) and (4.2) using independent normal and inverse-gamma priors on the parameters $(\mathcal{A}, \alpha, \beta, \sigma^2)$. The independent prior choice is a popular choice in Bayesian regression models with and without measurement error [14, 23, 30, 92]. For the EIV regression models (4.1) and

(4.2), the independent priors are chosen

$$\sigma^2 \sim \text{Inverse-gamma}(a_0, b_0) \tag{4.3a}$$

$$(\alpha, \beta)^T \sim N_{q+p}(j_0, J_0) \tag{4.3b}$$

where $a_0, b_0 \in (0, \infty)$, $j_0 \in \mathbb{R}^{q+p}$ and $J_0 \in \mathbb{R}^{(q+p)\times(q+p)}$ is a SPD matrix. The classical and Berkson error models assume either

$$\mathcal{A}_i \sim N_p(k_i, K_i) \text{ (Classical)} \quad \text{or} \quad \mathcal{A}_i \text{ flat prior (Berkson)} \tag{4.4}$$

where $k_i \in \mathbb{R}^p$ and $K_i \in \mathbb{R}^{p\times p}$ are SPD matrices. For example, an *exposure model* [44] utilized often in epidemiology would assume classical errors with a data-dependent prior on each $\mathcal{A}_i$ depending on Z_i. In the Berkson error model, each $\mathcal{A}_i|X_i$ is specified and it is natural to assume an improper flat prior on each $\mathcal{A}_i$.

Previous work has proposed Gibbs sampling [38] to draw samples from the posterior, denoted by Π_n, in Bayesian EIV regression models [8, 14, 23, 93]. However, trustworthy estimation from a Gibbs sampler requires the Markov chain to converge to the posterior distribution at a sufficiently fast rate. Consider a real-valued function f with $\int |f|^{2+\delta} d\Pi_n < \infty$ for some $\delta \in (0, \infty)$ and denote $\bar{f}_m$ as the time average of m samples from the Gibbs sampler. In order to be confident in the estimator $\bar{f}_m$ in applications, a standard error and confidence interval are essential. A Gibbs sampler is geometrically ergodic if initialized at points, its marginal distribution is converging to Π_n at an exponential rate in total variation. Geometrically ergodic Gibbs samplers provide rich theoretical guarantees which are of practical relevance in applications. These Gibbs samplers satisfy a central limit theorem [15, 57], that is,

$$\sqrt{m}\left(\bar{f}_m - \int f d\Pi_n\right)$$

is asymptotically normally distributed and under suitable assumptions, the variance in this normal distribution can be consistently estimated [33]. Further pertinent tools to ensuring reliable estimation such as estimates of the effective sample size, consistent confidence ellipsoids, and consistent confidence intervals for quantile estimation are also available [25, 119].

To the best of our knowledge, the rate of convergence for Gibbs sampling in EIV regression models has not been previously investigated. Related approaches have instead proposed variational Bayesian methods [11, 84] and the integrated nested Laplace approximation (INLA) [52, 79]. We construct a general density which in special cases, is the posterior for the 4 Bayesian EIV regression models (4.1) and (4.2) using the independent normal and inverse-gamma prior choice on the parameters (4.3) and (4.4). Our main contribution constructs a 3-variable deterministic scan Gibbs sampler for this general density, and we show it is *always* geometrically ergodic using a drift and minorization condition [45, 76]. The 3-variable Gibbs sampler we construct can be simulated efficiently on a computer without the need for complex Metropolis-Hastings or rejection sampling steps at each iteration.

The organization is as follows. In Section 4.2, we construct a general EIV regression density and construct a 3-variable Gibbs sampler for this density. We show the Gibbs sampler is always geometrically ergodic and apply this to the 4 Bayesian EIV regression models presented in this introduction. Finally in Section 4.3, we discuss our results and future research directions.

4.2 General Gibbs Sampler for EIV regression

For positive integers p, define p-norms by $\|\cdot\|_p$. The posteriors for the Bayesian EIV regression models (4.1) and (4.2) using independent prior choices (4.3) and (4.4) for both classical and Berkson errors share a common general form which we study in this

section. The posterior densities for Bayesian EIV regression models are special cases of the density (4.5) but will differ depending on the EIV modeling choice illustrated in the subsequent sections. For $i \in 1, \ldots, n$, define hyper-parameters $c_0 \in \mathbb{R}^{p+q}$, SPD matrices $C_0 \in \mathbb{R}^{(p+q)\times(p+q)}$, $D_i \in \mathbb{R}^{p\times p}$, and $d_i \in \mathbb{R}^p$, $r \in \mathbb{R}^n$, and $M \in \mathbb{R}^{n\times q}$. For $\mathcal{A} \in \mathbb{R}^{n\times p}$, $\mu \in \mathbb{R}^q$, $\beta \in \mathbb{R}^p$, $\sigma^2 \in (0, \infty)$, define the density

$$\pi_n(\mathcal{A}, \mu, \beta, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{n/2+a_0+1} \exp\left[-\frac{1}{\sigma^2}\left(\frac{1}{2}\left\|r - M\mu - \mathcal{A}\beta\right\|_2^2 + b_0\right)\right] \tag{4.5a}$$

$$\times \prod_{i=1}^{n} \exp\left(-\frac{1}{2}(\mathcal{A}_i - d_i)^T D_i^{-1}(\mathcal{A}_i - d_i)\right) \tag{4.5b}$$

$$\times \exp\left(-\frac{1}{2}((\mu, \beta)^T - c_0)^T C_0^{-1}((\mu, \beta)^T - c_0)\right). \tag{4.5c}$$

We will construct a 3-variable deterministic scan Gibbs sampler using a specific update order for the density (4.5). We also derive the conditional densities for the Gibbs sampler which can be sampled directly. Initialize $(\mu_0, \beta_0, \sigma_0^2)$ and $\mathcal{A}_0 = (\mathcal{A}_{1,0}, \ldots, \mathcal{A}_{n,0})$ from an initial distribution. For $t \in 1, \ldots$, first generate

$$\begin{aligned}&\sigma_t^2 | \mathcal{A}_{t-1}, \mu_{t-1}, \beta_{t-1} \\ &\sim \text{Inverse-gamma}\left(n/2 + a_0, \left\|r - M\mu_{t-1} - \mathcal{A}_{t-1}\beta_{t-1}\right\|_2^2/2 + b_0\right).\end{aligned}$$

Next, generate $(\mu_t, \beta_t)^T | \mathcal{A}_{t-1}, \sigma_t^2 \sim N_{p+q}(c_n(\mathcal{A}_{t-1}, \sigma_t^2), C_n(\mathcal{A}_{t-1}, \sigma_t^2))$ where

$$\begin{aligned}C_n(\mathcal{A}_{t-1}, \sigma_t^2) &= \left(\begin{pmatrix} M & \mathcal{A}_{t-1}\end{pmatrix}^T \begin{pmatrix} M & \mathcal{A}_{t-1}\end{pmatrix}/\sigma_t^2 + C_0^{-1}\right)^{-1} \\ c_n(\mathcal{A}_{t-1}, \sigma_t^2) &= C_n(\mathcal{A}_{t-1}, \sigma_t^2)\left[\begin{pmatrix} M & \mathcal{A}_{t-1}\end{pmatrix}^T r/\sigma_t^2 + C_0^{-1}c_0\right].\end{aligned}$$

Finally, generate independently $\mathcal{A}_{i,t} | \mu_t, \beta_t, \sigma_t^2 \sim N_p(d_{n,i}(\mu_t, \beta_t, \sigma_t^2), D_{n,i}(\mu_t, \beta_t, \sigma_t^2))$

where

$$D_{n,i}(\beta_t, \sigma_t^2) = \left(\beta_t \beta_t^T / \sigma_t^2 + D_i^{-1}\right)^{-1}$$
$$d_{n,i}(\mu_t, \beta_t, \sigma_t^2) = D_{n,i}(\beta_t, \sigma_t^2)\left[D_i^{-1} d_i + \left(r_i - M_i^T \mu_t\right)\beta_t / \sigma_t^2\right]$$

to obtain $\mathcal{A}_t = (\mathcal{A}_{1,t}, \ldots, \mathcal{A}_{n,t})^T$.

For points $(\mathcal{A}, \mu, \beta, \sigma^2)$ and $(\mathcal{A}', \mu', \beta', \sigma^{2'})$, the Gibbs sampler has Markov transition density

$$p\left((\mathcal{A}, \mu, \beta, \sigma^2), (\mathcal{A}', \mu', \beta', \sigma^{2'})\right) = \pi_n(\mathcal{A}'|\mu', \beta', \sigma^{2'})\pi_n(\mu', \beta'|\mathcal{A}, \sigma^{2'})\pi_n(\sigma^{2'}|\mathcal{A}, \mu, \beta)$$

and Markov transition kernel defined for suitable sets B by

$$P\left((\mathcal{A}, \mu, \beta, \sigma^2), B\right) = \int\int\int_B p\left((\mathcal{A}, \mu, \beta, \sigma^2), (\mathcal{A}', \mu', \beta', \sigma^{2'})\right) d\mathcal{A}' d(\mu', \beta')^T d\sigma^{2'}.$$

The Markov kernel at larger iteration times $t \geq 2$ is defined recursively with $P^1 \equiv P$ by

$$P^t\left((\mathcal{A}, \mu, \beta, \sigma^2), B\right) = \int P^{t-1}(\cdot, B)\, dP\left((\mathcal{A}, \mu, \beta, \sigma^2), \cdot\right).$$

We will use the following drift function defined by

$$V(\mathcal{A}, \mu, \beta) = \frac{1}{2}\sum_{i=1}^{n}(\mathcal{A}_i - d_i)^T D_i^{-1}(\mathcal{A}_i - d_i) + \frac{1}{2}(\mu, \beta)C_0^{-1}(\mu, \beta)^T$$

combined with a minorization condition (see [76, Chapter 15] and [45]) to show there is a $\rho \in (0, 1)$ and $M_0 \in (0, \infty)$ so that for any initialization $\mathcal{A}, \mu, \beta, \sigma^2$,

$$\sup_{|\varphi| \leq 1 + M_0 V} \left| \int \varphi dP^t\left((\mathcal{A}, \mu, \beta, \sigma^2), \cdot\right) - \int \varphi d\Pi_n \right| \leq M(\mathcal{A}, \mu, \beta)\rho^t \tag{4.6}$$

where $M(\mathcal{A}, \mu, \beta) = 2 + M_0 V(\mathcal{A}, \mu, \beta) + M_0 \int V d\Pi_n$. The condition (4.6) implies the

Gibbs sampler is geometrically ergodic. We now state our main result.

Theorem 6. *The 3-variable deterministic scan Gibbs sampler* $(A_t, \mu_t, \beta_t, \sigma_t^2)_{t=0}^{\infty}$ *for the general density* (4.5) *is geometrically ergodic.*

Proof. We first show a minorization condition. We will use a special property of this Gibbs Markov kernel P that for suitable sets B, $P(\cdot, B)$ is a function of only the parameters $(\mathcal{A}, \mu, \beta)$ and does not depend on σ^2. For $R \in (0, \infty)$, define the function g_R by

$$g_R(\mathcal{A}', \mu', \beta', \sigma^{2'}) = \inf_{V(\mathcal{A},\mu,\beta) \leq R} \pi_n(\mathcal{A}'|\mu', \beta', \sigma^{2'}) \pi_n(\mu', \beta'|\mathcal{A}, \sigma^{2'}) \pi_n(\sigma^{2'}|\mathcal{A}, \mu, \beta)$$

and the constant $Z_{g_R} = \int g_R(\mathcal{A}', \mu', \beta', \sigma^{2'}) d\mathcal{A}' d\mu' d\beta' d\sigma^{2'}$. The drift function V is continuous and a strongly convex function so its sublevel sets are closed and bounded [80, Corollary 3.2.2]. For fixed $\mu', \beta', \sigma^{2'}$, the function

$$(\mathcal{A}, \mu, \beta) \mapsto \pi_n(\mu', \beta'|\mathcal{A}, \sigma^{2'}) \pi_n(\sigma^{2'}|\mathcal{A}, \mu, \beta)$$

is continuous and achieves its minimum over compact sets. Thus, Z_{g_R} is not 0 and we can define the probability measure $\nu_R(\cdot) = Z_{g_R}^{-1} \int_{\cdot} g_R(\mathcal{A}', \mu', \beta', \sigma^{2'})) d\mathcal{A}' d\mu' d\beta' d\sigma^{2'}$. For any $R \in (0, \infty)$ and any suitable set B,

$$\begin{aligned}
&\inf_{\substack{\sigma^2 \in (0,\infty), \\ V(\mathcal{A},\mu,\beta) \leq R}} P\left((\mathcal{A}, \mu, \beta, \sigma^2), B\right) \\
&= \inf_{V(\mathcal{A},\mu,\beta) \leq R} \int_B \pi_n(\mathcal{A}'|\mu', \beta', \sigma^{2'}) \pi_n(\mu', \beta'|\mathcal{A}, \sigma^{2'}) \pi_n(\sigma^{2'}|\mathcal{A}, \mu, \beta) d\mathcal{A}' d\mu' d\beta' d\sigma^{2'} \\
&\geq \int_B g_R(\mathcal{A}', \mu', \beta', \sigma^{2'}) d\mathcal{A}' d\mu' d\beta' d\sigma^{2'} \\
&= Z_{g_R} \nu_R(B).
\end{aligned}$$

It remains to show a drift condition.

Fix $\mathcal{A}_0, \mu_0, \beta_0$, and fix $i \in 1, \ldots, n$. Since D_i is SPD, let $D_i = D_i^{1/2} D_i^{1/2}$, $D_i^{-1} = D_i^{-1/2} D_i^{-1/2}$ where $D_i^{1/2}, D_i^{-1/2}$ are SPD. Using the identity

$$d_{n,i}(\mu, \beta, \sigma^2) = d_i + \left(r_i - M_i^T \mu - d_i^T \beta\right) \left(\beta\beta^T + \sigma^2 D_i^{-1}\right)^{-1} \beta$$

and taking the expectation with respect to $\mathcal{A}_i|, \mathcal{A}_0, \mu_0, \beta_0, \mu, \beta, \sigma^2$

$$\mathbb{E}\left[\frac{1}{2}\left\|D_i^{-1/2}(\mathcal{A}_i - d_i)\right\|_2^2 |\mathcal{A}_0, \mu_0, \beta_0, \mu, \beta, \sigma^2\right] \tag{4.7a}$$

$$= \frac{1}{2}\left(r_i - M_i^T \mu - d_i^T \beta\right)^2 \left\|D_i^{-1/2}\left(\beta\beta^T + \sigma^2 D_i^{-1}\right)^{-1} \beta\right\|_2^2 \tag{4.7b}$$

$$+ \frac{1}{2}\mathrm{tr}\left[D_i^{-1/2}\left(\beta\beta^T/\sigma^2 + D_i^{-1}\right)^{-1} D_i^{-1/2}\right]. \tag{4.7c}$$

Using singular value decomposition [50, Theorem 2.6.3], we can write $D_i^{1/2}\beta = U\Sigma_\beta$ where $U \in \mathbb{R}^{p\times p}$ with $U^T U = UU^T = I_p$ and $\Sigma_\beta = \left(\left\|D_i^{1/2}\beta\right\|_2, 0, \ldots, 0\right)^T$ to get

$$\begin{aligned}\left(\beta\beta^T/\sigma^2 + D_i^{-1}\right)^{-1} &= D_i^{1/2}(D_i^{1/2}\beta)(D_i^{1/2}\beta)^T/\sigma^2 + I_p)^{-1} D_i^{1/2} \\ &= D_i^{1/2} U(\Sigma_\beta\Sigma_\beta^T/\sigma^2 + I_p)^{-1} U^T D_i^{1/2}. \end{aligned} \tag{4.8}$$

Using (4.8) and properties of the trace

$$\begin{aligned}\frac{1}{2}\mathrm{tr}\left[D_i^{-1/2}\left(\beta\beta^T/\sigma^2 + D_i^{-1}\right)^{-1} D_i^{-1/2}\right] &= \frac{1}{2}\mathrm{tr}\left[\left(\Sigma_\beta\Sigma_\beta^T/\sigma^2 + I_p\right)^{-1} U^T U\right] \\ &\leq \frac{p}{2}.\end{aligned}$$

For $x \in [0, \infty)$ and $a \in (0, \infty)$, we have the inequality

$$x/(x^2 + a) \leq \frac{1}{2\sqrt{a}}. \tag{4.9}$$

Using inequalities (4.8) and (4.9), the matrix 2-norm is sub-multiplicative, and $\|U\|_2 =$

1, we have

$$\begin{aligned}\left\|D_i^{-1/2}\left(\beta\beta^T+\sigma^2D_i^{-1}\right)^{-1}\beta\right\|_2^2 &= \left\|U(\Sigma_\beta\Sigma_\beta^T+\sigma^2I_p)^{-1}U^TD_i^{1/2}\beta\right\|_2^2\\ &\le \|U\|_2^2\left\|(\Sigma_\beta\Sigma_\beta^T+\sigma^2I_p)^{-1}\Sigma_\beta\right\|_2^2\\ &\le \left[\frac{\left\|D_i^{1/2}\beta\right\|_2}{\left\|D_i^{1/2}\beta\right\|_2^2+\sigma^2}\right]^2\\ &\le \frac{1}{4\sigma^2}.\end{aligned}$$

Define the matrix $\tilde{X}=(d_1,\ldots,d_n)^T$. Applying these upper bounds to (4.7) and combining for each $i\in 1,\ldots,n$,

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{2}\sum_{i=1}^n\left\|D_i^{-1/2}(\mathcal{A}_i-d_i)\right\|_2^2\left|\mathcal{A}_0,\mu_0,\beta_0,\mu,\beta,\sigma^2\right.\right]\\ &\le \frac{1}{8\sigma^2}\left\|r-M\mu-\tilde{X}\beta\right\|_2^2+\frac{pn}{2}.\end{aligned}$$

By convexity, for every x,y, $\|x-y\|_2^2\le 2\|x\|_2^2+2\|y\|_2^2$. Since C_0 is SPD, let $C_0=C_0^{1/2}C_0^{1/2}$, $C_0^{-1}=C_0^{-1/2}C_0^{-1/2}$ where $C_0^{1/2},C_0^{-1/2}$ are SPD. Using convexity, and the matrix 2-norm is sub-multiplicative, we have

$$\frac{1}{2}\left\|r-M\mu-\tilde{X}\beta\right\|_2^2\le\|r\|_2^2+\left\|\begin{pmatrix}M & \tilde{X}\end{pmatrix}C_0^{1/2}\right\|_2^2\left\|C_0^{-1/2}(\mu,\beta)^T\right\|_2^2.$$

Therefore,

$$\mathbb{E}\left[\frac{1}{2}\sum_{i=1}^{n}\left\|D_i^{-1/2}(\mathcal{A}_i - d_i)\right\|_2^2 \middle| \mathcal{A}_0, \mu_0, \beta_0, \mu, \beta, \sigma^2\right] \tag{4.10a}$$

$$\leq \frac{1}{8\sigma^2}\left\|r - M\mu - \tilde{X}\beta\right\|_2^2 + \frac{pn}{2} \tag{4.10b}$$

$$\leq \frac{\|r\|_2^2}{4\sigma^2} + \frac{\left\|\begin{pmatrix} M & \tilde{X}\end{pmatrix} C_0^{1/2}\right\|_2^2}{4\sigma^2}\left\|C_0^{-1/2}(\mu,\beta)^T\right\|_2^2 + \frac{pn}{2}. \tag{4.10c}$$

Taking the expectation with respect to $\mu, \beta | \mathcal{A}_0, \mu_0, \beta_0, \sigma^2$

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{2}\left\|C_0^{-1/2}(\mu,\beta)^T\right\|_2^2 \middle| \sigma^2, \mathcal{A}_0, \mu_0, \beta_0\right] \\ &= \frac{1}{2}\left\|C_0^{-1/2} c_n(\mathcal{A}_0, \sigma^2)\right\|_2^2 + \frac{1}{2}\mathrm{tr}(C_0^{-1/2} C_n(\mathcal{A}_0, \sigma^2) C_0^{-1/2}).\end{aligned}$$

Using singular value decomposition [50, Theorem 2.6.3], choose matrices $U \in \mathbb{R}^{n\times n}$, $V \in \mathbb{R}^{(p+q)\times(p+q)}$ with $U^TU = UU^T = I_n$ and $V^TV = VV^T = I_{p+q}$ and a rectangular diagonal matrix $\Sigma_{\mathcal{A}_0} \in \mathbb{R}^{n\times(p+q)}$ with diagonal nonnegative singular values $(\sigma_{\mathcal{A}_0,i})_{i=1}^{p+q}$ so that $\begin{pmatrix} M & \mathcal{A}_0\end{pmatrix} C_0^{1/2} = U\Sigma_{\mathcal{A}_0}V^T$. We then have

$$\begin{aligned}C_n(\mathcal{A}_0, \sigma^2) &= \left(\begin{pmatrix} M & \mathcal{A}_0\end{pmatrix}^T \begin{pmatrix} M & \mathcal{A}_0\end{pmatrix}/\sigma^2 + C_0^{-1}\right)^{-1} \\ &= C_0^{1/2}\left(\left[\begin{pmatrix} M & \mathcal{A}_0\end{pmatrix} C_0^{1/2}\right]^T \left[\begin{pmatrix} M & \mathcal{A}_0\end{pmatrix} C_0^{1/2}\right]/\sigma^2 + I_{p+q}\right)^{-1} C_0^{1/2} \\ &= C_0^{1/2} V\left(\Sigma_{\mathcal{A}_0}^T\Sigma_{\mathcal{A}_0}/\sigma^2 + I_{p+q}\right)^{-1} V^T C_0^{1/2}. \end{aligned} \tag{4.11}$$

Using (4.11) and properties of the trace

$$\begin{aligned}\frac{1}{2}\mathrm{tr}(C_0^{-1/2} C_n(\mathcal{A}_0, \sigma^2) C_0^{-1/2}) &\leq \frac{1}{2}\max_i\left[\left(\sigma_{\mathcal{A}_0,i}^2/\sigma^2 + 1\right)^{-1}\right]\mathrm{tr}(V^TV) \\ &\leq \frac{p+q}{2}.\end{aligned}$$

Using convexity,

$$\begin{aligned}
&\frac{1}{2}\left\|C_0^{-1/2}c_n(\mathcal{A}_0,\sigma^2)\right\|_2^2 \\
&= \frac{1}{2}\left\|C_0^{-1/2}C_n(\mathcal{A}_0,\sigma^2)\left[\begin{pmatrix} M & \mathcal{A}_0\end{pmatrix}^T r/\sigma^2 + C_0^{-1}c_0\right]\right\|_2^2 \\
&\leq \left\|C_0^{-1/2}C_n(\mathcal{A}_0,\sigma^2)\begin{pmatrix} M & \mathcal{A}_0\end{pmatrix}^T r/\sigma^2\right\|_2^2 + \left\|C_0^{-1/2}C_n(\mathcal{A}_0,\sigma^2)C_0^{-1}c_0\right\|_2^2.
\end{aligned}$$

Using the inequality (4.9) and the identity (4.11),

$$\begin{aligned}
&\left\|C_0^{-1/2}C_n(\mathcal{A}_0,\sigma^2)\begin{pmatrix} M & \mathcal{A}_0\end{pmatrix}^T r/\sigma^2\right\|_2^2 \\
&= \left\|V\left(\Sigma_{\mathcal{A}_0}^T\Sigma_{\mathcal{A}_0}+\sigma^2 I_{p+q}\right)^{-1}V^T\left[\begin{pmatrix} M & \mathcal{A}_0\end{pmatrix}C_0^{1/2}\right]^T r\right\|_2^2 \\
&= \left\|V\left(\Sigma_{\mathcal{A}_0}^T\Sigma_{\mathcal{A}_0}+\sigma^2 I_{p+q}\right)^{-1}\Sigma_{\mathcal{A}_0}^T U^T r\right\|_2^2 \\
&\leq \|V\|_2^2\left\|\left(\Sigma_{\mathcal{A}_0}^T\Sigma_{\mathcal{A}_0}+\sigma^2 I_{p+q}\right)^{-1}\Sigma_{\mathcal{A}_0}^T\right\|_2^2\left\|U^T\right\|_2^2\|r\|_2^2 \\
&\leq \max_i\left(\frac{\sigma_{\mathcal{A}_0,i}}{\sigma_{\mathcal{A}_0,i}^2+\sigma^2}\right)^2\|r\|_2^2 \\
&\leq \frac{\|r\|_2^2}{4\sigma^2}.
\end{aligned}$$

Using (4.11),

$$\begin{aligned}
&\left\|C_0^{-1/2}C_n(\mathcal{A}_0,\sigma^2)C_0^{-1}c_0\right\|_2^2 \\
&= \left\|V\left(\Sigma_{\mathcal{A}_0}^T\Sigma_{\mathcal{A}_0}/\sigma^2+I_{p+q}\right)^{-1}V^T C_0^{-1/2}c_0\right\|_2^2 \\
&\leq \|V\|_2^2\left\|\left(\Sigma_{\mathcal{A}_0}^T\Sigma_{\mathcal{A}_0}/\sigma^2+I_{p+q}\right)^{-1}\right\|_2^2\left\|V^T\right\|_2^2 c_0^T C_0^{-1}c_0 \\
&\leq c_0^T C_0^{-1}c_0.
\end{aligned}$$

Combining the upper bounds

$$\mathbb{E}\left[\frac{1}{2}\left\|C_0^{-1/2}(\mu,\beta)^T\right\|_2^2 \big| \sigma^2, \mathcal{A}_0, \mu_0, \beta_0\right] \leq \frac{\|r\|_2^2}{4\sigma^2} + c_0^T C_0^{-1} c_0 + \frac{p+q}{2}. \tag{4.12}$$

Now using (4.10) and (4.12) and taking the iterated expectation with respect to $\mu, \beta | \mathcal{A}_0, \mu_0, \beta_0, \sigma^2$,

$$\mathbb{E}\left[\frac{1}{2}\sum_{i=1}^n \left\|D_i^{-1/2}(\mathcal{A}_i - d_i)\right\|_2^2 \big| \sigma^2, \mathcal{A}_0, \mu_0, \beta_0\right] \tag{4.13a}$$

$$\leq \frac{\|r\|_2^2}{4\sigma^2} + \frac{\left\|\begin{pmatrix} M & \tilde{X}\end{pmatrix} C_0^{1/2}\right\|_2^2}{2\sigma^2}\mathbb{E}\left[\frac{1}{2}\left\|C_0^{-1/2}(\mu,\beta)^T\right\|_2^2 \big| \sigma^2, \mathcal{A}_0, \mu_0, \beta_0\right] + \frac{pn}{2} \tag{4.13b}$$

$$\leq \frac{1}{\sigma^2}\left[\frac{\|r\|_2^2}{4} + \frac{\left\|\begin{pmatrix} M & \tilde{X}\end{pmatrix} C_0^{1/2}\right\|_2^2 (p+q)}{4} + \frac{\left\|\begin{pmatrix} M & \tilde{X}\end{pmatrix} C_0^{1/2}\right\|_2^2}{2} c_0^T C_0^{-1} c_0\right] \tag{4.13c}$$

$$+ \frac{1}{(\sigma^2)^2}\frac{\left\|\begin{pmatrix} M & \tilde{X}\end{pmatrix} C_0^{1/2}\right\|_2^2 \|r\|_2^2}{8} + \frac{pn}{2}. \tag{4.13d}$$

Since $1/\sigma^2 | \mathcal{A}_0, \mu_0, \beta_0$ has a gamma distribution, then using the identities for the moments, we have for every positive integer k:

$$\begin{aligned}\mathbb{E}\left[\left[1/\sigma^2\right]^k \big| \mathcal{A}_0, \mu_0, \beta_0\right] &= \frac{\Gamma(k + n/2 + a_0)}{\Gamma(n/2 + a_0)\left[\|r - M\mu_0 - \mathcal{A}_0\beta_0\|_2^2/2 + b_0\right]^k} \\ &\leq \frac{\Gamma(k + n/2 + a_0)}{\Gamma(n/2 + a_0) b_0^k}.\end{aligned}$$

Taking the iterated expectation with respect to $\sigma^2 | \mathcal{A}_0, \mu_0, \beta_0$ in (4.12) and (4.13), there is a constant $L \in (0, \infty)$ so that the drift condition is satisfied with

$$\mathbb{E}\left[V(\mathcal{A}, \mu, \beta) | \mathcal{A}_0, \mu_0, \beta_0\right] \leq L.$$

□

4.2.1 Bayesian EIV regression with errors in the features

Using Theorem 6, we develop geometrically ergodic Gibbs samplers for Bayesian EIV regression with additive Gaussian error in the features. For the remainder, we write the observed data as $Y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$, $X = (x_1, \ldots, x_n)^T \in \mathbb{R}^{n \times p}$, and $Z = (Z_1, \ldots, Z_n)^T \in \mathbb{R}^{n \times q}$. Consider the Bayesian EIV regression (4.1) with Berkson errors and priors (4.3) and (4.4). We will write the posterior density π_n for this Bayesian model as

$$\begin{aligned}\pi_n(\mathcal{A}, \alpha, \beta, \sigma^2) \propto & \left(\frac{1}{\sigma^2}\right)^{n/2+a_0+1} \exp\left[-\frac{1}{\sigma^2}\left[\|Y - Z\alpha - \mathcal{A}\beta\|_2^2/2 + b_0\right]\right] \\ & \times \prod_{i=1}^{n} \exp\left(-\frac{1}{2}(\mathcal{A}_i - x_i)^T V_i^{-1}(\mathcal{A}_i - x_i)\right) \\ & \times \exp\left(-\frac{1}{2}((\alpha, \beta)^T - j_0)^T J_0^{-1}((\alpha, \beta)^T - j_0)\right).\end{aligned}$$

This posterior density is a special case of the general density (4.5) choosing $\mu \equiv \alpha$, $M \equiv Z$, $r \equiv Y$, $c_0, C_0 \equiv j_0, J_0$, and $d_i, D_i \equiv x_i, V_i$.

We can define a 3-variable deterministic scan Gibbs sampler which generates a Markov chain $(\mathcal{A}_t, \alpha_t, \beta_t, \sigma_t^2)_{t=0}^{\infty}$ for this posterior density as a special case of the Gibbs sampler constructed in Section 4.2. Initialize $(\mathcal{A}_0, \alpha_0, \beta_0, \sigma_0^2)$ and for $t \in 1, \ldots,$

1. Generate $\sigma_t^2 | \mathcal{A}_{t-1}, \alpha_{t-1}, \beta_{t-1} \sim$ Inverse-Gamma $(n/2 + a_0, b_{n,t})$ where

$$b_{n,t} = \|Y - Z\alpha_{t-1} - \mathcal{A}_{t-1}\beta_{t-1}\|_2^2/2 + b_0$$

2. Generate $(\alpha_t, \beta_t)^T | \mathcal{A}_{t-1}, \sigma_t^2 \sim N_{p+q}(j_{n,t}, J_{n,t})$ where

$$\begin{aligned}J_{n,t} &= \left(\begin{pmatrix} Z & \mathcal{A}_{t-1} \end{pmatrix}^T \begin{pmatrix} Z & \mathcal{A}_{t-1} \end{pmatrix}/\sigma_t^2 + J_0^{-1}\right)^{-1} \\ j_{n,t} &= J_{n,t}\left[\begin{pmatrix} Z & \mathcal{A}_{t-1} \end{pmatrix}^T Y/\sigma_t^2 + J_0^{-1} j_0\right]\end{aligned}$$

3. Generate $\mathcal{A}_{i,t}|\alpha_t, \beta_t, \sigma_t^2 \sim N_p(k_{n,i,t}, K_{n,i,t}), i \in 1, \ldots, n$ where

$$K_{n,i,t} = \left(\beta_t \beta_t^T / \sigma_t^2 + V_i^{-1}\right)^{-1}$$
$$k_{n,i,t} = K_{n,i,t}\left[V_i^{-1} x_i + \left(y_i - Z_i^T \alpha_t\right) \beta_t / \sigma_t^2\right]$$

Applying Theorem 6, we have the following result.

Corollary 5. *The 3-variable Gibbs sampler* $(\mathcal{A}_t, \alpha_t, \beta_t, \sigma_t^2)_{t=0}^{\infty}$ *for the posterior in Bayesian EIV regression* (4.1) *with Berkson errors and priors* (4.3) *and* (4.4) *is geometrically ergodic.*

Now consider Bayesian EIV regression (4.1) with additive Gaussian error in X_i using classical errors and priors (4.3) and (4.4). The posterior has density

$$\begin{aligned}\pi_n(\mathcal{A}, \alpha, \beta, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{n/2+a_0+1} \exp\left[-\frac{1}{\sigma^2}\left[\|Y - Z\alpha - \mathcal{A}\beta\|_2^2/2 + b_0\right]\right] \\ \times \prod_{i=1}^{n} \exp\left(-\frac{1}{2}(\mathcal{A}_i - k_i')^T (V_i^{-1} + K_i^{-1})(\mathcal{A}_i - k_i')\right) \\ \times \exp\left(-\frac{1}{2}((\alpha, \beta)^T - j_0)^T J_0^{-1}((\alpha, \beta)^T - j_0)\right)\end{aligned}$$

where $k_i' = (V_i^{-1} + K_i^{-1})^{-1}\left[V_i^{-1} x_i + K_i^{-1} k_i\right]$. The posterior density is also a special case of the general density (4.5) when $\mu \equiv \alpha$, $Z \equiv M$, $r \equiv Y$, and $c_0, C_0 \equiv j_0, J_0$, and $d_i, D_i \equiv k_i', (V_i^{-1} + K_i^{-1})^{-1}$.

We define a 3-variable deterministic scan Gibbs sampler similarly. Initialize $(\mathcal{A}_0, \alpha_0, \beta_0, \sigma_0^2)$ and for $t \in 1, \ldots,$

1. Generate $\sigma_t^2|\mathcal{A}_{t-1}, \alpha_{t-1}, \beta_{t-1} \sim \text{Inverse-Gamma}\left(n/2 + a_0, b_{n,t}\right)$

2. Generate $(\alpha_t, \beta_t)^T|\mathcal{A}_{t-1}, \sigma_t^2 \sim N_{p+q}(j_{n,t}, J_{n,t})$

3. Generate $\mathcal{A}_{i,t}|\alpha_t, \beta_t, \sigma_t^2 \sim N_p(k'_{n,i,t}, K'_{n,i,t}), i \in 1, \ldots, n$ where

$$K'_{n,i,t} = \left(\beta_t \beta_t^T / \sigma_t^2 + V_i^{-1} + K_i^{-1}\right)^{-1}$$
$$k'_{n,i,t} = K'_{n,i,t}\left[V_i^{-1} x_i + K_i^{-1} k_i + \left(y_i - Z_i^T \alpha_t\right) \beta_t / \sigma_t^2\right]$$

We also have the following as a direct result of Theorem 6.

Corollary 6. *The 3-variable Gibbs sampler* $(\mathcal{A}_t, \alpha_t, \beta_t, \sigma_t^2)_{t=0}^{\infty}$ *for the posterior in Bayesian EIV regression* (4.1) *with classical errors and priors* (4.3) *and* (4.4) *is geometrically ergodic.*

4.2.2 Bayesian EIV regression with errors in the response and features

Similar to the previous section, we develop geometrically ergodic Gibbs samplers for Bayesian EIV regression with additional additive Gaussian error in the features and response. Consider the Bayesian EIV regression (4.2) with Berkson errors in X_i and additional error in Y_i along with priors (4.3) and (4.4). Let $U_0 = \text{diag}(u_i^2) \in \mathbb{R}^{n \times n}$. The Bayesian posterior Π_n has density

$$\begin{aligned}\pi_n(\mathcal{A}, \nu, \alpha, \beta, \sigma^2) \propto & \left(\frac{1}{\sigma^2}\right)^{n/2+a_0+1} \exp\left[-\frac{1}{\sigma^2}\left(\|\nu - Z\alpha - \mathcal{A}\beta\|_2^2/2 + b_0\right)\right] \\ & \times \exp\left(-\frac{1}{2}(\nu - Y)^T U_0^{-1}(\nu - Y)\right) \\ & \times \prod_{i=1}^{n} \exp\left(-\frac{1}{2}(\mathcal{A}_i - x_i)^T V_i^{-1}(\mathcal{A}_i - x_i)\right) \\ & \times \exp\left(-\frac{1}{2}((\alpha, \beta)^T - j_0)^T J_0^{-1}((\alpha, \beta)^T - j_0)\right).\end{aligned}$$

This posterior density is a special case of the density (4.5) when $\mu \equiv (\nu, \alpha)^T$, $M \equiv \begin{pmatrix} -I & Z \end{pmatrix}$, $r \equiv 0$, $c_0 = (Y, j_0)^T$,

$$C_0 \equiv \begin{pmatrix} U_0 & 0 \\ 0 & J_0 \end{pmatrix},$$

and $d_i, D_i \equiv x_i, V_i$.

We define a 3-variable deterministic scan Gibbs sampler which generates a Markov chain $(\mathcal{A}_t, \nu_t, \alpha_t, \beta_t, \sigma_t^2)_{t=0}^{\infty}$ for this posterior density. Initialize $(\mathcal{A}_0, \nu_0, \alpha_0, \beta_0, \sigma_0^2)$ and for $t \in 1, \ldots,$

1. Generate $\sigma_t^2 | \mathcal{A}_{t-1}, \nu_{t-1}, \alpha_{t-1}, \beta_{t-1} \sim$ Inverse-Gamma $\left(n/2 + a_0, b'_{n,t}\right)$ where

$$b'_{n,t} = \left\| \nu_{t-1} - Z\alpha_{t-1} - \mathcal{A}_{t-1}\beta_{t-1} \right\|_2^2 /2 + b_0$$

2. Generate $(\nu_t, \alpha_t, \beta_t)^T | \mathcal{A}_{t-1}, \sigma_t^2 \sim N_{p+q}(j'_{n,t}, J'_{n,t})$ where

$$J'_{n,t} = \left(\begin{pmatrix} -I & Z & \mathcal{A}_{t-1} \end{pmatrix}^T \begin{pmatrix} -I & Z & \mathcal{A}_{t-1} \end{pmatrix} /\sigma_t^2 + \begin{pmatrix} U_0^{-1} & 0 \\ 0 & J_0^{-1} \end{pmatrix} \right)^{-1}$$
$$j'_{n,t} = J'_{n,t} \begin{pmatrix} U_0^{-1} & 0 \\ 0 & J_0^{-1} \end{pmatrix} (Y, j_0)^T$$

3. Generate $\mathcal{A}_{i,t} | \nu_t, \alpha_t, \beta_t, \sigma_t^2 \sim N_p(k''_{n,i,t}, K''_{n,i,t}), i \in 1, \ldots, n$ where

$$K''_{n,i,t} = \left(\beta_t \beta_t^T /\sigma_t^2 + V_i^{-1} \right)^{-1}$$
$$k''_{n,i,t} = K''_{n,i,t} \left[V_i^{-1} x_i + \left(\nu_{i,t} - Z_i^T \alpha_t \right) \beta_t /\sigma_t^2 \right]$$

Using Theorem 6, we have the following result.

Corollary 7. *The 3-variable Gibbs sampler* $(\mathcal{A}_t, \nu_t, \alpha_t, \beta_t, \sigma_t^2)_{t=0}^{\infty}$ *for Bayesian EIV regression* (4.2) *with Berkson errors and priors* (4.3) *and* (4.4) *is geometrically ergodic.*

Now consider the Bayesian EIV regression (4.2) with classical errors in X_i and additional error in Y_i with priors (4.3) and (4.4). The posterior Π_n for this Bayesian model has density

$$\begin{aligned}\pi_n(\mathcal{A}, \nu, \alpha, \beta, \sigma^2) \propto & \left(\frac{1}{\sigma^2}\right)^{n/2+a_0+1} \exp\left[-\frac{1}{\sigma^2}\left(\|\nu - Z\alpha - \mathcal{A}\beta\|_2^2/2 + b_0\right)\right] \\ & \times \exp\left(-\frac{1}{2}(\nu - Y)^T U_0^{-1}(\nu - Y)\right) \\ & \times \prod_{i=1}^{n} \exp\left(-\frac{1}{2}(\mathcal{A}_i - k_i')^T (V_i^{-1} + K_i^{-1})(\mathcal{A}_i - k_i')\right) \\ & \times \exp\left(-\frac{1}{2}((\alpha, \beta)^T - j_0)^T J_0^{-1}((\alpha, \beta)^T - j_0)\right).\end{aligned}$$

This posterior density is also a special case of the density (4.5) when $\mu \equiv (\nu, \alpha)^T$, $M \equiv \begin{pmatrix} -I & Z \end{pmatrix}$, $r \equiv 0$, $c_0 = (Y, j_0)^T$,

$$C_0 \equiv \begin{pmatrix} U_0 & 0 \\ 0 & J_0 \end{pmatrix},$$

and $d_i, D_i \equiv k_i', (V_i^{-1} + K_i^{-1})^{-1}$.

We define a 3-variable deterministic scan Gibbs sampler similarly. Initialize $(\mathcal{A}_0, \nu_0, \alpha_0, \beta_0, \sigma_0^2)$ and for $t \in 1, \ldots,$

1. Generate $\sigma_t^2 | \mathcal{A}_{t-1}, \nu_{t-1}, \alpha_{t-1}, \beta_{t-1} \sim$ Inverse-Gamma $\left(n/2 + a_0, b_{n,t}'\right)$
2. Generate $(\nu_t, \alpha_t, \beta_t)^T | \mathcal{A}_{t-1}, \sigma_t^2 \sim N_{p+q}(j_{n,t}', J_{n,t}')$

3. Generate $\mathcal{A}_{i,t}|\nu_t, \alpha_t, \beta_t, \sigma_t^2 \sim N_p(k'''_{n,i,t}, K'''_{n,i,t}), i \in 1, \ldots, n$ where

$$K'''_{n,i,t} = \left(\beta_t \beta_t^T/\sigma_t^2 + V_i^{-1} + K_i^{-1}\right)^{-1}$$
$$k'''_{n,i,t} = K'''_{n,i,t}\left[V_i^{-1}x_i + K_i^{-1}k_i + \left(\nu_{i,t} - Z_i^T\alpha_t\right)\beta_t/\sigma_t^2\right]$$

Using Theorem 6, we have the following result.

Corollary 8. *The 3-variable Gibbs sampler* $(\mathcal{A}_t, \nu_t, \alpha_t, \beta_t, \sigma_t^2)_{t=0}^{\infty}$ *for Bayesian EIV regression* (4.2) *with classical errors and priors* (4.3) *and* (4.4) *is geometrically ergodic.*

4.3 Conclusion and Future Directions

We showed using a 3-variable deterministic scan Gibbs sampler to sample the posterior in 4 different Bayesian EIV regression models with additive Gaussian errors and independent priors is always geometrically ergodic. This is of pragmatic importance to practitioners as trustworthy estimation from a Gibbs sampler is dependent on the speed of convergence of the Markov chain. More specifically, time averages from the Markov chains have many practically relevant theoretical guarantees such as a central limit theorem. Secondly, these Gibbs samplers can be simulated efficiently without the need for complex, intermediate Metropolis-Hastings or rejection sampling steps.

There are many future research directions in studying the convergence of Gibbs samplers in EIV models. It appears reasonable that some Gibbs samplers for generalized linear models such as the Pólya-Gamma sampler will also be geometrically ergodic [19, 87, 125]. It seems also interesting to look at alternative errors in the variables such as non-Gaussian or non-additive errors.

References

[1] Achic, B. G. B., Wang, T., Su, Y., Kipnis, V., Dodd, K., and Carroll, R. J. (2018). Categorizing a continuous predictor subject to measurement error. *Electronic Journal of Statistics*, 12(2):4032 – 4056.

[2] Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

[3] Bach, F. (2010). Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384 – 414.

[4] Bai, Z. D. and Yin, Y. Q. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, 21(3):1275 – 1294.

[5] Barker, A. A. (1965). Monte Carlo calculations of the radial distribution functions for a proton? electron plasma. *Australian Journal of Physics*, 18(2):119–134.

[6] Belloni, A. and Chernozhukov, V. (2009). On the computational complexity of MCMC-based estimators in large samples. *The Annals of Statistics*, 37(4):2011–2055.

[7] Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association*, 45(250):164–180.

[8] Bhadra, A. and Carroll, R. J. (2016). Exact sampling of the unobserved covariates in Bayesian spline models for measurement error problems. *Statistics and Computing*, 26(4):827–840.

[9] Bogachev, V. I. (1998). *Gaussian Measures*. American Mathematical Society.

[10] Bou-Rabee, N. and Eberle, A. (2020). Markov chain Monte Carlo methods.

[11] Bresson, G., Chaturvedi, A., Rahman, M. A., and Shalabh (2021). Seemingly unrelated regression with measurement error: estimation via Markov chain Monte Carlo and mean field variational Bayes approximation. *The International Journal of Biostatistics*, 17(1):75–97.

[12] Brown, A. and Jones, G. L. (2021). Exact convergence analysis for Metropolis-Hastings independence samplers in Wasserstein distances. *preprint arXiv:2111.10406*.

[13] Buonaccorsi, J. (2010). *Measurement Error: Models, Methods, and Applications*. Chapman and Hall/CRC, 1 edition.

[14] Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC, 2 edition.

[15] Chan, K. S. and Geyer, C. J. (1994). Discussion: Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1747–1758.

[16] Charisse Farr, A., Mengersen, K., Ruggeri, F., Simpson, D., Wu, P., and Yarlagadda, P. (2020). Combining opinions for use in Bayesian networks: A measurement error approach. *International Statistical Review*, 88(2):335 – 353.

[17] Chen, M.-H. and Shao, Q.-M. (2001). Propriety of posterior distribution for dichotomous quantal response models. *Proceedings of the American Mathematical Society*, 129(1):293–302.

[18] Chewi, S., Lu, C., Ahn, K., Cheng, X., Gouic, T. L., and Rigollet, P. (2021). Optimal dimension dependence of the Metropolis-Adjusted Langevin algorithm. *Proceedings of Thirty Fourth Conference on Learning Theory, PMLR*.

[19] Choi, H. M. and Hobert, J. P. (2013). The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics*, 7:2054–2064.

[20] Clayton, D. et al. (1992). Models for the analysis of cohort and case-control studies with inaccurately measured exposures. *Statistical models for longitudinal studies of health*, pages 301–331.

[21] Dalalyan, A. S. (2017). Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society, Series B*, 79:651–676.

[22] Damgaard, C. (2020). Measurement uncertainty in ecological and environmental models. *Trends in Ecology and Evolution*, 35(10):871–873.

[23] Dellaportas, P. and Stephens, D. A. (1995). Bayesian analysis of errors-in-variables regression models. *Biometrics*, 51(3):1085–1095.

[24] Demidenko, E. (2001). Computational aspects of probit model. *Mathematical Communications*, 6:233–247.

[25] Doss, C. R., Flegal, J. M., Jones, G. L., and Neath, R. C. (2014). Markov chain Monte Carlo estimation of quantiles. *Electronic Journal of Statistics*, 8(2):2448–2478.

[26] 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.

[27] Durmus, A. and Moulines, É. (2019). High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25:2854–2882.

[28] Dwivedi, R., Chen, Y., Wainwright, M. J., and Yu, B. (2018). Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 793–797.

[29] Eberle, A. (2014). Error bounds for Metropolis–Hastings algorithms applied to perturbations of Gaussian measures in high dimensions. *The Annals of Applied Probability*, 24(1):337 – 377.

[30] Ekvall, K. O. and Jones, G. L. (2021). Convergence analysis of a collapsed Gibbs sampler for Bayesian vector autoregressions. *Electronic Journal of Statistics*, 15:691–721.

[31] Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1):342 – 368.

[32] Fang, X., Li, B., Alkhatib, H., Zeng, W., and Yao, Y. (2017). Bayesian inference for the errors-in-variables model. *Studia Geophysica et Geodaetica*, 61(1):1573–1626.

[33] Flegal, J. M. and Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics*, 38(2):1034 – 1070.

[34] Folland, G. B. (1999). *Real Analysis: Modern Techniques and Their Applications*. Wiley, 2 edition.

[35] Fort, G., Moulines, E., Roberts, G. O., and Rosenthal, J. S. (2003). On the geometric ergodicity of hybrid samplers. *Journal of Applied Probability*, 40(1):123–146.

[36] Fuller, W. A. (1987). *Measurement Error Models*. John Wiley.

[37] Geman, S. (1980). A limit theorem for the norm of random matrices. *The Annals of Probability*, 8:252–261.

[38] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 6(6):721–741.

[39] Gibbs, A. L. (2004). Convergence in the Wasserstein metric for Markov chain Monte Carlo algorithms with applications to image restoration. *Stochastic Models*, 20(4):473–492.

[40] Giraudo, D. (2014). Product measure with a Dirac delta marginal. Mathematics Stack Exchange.

[41] Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 73(2):123–214.

[42] Goodfellow, I., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

[43] Groß, M. (2016). Modeling body height in prehistory using a spatio-temporal Bayesian errors-in-variables model. *AStA Advances in Statistical Analysis*, 100(3):289–311.

[44] Gustafson, P. (2003). *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. CRC Press.

[45] Hairer, M. and Mattingly, J. C. (2011). Yet another look at Harris' ergodic theorem for Markov chains. *Seminar on Stochastic Analysis, Random Fields and Applications VI*, 63.

[46] Hairer, M., Stuart, A. M., and Vollmer, S. J. (2014). Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *The Annals of Applied Probability*, 24:2455–2490.

[47] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

[48] Hilbe, J. M., de Souza, R. S., and Ishida, E. E. O. (2017). *Bayesian Models for Astrophysical Data: Using R, JAGS, Python, and Stan*. Cambridge University Press.

[49] Hiriart-Urruty, J.-B. and Lemaéchal, C. (2001). *Fundamentals of Convex Analysis*. Springer-Verlag Berlin Heidelberg, 1 edition.

[50] Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press.

[51] Huang, H.-J. (2010). Bayesian analysis of errors-in-variables growth curve models. *PhD Dissertation*.

[52] Håvard Rue, Sara Martino, N. C. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71(2):319–392.

[53] Jarner, S. F. and Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stochastic Processes and their Applications*, 85:341–361.

[54] Jin, R. and Tan, A. (2020). Central limit theorems for Markov chains based on their convergence rates in Wasserstein distance. *preprint arXiv:2002.09427*.

[55] Johndrow, J. E., Smith, A., Pillai, N., and Dunson, D. B. (2019). MCMC for imbalanced categorical data. *Journal of the American Statistical Association*, 114:1394–1403.

[56] Johnson, L. T. and Geyer, C. J. (2012). Variable transformation to obtain geometric ergodicity in the random-walk Metropolis algorithm. *The Annals of Statistics*, 40(6):3050 – 3076.

[57] Jones, G. L. (2004). On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320.

[58] Jones, G. L. and Hobert, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, 16:312–334.

[59] Jones, G. L. and Hobert, J. P. (2004). Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. *The Annals of Statistics*, 32(2):784 – 817.

[60] Joulin, A. and Ollivier, Y. (2010). Curvature, concentration and error estimates for Markov chain Monte Carlo. *The Annals of Probability*, 38:2418–2442.

[61] Kantorovich, L. V. and Rubinstein, G. S. (1957). On a function space in certain extremal problems. *Dokl. Akad. Nauk USSR*, 115(6):1058–1061.

[62] Kass, R. E., Tierney, L., and Kadane, J. B. (1990). The validity of posterior expansions based on Laplace's method. *Bayesian and Likelihood Methods in Statistics and Econometrics*, pages 473–488.

[63] Kelly, B. C. (2012). Measurement error models in astronomy. In *Statistical challenges in modern astronomy V*, pages 147–162. Springer.

[64] Keru Wu, Scott Schmidler, Y. C. (2021). Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling. *preprint arXiv:2109.13055*.

[65] Khare, K. and Hobert, J. P. (2012). Geometric ergodicity of the Gibbs sampler for Bayesian quantile regression. *Journal of Multivariate Analysis*, 112:108–116.

[66] Komorowski, T. and Walczuk, A. (2011). Central limit theorem for Markov processes with spectral gap in the Wasserstein metric. *Stochastic Processes and their Applications*, 122:2155–2184.

[67] Kröger, H., Hoffmann, R., and Pakpahan, E. (2016). Consequences of measurement error for inference in cross-lagged panel design—the example of the reciprocal causal relationship between subjective health and socio-economic status. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 179(2):607–628.

[68] Łatuszyński, K., Miasojedow, B., and Niemiro, W. (2013). Nonasymptotic bounds on the estimation error of MCMC algorithms. *Bernoulli*, 19(5A):2033 – 2066.

[69] Lawler, G. F. and Sokal, A. D. (1988). Bounds on the l^2 spectrum for Markov chains and Markov processes: A generalization of Cheeger's inequality. *Transactions of the American Mathematical Society*, 309(2):557–580.

[70] Lee, Y. T., Shen, R., and Tian, K. (2021). Lower bounds on Metropolized sampling methods for well-conditioned distributions. *preprint arXiv:2106.05480*.

[71] Liu, J. S. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, 6(2):113–119.

[72] Madras, N. and Sezer, D. (2010). Quantitative bounds for Markov chain convergence: Wasserstein and total variation distances. *Bernoulli*, 16(3):882 – 908.

[73] Mallick, B. K. and Gelfand, A. E. (1996). Semiparametric errors-in-variables models a Bayesian approach. *Journal of Statistical Planning and Inference*, 52(3):307–321.

[74] Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24:101–121.

[75] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087–1092.

[76] Meyn, S. P. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*. Cambridge University Press, USA, 2 edition.

[77] Michalek, J. E. and Tripathi, R. C. (1980). The effect of errors in diagnosis and measurement on the estimation of the probability of an event. *Journal of the American Statistical Association*, 75(371):713–721.

[78] Miller, J. W. (2021). Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22:1–53.

[79] Muff, S., Riebler, A., Held, L., Rue, H., and Saner, P. (2015). Bayesian analysis of measurement error models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 64(2):231–252.

[80] Nesterov, Y. (2018). *Lectures on Convex Optimization*. Springer International Publishing, 2 edition.

[81] Papaspiliopoulos, O., Roberts, G. O., and Zanella, G. (2019). Scalable inference for crossed random effects models. *Biometrika*, 107(1):25–40.

[82] Papaspiliopoulos, O., Stumpf-Fétizon, T., and Zanella, G. (2021). Scalable computation for Bayesian hierarchical models. *preprint arXiv:2103.10875*.

[83] Peskun, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612.

[84] Pham, T. H., Ormerod, J. T., and Wand, M. P. (2013). Mean field variational Bayesian inference for nonparametric regression with measurement error. *Computational Statistics and Data Analysis*, 68:375–387.

[85] Pierre, J., Robert, C. P., and Smith, M. H. (2011). Using parallel computation to improve independent Metropolis–Hastings based estimation. *Journal of Computational and Graphical Statistics*, 20(3):616–635.

[86] Pollice, A., Jona Lasinio, G., Rossi, R., Amato, M., Kneib, T., and Lang, S. (2019). Bayesian measurement error correction in structured additive distributional regression with an application to the analysis of sensor data on soil–plant variability. *Stochastic Environmental Research and Risk Assessment*, 33(3):747–763.

[87] Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108:1339–1349.

[88] Qin, Q. and Hobert, J. P. (2019). Convergence complexity analysis of Albert and Chib's algorithm for Bayesian probit regression. *Annals of Statistics*, 47:2320–2347.

[89] Qin, Q. and Hobert, J. P. (2021). On the limitations of single-step drift and minorization in Markov chain convergence analysis. *The Annals of Applied Probability*, 31(4):1633 – 1659.

[90] Qin, Q. and Hobert, J. P. (2022a). Geometric convergence bounds for markov chains in wasserstein distance based on generalized drift and contraction conditions. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 58(2):872 – 889.

[91] Qin, Q. and Hobert, J. P. (2022b). Wasserstein-based methods for convergence complexity analysis of MCMC with applications. *The Annals of Applied Probability*, 32(1):124 – 166.

[92] 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. *preprint arXiv:1508.00947*.

[93] Richardson S, G. W. (1993). A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *American journal of epidemiology*, 138(6):430–42.

[94] Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, Berlin, Heidelberg, 2 edition.

[95] Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110 – 120.

[96] Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255—-268.

[97] Roberts, G. O. and Rosenthal, J. S. (2011). Quantitative non-geometric convergence bounds for independence samplers. *Methodology and Computing in Applied Probability volume*, 13:391—-403.

[98] Roberts, G. O. and Tweedie, R. L. (1996a). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363.

[99] Roberts, G. O. and Tweedie, R. L. (1996b). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83:95–110.

[100] Robertson, N., Flegal, J. M., Vats, D., and Jones, G. L. (2021). Assessing and visualizing simultaneous simulation error. *Journal of Computational and Graphical Statistics*, 30:324–334.

[101] Rodrigues, J. and Bolfarine, H. (2007). Bayesian inference for an extended simple regression measurement error model using skewed priors. *Bayesian Analysis*, 61(2):349–364.

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

[103] Roy, V. and Zhang, L. (2021). Convergence of position-dependent MALA with application to conditional simulation in GLMMs. *preprint arXiv:2108.12662*.

[104] Rudolf, D. (2012). Explicit error bounds for Markov chain Monte Carlo. *Dissertationes Mathematicae*, 485:1 – 93.

[105] Schmidler, S. C. and Woodard, D. B. (2011). Lower bounds on the convergence rates of adaptive MCMC methods. *Technical Report, Duke University*.

[106] Shephard, N. and Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84(3):653–667.

[107] Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(4):749–760.

[108] Stefanski, L. A. and Carroll, R. J. (1985). Covariate measurement error in logistic regression. *The Annals of Statistics*, 13(4):1335 – 1351.

[109] Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116:14516–14525.

[110] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. In *International Conference on Learning Representations*.

[111] Tang, N.-S., Li, D.-W., and Tang, A.-M. (2017). Semiparametric Bayesian inference on generalized linear measurement error models. *Statistical Papers*, 58(4):1091–1113.

[112] Tang, Y. and Reid, N. (2021). Laplace and Saddlepoint Approximations in High Dimensions. *preprint arXiv:2107.10885*.

[113] Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22:1701–1728.

[114] Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces. *The Annals of Applied Probability*, 8:1–9.

[115] Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

[116] Tierney, L., Kass, R. E., and Kadane, J. B. (1989). Approximate marginal densities of nonlinear functions. *Biometrika*, 76(3):425–433.

[117] Torabi, M., Ghosh, M., Myung, J., and Steel, M. (2021). Measurement error in linear regression models with fat tails and skewed errors. *Communications in Statistics - Theory and Methods*, 0(0):1–20.

[118] Vaart, A. W. v. d. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[119] Vats, D., Flegal, J. M., and Jones, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, 106:321–337.

[120] Vidal, I. and Arellano-Valle, R. B. (2010). Bayesian inference for dependent elliptical measurement error models. *Journal of Multivariate Analysis*, 101(10):2587–2597.

[121] Vidal, I. and Iglesias, P. (2008). Comparison between a measurement error model and a linear model without measurement error. *Computational Statistics and Data Analysis*, 53(1):92–102.

[122] Villani, C. (2003). *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society.

[123] Villani, C. (2009). *Optimal Transport: Old and New*. Springer Berlin, Heidelberg, 1 edition.

[124] Wang, G. (2022). Exact convergence analysis of the independent Metropolis-Hastings algorithms. *Bernoulli*, 28(3):2012 – 2033.

[125] Wang, X. and Roy, V. (2018). Geometric ergodicity of Pólya-gamma Gibbs sampler for Bayesian logistic regression with a flat prior. *Electronic Journal of Statistics*, 12(2):3295 – 3311.

[126] Woodard, D., Schmidler, S., and Huber, M. (2009a). Sufficient conditions for torpid mixing of parallel and simulated tempering. *Electronic Journal of Probability*, 14:780 – 804.

[127] Woodard, D. B. (2015). A lower bound on the mixing time of uniformly ergodic Markov chains in terms of the spectral radius. *preprint arXiv:1405.0028*.

[128] Woodard, D. B., Schmidler, S. C., and Huber, M. (2009b). Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *The Annals of Applied Probability*, 19(2):617 – 640.

[129] Yang, J. and Rosenthal, J. S. (2017). Complexity results for MCMC derived from quantitative bounds. *preprint arXiv:1708.00829*.

[130] 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.

[131] Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243.

[132] Zhuo, B. and Gao, C. (2021). Mixing time of Metropolis-Hastings for Bayesian community detection. *Journal of Machine Learning Research*, 22(10):1–89.

Appendix A

Exact Convergence Analysis for Metropolis-Hastings Independence Samplers in Wasserstein Distances

A.1 Proof of Theorem 1

The proof will proceed by establishing the upper and lower bounds separately in Lemmas 8 and 9, respectively. This is done largely because the conditions for the upper bound are weaker than those for the lower bound. The following definitions will be used in the proofs of Lemmas 8 and 9. First, for $\theta \in \Theta$, real-valued measurable functions f, and a Markov kernel K, we will use the notation $K^t f(\theta) = \int f dK^t(\theta, \cdot) = \int f(\theta') K^t(\theta, d\theta')$ and $K^0 f(\theta) = f(\theta)$. Second, recall that for functions $\varphi : \mathbb{R}^d \to \mathbb{R}$,

$$\|\varphi\|_{\mathrm{Lip}(\rho)} = \sup_{x,y,x\neq y} \{|\varphi(x) - \varphi(y)|/\rho(x,y)\}.$$

Lemma 8. *Let* $\epsilon^* = \inf_{\theta\in\Theta}\{q(\theta)/\pi(\theta)\}$. *Then*

$$\sup_{\theta\in\Theta} \mathcal{W}_\rho(P^t(\theta,\cdot),\Pi) \leq (1-\epsilon^*)^t \sup_{\theta\in\Theta} \int \rho(\theta,\cdot) d\Pi.$$

Proof. Let $\theta \in \Theta$ and let φ satisfy $\|\varphi\|_{\mathrm{Lip}(\rho)} \leq 1$. The existence of ϵ^* implies the

minorization condition $P(\theta, \cdot) \geq \epsilon^* \Pi(\cdot)$ [113, Corollary 4] which, in turn, ensures the residual kernel $R(\theta, \cdot) = [P(\theta, \cdot) - \epsilon^* \Pi(\cdot)]/(1 - \epsilon^*)$ is a Markov kernel with invariant distribution Π. It then follows that

$$\begin{aligned}
\int \varphi dP^t(\theta, \cdot) - \int \varphi d\Pi &= (1 - \epsilon^*) \left[\int R\varphi dP^{t-1}(\theta, \cdot) - \int \varphi d\Pi \right] \\
&= (1 - \epsilon^*) \left[\int R\varphi dP^{t-1}(\theta, \cdot) - \int R\varphi d\Pi \right] \\
&\cdots \\
&= (1 - \epsilon^*)^t \left[\int \varphi dR^t(\theta, \cdot) - \int \varphi d\Pi \right].
\end{aligned}$$

Since φ is Lipschitz with respect to ρ, we then have

$$\begin{aligned}
\left| \int \varphi dR^t(\theta, \cdot) - \int \varphi d\Pi \right| &= \left| \int \int \left[\varphi(\theta') - \varphi(\omega) \right] d\Pi(\omega) dR^t(\theta, \theta') \right| \\
&\leq \int \int \rho(\theta', \omega) d\Pi(\omega) dR^t(\theta, \theta') \\
&\leq \sup_{\theta' \in \Theta} \int \rho(\theta', \cdot) d\Pi.
\end{aligned}$$

Taking the supremum with respect to φ and using the Kantorovich-Rubinstein theorem [122, Theorem 1.14],

$$\begin{aligned}
\sup_{\theta \in \Theta} \mathcal{W}_\rho \left(P^t(\theta, \cdot), \Pi \right) &= \sup_{\theta \in \Theta} \sup_{\|\varphi\|_{\mathrm{Lip}(\rho)} \leq 1} \left[\int \varphi dP^t(\theta, \cdot) - \int \varphi d\Pi \right] \\
&\leq (1 - \epsilon^*)^t \sup_{\theta \in \Theta} \int \rho(\theta, \cdot) d\Pi.
\end{aligned}$$

□

We now turn our attention to establishing the lower bound.

Lemma 9. *Let $\epsilon^* = \inf_{\theta \in \Theta} \{q(\theta)/\pi(\theta)\}$. Suppose q is lower semicontinuous and π is upper semicontinuous on Θ. Suppose there is a sequence of compact sets $B_n \subseteq \Theta$*

increasing in diameter to Θ. *If* $\rho(\cdot,\cdot) \leq 1$, *then*

$$\sup_{\theta\in\Theta} \mathcal{W}_\rho(P^t(\theta,\cdot),\Pi) \geq (1-\epsilon^*)^t \inf_{\theta\in\Theta} \int \rho(\cdot,\theta) d\Pi.$$

Proof. We can assume $\Pi(B_n) > 0$ or else we can take n large enough so this holds. Since $\pi, q > 0$ and π is upper semicontinuous on Θ, then q/π is lower semicontinuous on Θ. We have that $\inf_{\theta\in B_n}\{q(\theta)/\pi(\theta)\}$ is monotonically non-increasing to $\epsilon^* = \inf_{\theta\in\Theta}\{q(\theta)/\pi(\theta)\}$. Since we have assumed lower semicontinuity, the $\inf_{\theta\in K}\{q(\theta)/\pi(\theta)\}$ is attained over any compact set $K \subseteq \Theta$. Then define the sequence

$$\theta_n^* = \text{argmin}_{\theta\in B_n}\{q(\theta)/\pi(\theta)\}. \tag{A.1}$$

We can then define the sequence

$$\epsilon_{\theta_n^*} = \inf_{\theta\in B_n}\{q(\theta)/\pi(\theta)\} = q(\theta_n^*)/\pi(\theta_n^*)$$

and this is monotonically non-increasing to ϵ^*.

Define P_n to be the Metropolis-Hastings independence kernel with independent proposal Q with density q and target distribution $\Pi(\cdot|B_n)$ with density $\pi(\cdot|B_n) = \pi(\cdot)I_{B_n}(\cdot)/\Pi(B_n)$. By construction, $\Pi(B_n) > 0$ and this is well-defined. The key part of the proof is that if we start at any $\theta_n \in B_n$, this kernel P_n and the kernel P only disagree outside of B_n. For $\theta_n \in B_n$, we have $\pi(\theta_n) > 0$, $I_{B_n}(\theta_n) = 1$, and since $\Theta \equiv \text{supp}(q)$ by assumption, then $q(\theta_n) > 0$. Also, if $y \in B_n^c \cap \Theta$, then $\min\left\{\frac{\pi(y)I_{B_n}(y)q(\theta_n)}{\pi(\theta_n)q(y)}, 1\right\} = 0$. Let $M_1(\mathbb{R}^d)$ be the set of measurable functions $\varphi : \mathbb{R}^d \to$

$\mathbb{R}$ with $\sup_{x\in\mathbb{R}^d}|\varphi(x)|\leq 1$. Therefore, for any $\theta_n\in B_n$ and any function $\varphi\in M_1(\mathbb{R}^d)$,

$$\begin{aligned}\int_{\mathbb{R}^d}\varphi dP_n(\theta_n,\cdot) &= \int_{B_n}\varphi(y)\min\left\{\frac{\pi(y)q(\theta_n)}{\pi(\theta_n)q(y)},1\right\}q(y)dy\\ &\quad+\varphi(\theta_n)\left(1-\int_{B_n}\min\left\{\frac{\pi(y)q(\theta_n)}{\pi(\theta_n)q(y)},1\right\}q(y)dy\right).\end{aligned}$$

Let $\epsilon\in(0,1-\epsilon^*)$. Since Q and Π are probability measures, we may then choose n_ϵ sufficiently large such that for all $n\geq n_\epsilon$,

$$2\max\left\{\Pi(B_n^c),Q(B_n^c)\right\}\leq\epsilon/2.$$

We then have

$$\begin{aligned}&\sup_{\theta_n\in B_n}\sup_{\varphi\in M_1(\mathbb{R}^d)}\left|\int_{\mathbb{R}^d}\varphi dP_n(\theta_n,\cdot)-\int_{\mathbb{R}^d}\varphi dP(\theta_n,\cdot)\right|\\ &=\sup_{\theta_n\in B_n}\sup_{\varphi\in M_1(\mathbb{R}^d)}\left|\int_{B_n^c\cap\Theta}\varphi(y)\min\left\{\frac{\pi(y)q(\theta_n)}{\pi(\theta_n)q(y)},1\right\}q(y)dy\right.\\ &\qquad\qquad\left.+\varphi(\theta_n)\int_{B_n^c\cap\Theta}\min\left\{\frac{\pi(y)q(\theta_n)}{\pi(\theta_n)q(y)},1\right\}q(y)dy\right|\\ &\leq 2\int_{B_n^c}q(y)dy\\ &\leq\epsilon/2.\end{aligned}\tag{A.2}$$

Similarly,

$$\begin{aligned}
&\sup_{\varphi\in M_1(\mathbb{R}^d)}\left|\int_{\mathbb{R}^d}\varphi d\Pi(\cdot|B_n)-\int_{\mathbb{R}^d}\varphi d\Pi\right| \\
&=\sup_{\varphi\in M_1(\mathbb{R}^d)}\left|\int_{\mathbb{R}^d}\varphi\left(1-\Pi(B_n)\right)d\Pi(\cdot|B_n)-\int_{\mathbb{R}^d}\varphi d\Pi(\cdot|B_n^c)\Pi(B_n^c)\right| \\
&=\Pi(B_n^c)\sup_{\varphi\in M_1(\mathbb{R}^d)}\left|\int_{\mathbb{R}^d}\varphi d\Pi(\cdot|B_n)-\int_{\mathbb{R}^d}\varphi d\Pi(\cdot|B_n^c)\right| \\
&\leq 2\Pi(B_n^c) \\
&\leq \epsilon/2. \qquad \text{(A.3)}
\end{aligned}$$

With θ_n^* as in (A.1), let $\psi_n(\cdot)=-\rho(\cdot,\theta_n^*)$. Then for any $x,y\in\mathbb{R}^d$,

$$|\psi_n(x)-\psi_n(y)|\leq\rho(x,y) \qquad \text{(A.4)}$$

and $\psi_n\in M_1(\mathbb{R}^d)$. Since Π is invariant for the kernel P,

$$\int_{\mathbb{R}^d}\psi_n dP^t(\theta_n^*,\cdot)-\int_{\mathbb{R}^d}\psi_n d\Pi=\int_{\mathbb{R}^d}P^{t-1}\psi_n(\cdot)dP(\theta_n^*,\cdot)-\int_{\mathbb{R}^d}P^{t-1}\psi_n(\cdot)d\Pi(\cdot). \qquad \text{(A.5)}$$

Now for any integer s with $1\leq s\leq t$, the function $P^s\psi_n\in M_1(\mathbb{R}^d)$ since P is a Markov kernel. Since $\theta_n^*\in B_n$ and $\pi(\theta_n^*)>0$, using (A.2), (A.3), and (A.5),

$$\begin{aligned}
&\int_{\mathbb{R}^d}\psi_n dP^t(\theta_n^*,\cdot)-\int_{\mathbb{R}^d}\psi_n d\Pi \\
&\geq\int_{\mathbb{R}^d}P^{t-1}\psi_n dP_n(\theta_n^*,\cdot)-\int_{\mathbb{R}^d}P^{t-1}\psi_n d\Pi(\cdot|B_n)-\epsilon. \qquad \text{(A.6)}
\end{aligned}$$

By construction of θ_n^* in (A.1), we have

$$\begin{aligned}\inf_{\theta\in B_n}\{q(\theta)/\pi(\theta|B_n)\} &= \Pi(B_n)\inf_{\theta\in B_n}\{q(\theta)/\pi(\theta)\}\\ &= \Pi(B_n)q(\theta_n^*)/\pi(\theta_n^*)\\ &= \epsilon_{\theta_n^*}\Pi(B_n)\\ &= q(\theta_n^*)/\pi(\theta_n^*|B_n).\end{aligned}$$

For measurable $A\subset\mathbb{R}^d$ [124, Remark 1, Theorem 2], we then have the identity

$$P_n(\theta_n^*,A) = \epsilon_{\theta_n^*}\Pi(B_n)\Pi(A|B_n) + \left(1-\epsilon_{\theta_n^*}\Pi(B_n)\right)\delta_{\theta_n^*}(A). \tag{A.7}$$

Since $P^{t-1}\psi_n$ is a bounded measurable function, (A.7) gives the identity:

$$\begin{aligned}&\int_{\mathbb{R}^d} P^{t-1}\psi_n(\cdot)dP_n(\theta_n^*,\cdot) - \int_{\mathbb{R}^d} P^{t-1}\psi_n(\cdot)d\Pi(\cdot|B_n)\\ &= (1-\epsilon_{\theta_n^*}\Pi(B_n))\left(P^{t-1}\psi_n(\theta_n^*) - \int_{\mathbb{R}^d} P^{t-1}\psi_n(\cdot)d\Pi(\cdot|B_n)\right).\end{aligned} \tag{A.8}$$

Using (A.6) in the first inequality, (A.8) in the second inequality, (A.3) in the third inequality, and using the invariance of Π for the Markov kernel P in the last inequality,

$$\begin{aligned}&\int_{\mathbb{R}^d} P^{t-1}\psi_n P(\theta_n^*,\cdot) - \int_{\mathbb{R}^d} P^{t-1}\psi_n d\Pi\\ &\geq \int_{\mathbb{R}^d} P^{t-1}\psi_n dP_n(\theta_n^*,\cdot) - \int_{\mathbb{R}^d} P^{t-1}\psi_n d\Pi(\cdot|B_n) - \epsilon.\\ &\geq (1-\epsilon_{\theta_n^*}\Pi(B_n))\left(P^{t-1}\psi_n(\theta_n^*) - \int_{\mathbb{R}^d} P^{t-1}\psi_n d\Pi(\cdot|B_n)\right) - \epsilon\\ &\geq (1-\epsilon_{\theta_n^*}\Pi(B_n))\left(P^{t-1}\psi_n(\theta_n^*) - \int_{\mathbb{R}^d} P^{t-1}\psi_n d\Pi\right) - 2\epsilon\\ &\geq (1-\epsilon_{\theta_n^*}\Pi(B_n))\left(\int_{\mathbb{R}^d} P^{t-2}\psi_n dP(\theta_n^*,\cdot) - \int_{\mathbb{R}^d} P^{t-2}\psi_n d\Pi\right) - 2\epsilon.\end{aligned}$$

Applying this inequality recursively and using the definition of ψ_n

$$\begin{aligned}
&\int_{\mathbb{R}^d} \psi_n dP^t(\theta_n^*, \cdot) - \int_{\mathbb{R}^d} \psi_n d\Pi \\
&= \int_{\mathbb{R}^d} P^{t-1}\psi_n dP(\theta_n^*, \cdot) - \int_{\mathbb{R}^d} P^{t-1}\psi_n d\Pi \\
&\geq (1 - \epsilon_{\theta_n^*}\Pi(B_n))^t \left(\psi_n(\theta_n^*) - \int_{\mathbb{R}^d} \psi_n d\Pi\right) - 2\epsilon \sum_{s=0}^{t-1} (1 - \epsilon_{\theta_n^*}\Pi(B_n))^s \\
&= (1 - \epsilon_{\theta_n^*}\Pi(B_n))^t \int_{\mathbb{R}^d} \rho(\theta, \theta_n^*) d\Pi - 2\epsilon \sum_{s=0}^{t-1} (1 - \epsilon_{\theta_n^*}\Pi(B_n))^s. \qquad \text{(A.9)}
\end{aligned}$$

Since $\Pi(B_n) \to 1$ and $\epsilon_{\theta_n^*} \to \epsilon^*$, we may take n large enough so that

$$\left|\epsilon_{\theta_n^*}\Pi(B_n) - \epsilon^*\right| \leq \epsilon.$$

For all large enough n and since $\epsilon < 1 - \epsilon^*$, we lower bound (A.9) to get

$$\begin{aligned}
&\int_{\mathbb{R}^d} \psi_n dP^t(\theta_n^*, \cdot) - \int_{\mathbb{R}^d} \psi_n d\Pi \\
&\geq (1 - \epsilon^* - \epsilon)^t \inf_{\theta \in \Theta} \int \rho(\cdot, \theta) d\Pi - 2\epsilon \sum_{s=0}^{t-1} (1 - \epsilon^* + \epsilon)^s. \qquad \text{(A.10)}
\end{aligned}$$

Combining (A.4) and (A.10), we lower bound the Wasserstein distance (we could instead use more sophisticated duality results [122, Proposition 1.5, Theorem 1.14]) with

$$\begin{aligned}
&\sup_{\theta \in \Theta} \mathcal{W}_\rho(P^t(\theta, \cdot), \Pi) \\
&\geq \mathcal{W}_\rho(P^t(\theta_n^*, \cdot), \Pi) \\
&\geq (1 - \epsilon^* - \epsilon)^t \inf_{\theta \in \Theta} \int \rho(\cdot, \theta) d\Pi - 2\epsilon \sum_{s=0}^{t-1} (1 - \epsilon^* + \epsilon)^s.
\end{aligned}$$

Since this holds for all small ϵ, the proof is complete by taking the limit as $\epsilon \downarrow 0$. □

A.2 Proof of Theorem 2

The following lemma uses a similar argument as [80, Lemma 3.2.3] where the standard Euclidean norm is used.

Lemma 10. *Let $C \in \mathbb{R}^{d\times d}$ be a positive-definite, symmetric matrix and $\alpha \in (0,\infty)$. Let $f : \mathbb{R}^d \to \mathbb{R}$ and suppose $\theta \mapsto f(\theta) - \alpha\theta^T C^{-1}\theta/2$ is convex for all points on $\mathbb{R}^d$. Then there exists $\theta^* = argmin_{\theta\in\mathbb{R}^d} f(\theta)$ and*

$$f(\theta) \geq f(\theta^*) + \frac{\alpha}{2}\left(\theta - \theta^*\right)^T C^{-1}(\theta - \theta^*).$$

Proof. Since the function $f(\theta) - \alpha\theta^T C^{-1}\theta/2$ is convex for all points on $\mathbb{R}^d$, it follows that for any $\lambda \in [0,1]$ and any $(\theta,\theta') \in \mathbb{R}^d \times \mathbb{R}^d$,

$$f(\lambda\theta + (1-\lambda)\theta') \leq \lambda f(\theta) + (1-\lambda)f(\theta') - \frac{\alpha}{2}\lambda(1-\lambda)\left(\theta' - \theta\right)^T C^{-1}(\theta' - \theta).$$

Since C^{-1} is positive-definite, then $\alpha\lambda(1-\lambda)\left(\theta' - \theta\right)^T C^{-1}(\theta' - \theta)/2$ is nonnegative and this implies that f is a convex function. It can also be shown that $\lim_{\|\theta\|\to+\infty} f(\theta) = +\infty$ and since f is lower semicontinuous, then f attains its minimum $\theta^* \in \mathbb{R}^d$. The right directional derivative

$$f'(\theta^*;\theta) = \lim_{t\downarrow 0} t^{-1}\left[f(\theta^* + t\theta) - f(\theta^*)\right]$$

exists for all points $\theta \in \mathbb{R}^d$ [80, Theorem 3.1.12]. For $\lambda \in (0,1)$, we have

$$\begin{aligned}&\frac{1}{(1-\lambda)}\frac{1}{\lambda}\left[f(\theta^* + \lambda(\theta - \theta^*)) - f(\theta^*)\right] - \frac{1}{(1-\lambda)}\left(f(\theta) - f(\theta^*)\right)\\ &\leq -\frac{\alpha}{2}(\theta - \theta^*)^T C^{-1}(\theta - \theta^*).\end{aligned}$$

Taking the limit with $\lambda \downarrow 0$, we have that

$$f'(\theta^*; \theta - \theta^*) - f(\theta) + f(\theta^*) \leq -\frac{\alpha}{2}(\theta - \theta^*)^T C^{-1}(\theta - \theta^*).$$

Since θ^* is the minimum of f, then the right directional derivative satisfies $f'(\theta^*; \theta - \theta^*) \geq 0$ for all $\theta \in \mathbb{R}^d$. Therefore for all $\theta \in \mathbb{R}^d$,

$$f(\theta) \geq f(\theta^*) + \frac{\alpha}{2}(\theta - \theta^*)^T C^{-1}(\theta - \theta^*).$$

□

Proof of Theorem 2. We may define the function $f : \mathbb{R}^d \times (0, \infty) \to \mathbb{R}$ by

$$f(\beta, \sigma) = \frac{s_0 + s_{n,r}(\beta)}{\sigma} + (n + v_0 + 1 + d/2)\log(\sigma)$$

and write the posterior density as $\pi(\beta, \sigma | X, Y) = Z^{-1}_{\Pi(\cdot|X,Y)} \exp\left(-f(\beta, \sigma)\right)$. Since the function $\beta \mapsto s_{n,r}(\beta) - \beta^T C^{-1}\beta/2$ is a convex function on $\mathbb{R}^d$, then by Lemma 10 for every $\beta \in \mathbb{R}^d$,

$$s_{n,r}(\beta) \geq s_{n,r}(\beta^*) + \frac{1}{2}(\beta - \beta^*)^T C^{-1}(\beta - \beta^*).$$

For any $(\beta, \sigma) \in \mathbb{R}^d \times (0, \infty)$, we then have the lower bound

$$\begin{aligned} & f(\beta, \sigma) \\ & = \frac{s_0 + s_{n,r}(\beta)}{\sigma} + (n + v_0 + 1 + d/2)\log(\sigma) \\ & \geq \frac{s_0 + s_{n,r}(\beta^*)}{\sigma} + (n + v_0 + 1 + d/2)\log(\sigma) + \frac{1}{2\sigma}\left(\beta - \beta^*\right)^T C^{-1}\left(\beta - \beta^*\right). \end{aligned}$$

This implies

$$\begin{aligned} & f(\beta, \sigma) \\ & - \frac{1}{2\sigma}\left(\beta - \beta^*\right)^T C^{-1}\left(\beta - \beta^*\right) - \frac{s_0 + s_{n,r}(\beta^*)}{\sigma} - (n + v_0 + 1 + d/2)\log(\sigma) \\ & \geq 0. \end{aligned}$$

Let q denote the proposal's normal-inverse-gamma density. For any $\sigma_0 \in (0, \infty)$ and for every $(\beta, \sigma) \in \mathbb{R}^d \times (0, \infty)$, we have shown

$$\begin{aligned} \frac{q(\beta, \sigma)}{\pi(\beta, \sigma)} & \geq Z_{\Pi(\cdot|X,Y)}\,(2\pi)^{-\frac{d}{2}}\det(C)^{-1/2}\left(s_0 + s_{n,r}(\beta^*)\right)^{n+v_0}\Gamma(n + v_0)^{-1} \\ & = \frac{q(\beta^*, \sigma_0)}{\pi(\beta^*, \sigma_0)}. \end{aligned}$$

An application of Proposition 1 completes the proof. □

A.3 Proof of Theorem 3

Under our assumption, we may write the matrix $X = n^{-1/2}G$ where G is a matrix with i.i.d Gaussian entries with mean 0 and variance σ^2. Denote the largest eigenvalue of the matrix X^TX by $\lambda_{max}(X^TX)$. Therefore, as $d, n \to \infty$ in such a way that

$d/n \to \gamma \in (0, +\infty)$,

$$\lambda_{max}\left(X^TX\right) = \lambda_{max}\left(\frac{1}{n}G^TG\right) = \frac{1}{n}\sup_{v\in\mathbb{R}^d, \|v\|_2=1}\left\|G^TGv\right\|_2 \to (1+\gamma^{1/2})^2\sigma^2$$

almost surely [37, Theorem 1].

Define the function $f : \mathbb{R}^d \to \mathbb{R}$ by $f(\beta) = \ell_n(\beta) + \alpha/2\beta^TC^{-1}\beta$ where ℓ_n is the negative log-likelihood loss function and define $Z_Q = (2\pi)^{d/2}\det\left(\alpha^{-1}C\right)^{1/2}$. We will first lower bound the quantity $\exp(f(\beta^*))Z_{\Pi(\cdot|X,Y)}/Z_Q$. We have that for any $\beta \in \mathbb{R}^d$ and any $v \in \mathbb{R}^d$,

$$v^TH_{\ell_n}(\beta)v \leq r_0\lambda_{max}(X^TX)\left\|v\right\|_2^2.$$

This implies that for any $\beta \in \mathbb{R}^d$ and any $v \in \mathbb{R}^d$, the Hessian of f, denoted by H_f, satisfies

$$v^TH_f(\beta)v \leq v^T\left(r_0\lambda_{max}(X^TX)I_d + \alpha C^{-1}\right)v.$$

Since the function ℓ_n is twice continuously differentiable, then f is also twice continuously differentiable. Since both the gradient ∇f and Hessian H_f are continuous and $\nabla f(\beta^*) = 0$, we use a Taylor expansion to obtain the upper bound

$$\begin{aligned} f(\beta) &= f(\beta^*) + \int_0^1\int_0^t (\beta-\beta^*)^TH_f(\beta^* + s(\beta-\beta^*))(\beta-\beta^*)dsdt \\ &\leq f(\beta^*) + \frac{1}{2}(\beta-\beta^*)^T\left(r_0\lambda_{max}(X^TX)I_d + \alpha C^{-1}\right)(\beta-\beta^*). \end{aligned}$$

We then have a lower bound on the normalizing constant of the target posterior

$$Z_{\Pi(\cdot|X,Y)} = \int_{\mathbb{R}^d}\exp\left(-f(\beta)\right)d\beta \geq \frac{\exp\left(-f(\beta^*)\right)(2\pi)^{d/2}}{\det\left(r_0\lambda_{max}(X^TX)I_d + \alpha C^{-1}\right)^{1/2}}.$$

This in turn yields a lower bound on the ratio

$$\frac{Z_{\Pi(\cdot|X,Y)}}{Z_Q}\exp\left(f(\beta^*)\right) \geq \frac{\det\left(\alpha C^{-1}\right)^{1/2}}{\det\left(r_0\lambda_{max}(X^TX)I_d + \alpha C^{-1}\right)^{1/2}}. \tag{A.11}$$

The matrix C is symmetric and positive-definite and so its eigenvalues exist and are positive. Let $(\lambda_i(C))_{i=1}^d$ be the eigenvalues of C. It is readily verified that the eigenvalues of the matrix $r_0\lambda_{max}(X^TX)I_d+\alpha C^{-1}$ exist and are $\left(r_0\lambda_{max}(X^TX) + \frac{\alpha}{\lambda_i(C)}\right)_{i=1}^d$. Then

$$\begin{aligned}
&\frac{\det\left(\alpha C^{-1}\right)}{\det\left(r_0\lambda_{max}(X^TX)I_d + \alpha C^{-1}\right)} \qquad (A.12)\\
&= \frac{\prod_{i=1}^d \frac{\alpha}{\lambda_i(C)}}{\prod_{i=1}^d \left(r_0\lambda_{max}\left(X^TX\right) + \frac{\alpha}{\lambda_i(C)}\right)}\\
&= \prod_{i=1}^d \frac{\frac{\alpha}{\lambda_i(C)}}{r_0\lambda_{max}\left(X^TX\right) + \frac{\alpha}{\lambda_i(C)}}\\
&= \prod_{i=1}^d \frac{1}{\frac{r_0}{\alpha}\lambda_{max}\left(X^TX\right)\lambda_i(C) + 1}\\
&= \exp\left(-\sum_{i=1}^d \log\left(\frac{r_0}{\alpha}\lambda_{max}\left(X^TX\right)\lambda_i(C) + 1\right)\right). \qquad (A.13)
\end{aligned}$$

We have the basic inequality $\log(x+1) \leq x$ for any $x \in [0, +\infty)$. Since the eigenvalues of C are positive and $\lambda_{max}\left(X^TX\right)$ is nonnegative, we have the upper bound

$$\sum_{i=1}^d \log\left(\frac{r_0}{\alpha}\lambda_{max}\left(X^TX\right)\lambda_i(C) + 1\right) \leq \frac{r_0}{\alpha}\lambda_{max}\left(X^TX\right)\sum_{i=1}^d \lambda_i(C). \tag{A.14}$$

This yields a lower bound on (A.13). Define the doubly-indexed sequence $(a_{d,n})$ by

$$a_{d,n} = \frac{r_0}{2\alpha}\lambda_{max}\left(X^T X\right)\sum_{i=1}^{d}\lambda_i(C).$$

We have then shown that

$$\frac{Z_{\Pi(\cdot|X,Y)}}{Z_Q}\exp\left(f(\beta^*)\right) \geq \exp\left(-a_{d,n}\right). \tag{A.15}$$

By our assumption, $tr(C) \to s_0$ as $d \to \infty$. That is to say that

$$\lim_{d\to+\infty}\sum_{i=1}^{d}\lambda_i(C) = s_0.$$

Then as $d, n \to \infty$ in such a way that $d/n \to \gamma \in (0, +\infty)$, by continuity

$$a_{d,n} \to \frac{r_0}{2\alpha}(1+\gamma^{1/2})^2\sigma^2 s_0$$

almost surely. This implies using continuity that almost surely,

$$\lim_{\substack{d,n\to\infty \\ \frac{d}{n}\to\gamma}}(1-\exp\left(-a_{n,d}\right))^t = (1-\exp\left(-a_0\right))^t.$$

By Corollary 1, we have the upper bound on the Wasserstein distance for each d and each n:

$$\begin{aligned}
&\mathcal{W}_\rho\left(P^t(\beta^*,\cdot),\Pi(\cdot|X,Y)\right) \\
&= \left(1-\exp(f(\beta^*))\frac{Z_{\Pi(\cdot|X,Y)}}{Z_Q}\right)^t\int_{\mathbb{R}^d}\rho(\beta,\beta^*)d\Pi(\beta|X,Y) \\
&\leq (1-\exp\left(-a_{n,d}\right))^t\int_{\mathbb{R}^d}\rho(\beta,\beta^*)d\Pi(\beta|X,Y).
\end{aligned}$$

Suppose that $\limsup_{\substack{d,n\to\infty \\ \frac{d}{n}\to\gamma}} \int_{\mathbb{R}^d} \rho(\beta,\beta^*)d\Pi(\beta|X,Y) < \infty$. Using properties of the limit superior,

$$
\begin{aligned}
&\limsup_{\substack{d,n\to\infty \\ \frac{d}{n}\to\gamma}} \mathcal{W}_\rho\left(P^t(\beta^*,\cdot),\Pi(\cdot|X,Y)\right) \\
&\leq \limsup_{\substack{d,n\to\infty \\ \frac{d}{n}\to\gamma}} (1-\exp\left(-a_{n,d}\right))^t \limsup_{\substack{d,n\to\infty \\ \frac{d}{n}\to\gamma}} \int_{\mathbb{R}^d} \rho(\beta,\beta^*)d\Pi(\beta|X,Y) \\
&= \lim_{\substack{d,n\to\infty \\ \frac{d}{n}\to\gamma}} (1-\exp\left(-a_{n,d}\right))^t \limsup_{\substack{d,n\to\infty \\ \frac{d}{n}\to\gamma}} \int_{\mathbb{R}^d} \rho(\beta,\beta^*)d\Pi(\beta|X,Y) \\
&= (1-\exp\left(-a_0\right))^t \limsup_{\substack{d,n\to\infty \\ \frac{d}{n}\to\gamma}} \int_{\mathbb{R}^d} \rho(\beta,\beta^*)d\Pi(\beta|X,Y).
\end{aligned}
$$

The other case when $\limsup_{\substack{d,n\to\infty \\ \frac{d}{n}\to\gamma}} \int_{\mathbb{R}^d} \rho(\beta,\beta^*)d\Pi(\beta|X,Y) = +\infty$ is trivial.