J. Appl. Probab. **61,** 33–54 (2024)
*doi:*10.1017/jpr.2023.21

EXACT CONVERGENCE ANALYSIS FOR METROPOLIS–HASTINGS INDEPENDENCE SAMPLERS IN WASSERSTEIN DISTANCES

AUSTIN BROWN,* *University of Warwick*
GALIN L. JONES,** *University of Minnesota*

Abstract

Under mild assumptions, we show that the exact convergence rate in total variation is also exact in weaker Wasserstein distances for the Metropolis–Hastings independence sampler. We develop a new upper and lower bound on the worst-case Wasserstein distance when initialized from points. For an arbitrary point initialization, we show that the convergence rate is the same and matches the convergence rate in total variation. We derive exact convergence expressions for more 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 that 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.

Keywords: Bayesian statistics; convergence analysis; convergence rate lower bounds; computational complexity; Markov chain Monte Carlo; Metropolis–Hastings

2020 Mathematics Subject Classification: Primary 60J05; 60J22
Secondary 60J20

1. Introduction

Applications in modern statistics often require generating Monte Carlo samples from a distribution defined on $\Theta \subseteq \mathbb{R}^d$ using a version of the Metropolis–Hastings algorithm [4, 16, 30]. Popular versions of Metropolis–Hastings include, among many others, random walk Metropolis–Hastings (RWM), the Metropolis-adjusted Langevin algorithm (MALA), the Metropolis–Hastings independence (MHI) sampler, and Hamiltonian Monte Carlo.

Convergence analyses of general state space Metropolis–Hastings algorithms have traditionally focused on studying their convergence rates in total variation distances [31, 44, 48]. 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 [22] and the validity of methods for assessing the reliability of the simulation effort [43, 49]. However, convergence analyses of Metropolis–Hastings Markov chains typically result in *qualitative* convergence rates [15, 18, 21, 29, 42]. Exact convergence rates in total variation in the general state space setting have been nonexistent until recently and these are for MHI samplers [46, 52]. Moreover,

Received 12 November 2022; revision received 22 March 2023.

* Postal address: Department of Statistics, University of Warwick, Coventry, UK. Email: austin.d.brown@warwick.ac.uk
** Postal address: School of Statistics, University of Minnesota, Minneapolis, MN, USA. Email: galin@umn.edu

exact convergence rates in total variation for MHI samplers have been studied only in trivial examples. This leaves practitioners with little guidance on the true convergence behavior and reliability of MHI samplers in practical applications, especially in high dimensions.

At the same time, there has been significant recent interest in the convergence properties of Monte Carlo Markov chains in high-dimensional settings [8, 11, 15, 20, 37, 41, 53] and traditional approaches can have limitations in this regime [38]. This has led to an interest in considering the convergence of Monte Carlo Markov chains using Wasserstein distances [13, 15, 19, 28, 39, 40] which may scale to large problem sizes where other approaches have had difficulties [7, 15, 40]. 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 [15, 19, 23, 26].

We study exact convergence rates of the MHI sampler in L_1-Wasserstein distances, which we refer to as just Wasserstein distances. There has been previous successful convergence analysis of Metropolis–Hastings algorithms using specific Wasserstein distances [7, 15]. We develop exact convergence rates which are universal across many different metrics used in Wasserstein distances for the MHI sampler, simultaneously. Under mild assumptions, we show that the exact convergence rate in total variation [52] is also exact for Wasserstein distances weaker than total variation for every initialization. We provide a new upper and lower bound on the worst-case Wasserstein distance when initialized from points. Under mild assumptions, similar to the ones used for the result in total variation [52], we show that the convergence rate at any point initialization is the same as the worst-case convergence rate. When the algorithm is started at a specific point, we give exact convergence expressions across more general Wasserstein distances, possibly stronger than total variation.

Our theoretical results on the exact convergence rate extend the results in total variation [52] to Wasserstein distances. However, only a trivial example was studied in total variation [52]. We 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. Previously, qualitative convergence results for a Gibbs sampler were developed [25].

Compared to methods used to approximate integrals such as importance sampling, MHI samplers can generate samples from the target distribution, which is often of interest for practitioners. MHI samplers can also be computationally efficient at each iteration, in contrast to more sophisticated Markov chain Monte Carlo algorithms, but can require many iterations to accept a proposed sample. Connections between the MHI sampler and rejection sampling are also well known [27, 48]. Exact convergence rates for MHI samplers may also provide insight into the convergence rates of more popular Metropolis–Hastings algorithms such as the MALA and RWM algorithms [5].

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 [35, 45, 48], 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 well-known condition [29, 48, 52], but 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 [1, 36], 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 centered MHI sampler for general Wasserstein distances, independent of the dimension d and sample size n, when they increase in such a way that the ratio $d/n \to \gamma \in (0, +\infty)$. In this case, we show informative convergence rates for practitioners for the MHI sampler 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 RWM on a compact set was studied [2]. 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 MALA and RWM for certain log-concave target distributions [9, 10]. We take into account the sample size, and upper bound the convergence rate to provide further theoretical guarantees for time averages of the Markov chain [22, 23].

Some related quantitative convergence complexity bounds look at the spectral gap, implying a convergence rate in total variation from a specific distribution initialization. For example, the convergence rate of a random walk algorithm for logistic regression in one dimension has been studied in terms of the sample size [20]. Other related results concern the convergence properties of some high-dimensional Gibbs samplers [33, 34] or the convergence properties of Gibbs samplers when the dimension or the sample size increase individually [11, 40, 41].

The remainder is organized as follows. In Section 2, we define the Metropolis–Hastings independence sampler and the Wasserstein distance. In Section 3, we develop exact convergence rates in the Wasserstein distance for the MHI sampler and apply this theory to Bayesian quantile regression. In Section 5, 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 6 contains some final remarks. Some technical details and proofs are deferred to the appendices.

2. MHI samplers and Wasserstein distances

As they will be considered here, MHI samplers 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

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

We consider MHI samplers 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

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

https://doi.org/10.1017/jpr.2023.21 Published online by Cambridge University Press

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

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

For $\theta \in \mathbb{R}^d$, define the Markov kernel at iteration time $t \geq 2$ recursively by

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

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 [24, 50, 51], which we will call simply the Wasserstein distance, is

$$\mathcal{W}_\rho(P^t(\theta, \cdot), \Pi) = \inf_{\xi \in \mathcal{C}(P^t(\theta, \cdot), \Pi)} \int \rho(\theta', \omega) \, \mathrm{d}\xi(\theta', \omega).$$

Notice that when the metric ρ is $\rho(\theta, \omega) = I_{\theta \neq \omega}$, then the Wasserstein distance is the total variation distance. More generally, for a lower semicontinuous function $V \geq 1$, $\rho(\theta, \omega) = [V(\theta) + V(\omega)] I_{\theta \neq \omega}$ defines a weighted total variation distance. Another example is $\rho(\theta, \omega) = \min\{\|\theta - \omega\|, 1\}$, which is always less than the Hamming metric used in total variation. More general L_p-Wasserstein distances with $p \geq 2$ are not studied in this work.

Wasserstein distances control the bias of integrals of $P^t(\theta, \cdot)$ over all ρ-Lipschitz functions, whereas total variation controls the bias for bounded, measurable functions. Often in applications, various types of Lipschitz functions are of interest. Consider, for example, functions such as the identity function and Winsorized functions such as $g(\omega) = (-R) \vee (R \wedge \omega)$ where $R \in (0, +\infty)$, which reduce sensitivity to extreme values. An alternative motivation for using Wasserstein distances is that convergence analysis in certain Wasserstein distances may improve the scaling to high dimensions [15, 39, 40]. Serious problems can arise with existing convergence analysis in total variation [48] in high dimensions even for seemingly trivial examples such as the one below.

Example 1. Consider a standard $d-$dimensional Gaussian target distribution $\Pi \equiv N(0, I_d)$ and Gaussian proposal $Q \equiv N(0, \sigma^2 I_d)$ with $\sigma^2 \in (1, \infty)$. With $\rho_{\mathrm{TV}}(\theta', \omega') = I_{\theta' \neq \omega'}$, current results on convergence analysis [48] show that, for every $\theta \in \mathbb{R}^d$, the ratio of the proposal and target densities $\inf_\theta \{q(\theta)/\pi(\theta)\} \geq \sigma^{-d}$ and $\mathcal{W}_{\rho_{\mathrm{TV}}}(P^t(\theta, \cdot), \Pi) \leq (1 - \sigma^{-d})^t$. In particular, $\sigma^{-d} \approx 0$ in high dimensions.

At a specific initialization, we may look for a smaller convergence rate which may scale to high dimensions in a Wasserstein distance weaker than total variation. We will not only develop an exact convergence analysis which controls the bias for Lipschitz functions, but, roughly speaking, we also discover that the convergence rate in Example 1 is exact for Wasserstein distances and the same for every initialization. Nevertheless, we show that convergence rates can scale to large problem sizes using a novel proposal and exact convergence analysis.

3. Exact convergence rates for MHI samplers

When the ratio of the proposal and target densities is bounded below by a positive number, i.e. $\varepsilon^* = \inf_{\theta \in \Theta} \{q(\theta)/\pi(\theta)\} > 0$, the MHI sampler is uniformly ergodic in total variation with convergence rate upper bounded by $1-\varepsilon^*$ [48, 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 [29, Theorem 2.1]. More recently, it was shown that the convergence rate cannot be improved [52, Theorem 1]. We show this to be the case even in weaker Wasserstein distances where the lower bound does not follow trivially from that of the total variation lower bound [52, Theorem 1].

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

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

If, in addition, q is lower semicontinuous on Θ, π is upper semicontinuous on Θ, and Θ can be expressed as a countable union of compact sets, then

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

Proof. The proof is provided in Appendix A. □

The semicontinuity assumption is not required when working with the total variation distance [52, Theorem 1], but it is a mild assumption that holds in many practical applications. The upper bound constant can improve upon upper bounds in total variation [48] if, for example, ρ is continuous and Θ is compact. If $\varepsilon^* = 0$, Theorem 1 also gives the lower bound

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

which shows that 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 sampler in many Wasserstein distances.

Exact convergence expressions are available when the Markov chain is initialized at $\theta^* = \operatorname{argmin}\{q(\theta)/\pi(\theta) : \theta \in \Theta\}$ using techniques from [52].

Proposition 1. *Suppose there exists a solution* $\theta^* = \operatorname{argmin}\{q(\theta)/\pi(\theta) : \theta \in \Theta\}$. *Then*

$$\mathcal{W}_\rho(P^t(\theta^*, \cdot), \Pi) = (1 - q(\theta^*)/\pi(\theta^*))^t \int \rho(\theta, \theta^*)\, \mathrm{d}\Pi(\theta).$$

Proof. Define $\varepsilon_{\theta^*} = q(\theta^*)/\pi(\theta^*) = \varepsilon^*$. 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^* [52, Remark 1, Theorem 2], that is,

$$P^t(\theta^*, \cdot) = (1-(1-\varepsilon_{\theta^*})^t)\Pi + (1-\varepsilon_{\theta^*})^t \delta_{\theta^*}.$$

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

$$\int_\Theta \psi\, \mathrm{d}P^t(\theta^*, \cdot) = (1-(1-\varepsilon_{\theta^*})^t) \int_\Theta \psi\, \mathrm{d}\Pi + (1-\varepsilon_{\theta^*})^t \psi(\theta^*).$$

Since the only coupling between Π^* and the Dirac measure δ_{θ^*} is the independent coupling [14], the Wasserstein distance takes the simple form $\mathcal{W}_\rho(\delta_{\theta^*}, \Pi) = \int \rho(\theta, \theta^*)\, \mathrm{d}\Pi(\theta)$.

https://doi.org/10.1017/jpr.2023.21 Published online by Cambridge University Press

Since q is not exactly π, $\varepsilon_{\theta^*} \in (0, 1)$. 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 [50, Theorem 1.14],

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

3.1. Application: Bayesian quantile regression

Fix $r \in (0, 1)$ and suppose, for $i = 1, \ldots, n$, that ε_i are independent and identically distributed (i.i.d.) with density $p_r(\varepsilon) = r(1-r)(\exp((1-r)\varepsilon)I_{\varepsilon<0} + \exp(-r\varepsilon)I_{\varepsilon\geq 0})$. Let $v_0, s_0 \in (0, \infty)$ and $C \in \mathbb{R}^{d\times d}$ be symmetric positive-definite. We parameterize the inverse gamma distribution so that if $\sigma \sim \mathrm{IG}(v, s)$ for some $v, c \in (0, \infty)$, 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

$$\sigma \sim \mathrm{IG}(v_0, s_0), \qquad \beta \mid \sigma \sim \mathrm{N}_d(0, \sigma C), \qquad Y_i = \beta^\top X_i + \sigma\varepsilon_i.$$

Let $\Pi(\cdot \mid X, Y)$ denote the posterior and $\pi(\cdot \mid X, Y)$ 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 [25] in this setting. We will study the MHI sampler with a normal–inverse-gamma proposal constructed as follows. Define the convex function by $\ell_r(u) = u(r - I_{u<0})$ and $s_{n,r} : \mathbb{R}^d \to \mathbb{R}$ by $s_{n,r}(\beta) = \sum_{i=1}^n \ell_r(Y_i - \beta^\top X_i) + \beta^\top C^{-1}\beta/2$. 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

$$\sigma \sim \mathrm{IG}(n + v_0, s_0 + s_{n,r}(\beta^*)), \qquad \beta \mid \sigma \sim \mathrm{N}_d(\beta^*, \sigma C).$$

Let $\Gamma : (0, \infty) \to \mathbb{R}$ be the usual Gamma function and define

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

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

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

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

https://doi.org/10.1017/jpr.2023.21 Published online by Cambridge University Press

Proof. 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 $\pi(\beta, \sigma \mid X, Y) = Z^{-1}_{\Pi(\cdot|X,Y)} \exp(-f(\beta, \sigma))$. Since the function $\beta \mapsto s_{n,r}(\beta) - \beta^\top C^{-1}\beta/2$ is a convex function on $\mathbb{R}^d$, by Lemma 4, for every $\beta \in \mathbb{R}^d$, $s_{n,r}(\beta) \geq s_{n,r}(\beta^*) + \frac{1}{2}(\beta - \beta^*)^\top C^{-1}(\beta - \beta^*)$. For any $(\beta, \sigma) \in \mathbb{R}^d \times (0, \infty)$, we then have the lower bound

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

This implies that

$$f(\beta, \sigma) - \frac{1}{2\sigma}(\beta - \beta^*)^\top C^{-1}(\beta - \beta^*) - \frac{s_0 + s_{n,r}(\beta^*)}{\sigma} - (n + v_0 + 1 + d/2) \log(\sigma) \geq 0.$$

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 that

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

An application of Proposition 1 completes the proof. □

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 and probit regression.

4. The convergence rate at arbitrary initializations

The previous section studies the worst-case convergence rate and the convergence rate at an individual point for the MHI sampler. We can study the convergence rate at every point, as was done for total variation [52]. The technique needed to prove this relies on the exact representation of the MHI sampler [46, Theorem 1], but new techniques are needed to show this in the Wasserstein distance. Similar to the convergence rate in total variation [52], we define the Wasserstein convergence rate for a point $\theta \in \Theta$ as $r_\rho(\theta) = \lim_{t\to\infty} \mathcal{W}_\rho(P^t(\theta, \cdot), \Pi)^{1/t}$.

When the distance metric is $\min\{\|\cdot\|, 1\}$ where $\|\cdot\|$ can be any norm, Theorem 3 shows we can obtain the convergence rate at every point under mild conditions. We require π and q to be locally $\|\cdot\|$-Lipschitz continuous and bounded, which is stronger than only locally Lipschitz as in total variation [52]. However, this additional condition is satisfied in many practical applications in statistics. Theorem 3 also lower bounds the convergence rate for Wasserstein L_p-distances, and the rate of convergence for these distances cannot be improved.

Theorem 3. *If π, q are locally $\|\cdot\|$-Lipschitz continuous and bounded on $\mathbb{R}^d$ and there exists a solution $\theta^* = \operatorname{argmin}\{q(\theta)/\pi(\theta) : \theta \in \Theta\}$, then, for any $\theta \in \Theta$, the Wasserstein convergence rate is the same with $r_{\min\{\|\cdot\|,1\}}(\theta) = 1 - q(\theta^*)/\pi(\theta^*)$.*

https://doi.org/10.1017/jpr.2023.21 Published online by Cambridge University Press

Proof. If the initialization is at θ^*, then the result follows from Proposition 1. Fix a point $\theta_0 \in \Theta$ such that $\theta_0 \neq \theta^*$. Using Lemma 1, we have an upper bound on the convergence rate by

$$\limsup_{t\to\infty} \mathcal{W}_{\min\{\|\cdot\|,1\}}(P^t(\theta_0, \cdot), \Pi)^{1/t} \leq 1 - q(\theta^*)/\pi(\theta^*).$$

It remains to lower bound the limit inferior.

Denote the standard p-norms for vectors $x \in \mathbb{R}^d$ by $\|x\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$. For $h \in (0, 1]$, define the function $\varphi_h(\theta) = (2h)^{-d} \exp(-h^{-1} \|\theta - \theta^*\|_1)$. This is the probability density function for a Laplace distribution and φ_h is $2^{-d}h^{-d-1}$ $\|\cdot\|_1$-Lipschitz, and so $2^d h^{d+1}\varphi_h(\theta) = h \exp(-h^{-1} \|\theta - \theta^*\|_1)$ is nonnegative, $\|\cdot\|_1$-Lipschitz with constant 1 and bounded by 1. In particular, it is readily shown that $2^d h^{d+1}\varphi_h(\theta)$ is $\min\{1, \|\cdot\|_1\}$-Lipschitz. Using Kantorovich–Rubinstein duality [50, Theorem 1.14], we have the lower bound

$$\mathcal{W}_{\min\{\|\cdot\|_1,1\}}(P^t(\theta_0, \cdot), \Pi) \geq 2^d h^{d+1}\left[\int \varphi_h \, \mathrm{d}\Pi - \int \varphi_h \, \mathrm{d}P^t(\theta_0, \cdot)\right]. \tag{1}$$

We will develop some approximation properties of φ_h. Since $\theta_0 \neq \theta^*$, it is readily shown that $\lim_{h\downarrow 0} \varphi_h(\theta_0) = 0$. Using a change of variables and since we have assumed $\sup_\theta \pi(\theta) < \infty$,

$$\left|\int \varphi_h(\theta')\pi(\theta') \, \mathrm{d}\theta' - \pi(\theta^*)\right| \leq \int_{\|\theta'\|_2 \leq t} |\pi(\theta^* + h\theta') - \pi(\theta^*)| 2^{-d} \exp(-\left\|\theta'\right\|_1) \, \mathrm{d}\theta' \tag{2}$$

$$+ 2 \sup_\theta \pi(\theta) \int_{\|\theta'\|_2 > t} 2^{-d} \exp(-\left\|\theta'\right\|_1) \, \mathrm{d}\theta'. \tag{3}$$

If $Y_1, \ldots, Y_d$ are i.i.d. Laplace, then we have the tail bound

$$\mathbb{P}(\|Y\|_2 \geq t) \leq \mathbb{P}\left(\max_i |Y_i| \geq \frac{t}{\sqrt{d}}\right) \leq \sum_{i=1}^d \mathbb{P}\left(|Y_i| \geq \frac{t}{\sqrt{d}}\right) \leq d \exp\left(-\frac{t}{\sqrt{d}}\right).$$

Since norms are equivalent on $\mathbb{R}^d$, π and q are locally Lipschitz with respect to any norm. Choosing $h = h_0/t^2$ for some $h_0 \in (0, 1)$, since π is locally $\|\cdot\|_2$-Lipschitz, we can find a universal constant $L \in (0, \infty)$ such that, for any $\left\|\theta'\right\|_2 \leq t$, $|\pi(\theta^* + h\theta') - \pi(\theta^*)| \leq h_0 L/t$. Applying these upper bounds to (2) and (3), for large enough t we have

$$\left|\int \varphi_h(\theta')\pi(\theta') \, \mathrm{d}\theta' - \pi(\theta^*)\right| \leq \frac{2h_0 L}{t}. \tag{4}$$

A similar argument with the assumptions on q yields, for large enough t,

$$\left|\int \varphi_h(\theta')q(\theta') \, \mathrm{d}\theta' - q(\theta^*)\right| \leq \frac{2h_0 L}{t}. \tag{5}$$

It remains to lower bound (1). We will use the exact representation of the independence sampler [46, Theorem 1, Lemma 3]. Define the importance sampling weight by $w(\theta) = \pi(\theta)/q(\theta)$ and its maximum $w^* = w(\theta^*) = \pi(\theta^*)/q(\theta^*)$. For $w \in (0, \infty)$, define

$$\lambda(w) = \int_{w(\theta') \leq w} \left[1 - \frac{w(\theta')}{w}\right] q(\theta') \, \mathrm{d}\theta', \qquad T_t(w) = \int_w^\infty \frac{t\lambda^{t-1}(v)}{v^2} \, \mathrm{d}v,$$

https://doi.org/10.1017/jpr.2023.21 Published online by Cambridge University Press

and, using the exact representation of the independence sampler [46, Theorem 1, Lemma 3], for measurable sets $B \subseteq \mathbb{R}^d$,

$$P^t(\theta_0, B) = \int_B T_t(\max\{w(\theta_0), w(\theta')\})\pi(\theta')\, \mathrm{d}\theta' + \lambda^t(w(\theta_0))\delta_{\theta_0}(B).$$

The proof of existing results [52, Theorem 4] shows that

$$T_t(w) \le 1 + (1 - 1/w^*)^t \left[\frac{t}{w^* - 1}\left(\frac{w^*}{w} - 1\right) - 1\right].$$

We now use the exact representation of the independence sampler to lower bound (1). We have $\lambda^t(w(\theta_0)) \le (1 - 1/w^*)^t$, so we then have the upper bound

$$\begin{aligned}\int \varphi_h\, \mathrm{d}P^t(\theta_0, \cdot) - \int \varphi_h\, \mathrm{d}\Pi &\le -(1 - 1/w^*)^t \int \varphi_h\, \mathrm{d}\Pi + \lambda^t(w(\theta_0))\varphi_h(\theta_0) \\ &\quad + (1 - 1/w^*)^t \frac{t}{w^* - 1} \int \left[\frac{w^*}{\max\{w(\theta_0), w(\theta')\}} - 1\right]\varphi_h(\theta')\pi(\theta')\, \mathrm{d}\theta' \\ &\le (1 - 1/w^*)^t \left[\varphi_h(\theta_0) - \int \varphi_h\, \mathrm{d}\Pi\right] \\ &\quad + (1 - 1/w^*)^t \frac{t}{w^* - 1}\left[w^* \int \varphi_h(\theta')q(\theta')\, \mathrm{d}\theta' - \int \varphi_h(\theta')\pi(\theta')\, \mathrm{d}\theta'\right].\end{aligned}$$

Using this upper bound with the approximations (4) and (5) yields

$$\begin{aligned}\int \varphi_h\, \mathrm{d}P^t(\theta_0, \cdot) - \int \varphi_h\, \mathrm{d}\Pi &\le (1 - 1/w^*)^t \left[\varphi_h(\theta_0) - \pi(\theta^*) + \frac{2h_0L}{t}\right] \\ &\quad + (1 - 1/w^*)^t \frac{t}{w^* - 1}\left[w^* q(\theta^*) - \pi(\theta^*) + (w^* + 1)\frac{2h_0L}{t}\right] \\ &\le (1 - 1/w^*)^t \left[\varphi_h(\theta_0) - \pi(\theta^*) + \frac{2h_0L}{t} + \frac{w^* + 1}{w^* - 1}2h_0L\right].\end{aligned}$$

Therefore, we can choose a small enough $h_0 \in (0, 1)$ independently of t such that we have the upper bound

$$\int \varphi_h\, \mathrm{d}P^t(\theta_0, \cdot) - \int \varphi_h\, \mathrm{d}\Pi \le -\frac{\pi(\theta^*)}{4}(1 - 1/w^*)^t.$$

Applying these bounds to (1) and using that we have chosen $h = h_0/t^2$, we have the lower bound

$$\begin{aligned}\mathcal{W}_{\min\{\|\cdot\|_1, 1\}}(P^t(\theta_0, \cdot), \Pi) &\ge 2^d h^{d+1}\left[\int \varphi_h\, \mathrm{d}\Pi - \int \varphi_h\, \mathrm{d}P^t(\theta_0, \cdot)\right] \\ &\ge 2^{d-2}h^{d+1}\pi(\theta^*)(1 - 1/w^*)^t \\ &\ge 2^{d-2}\left(\frac{h_0}{t^2}\right)^{d+1}\pi(\theta^*)(1 - 1/w^*)^t.\end{aligned}$$

Taking the limit,

$$\liminf_{t\to\infty} \mathcal{W}_{\min\{\|\cdot\|_1, 1\}}(P^t(\theta_0, \cdot), \Pi)^{1/t} \ge 1 - 1/w^* = 1 - q(\theta^*)/\pi(\theta^*).$$

Since all norms are equivalent on $\mathbb{R}^d$, this can be extended to any norm $\|\cdot\|$. □

https://doi.org/10.1017/jpr.2023.21 Published online by Cambridge University Press

5. MHI samplers with centered Gaussian proposals

We look to apply the exact convergence expression from Proposition 1 to practical applications since the convergence rate is the same at every initialization under mild assumptions. Recently, centered drift functions have been used to improve convergence analyses of some Monte Carlo Markov chains [11, 37, 40]. 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

$$\varepsilon_{\theta^*} = (2\pi)^{-d/2}\alpha^{d/2} \det(C)^{-1/2} Z_\Pi \exp(f(\theta^*)).$$

Proposition 2. *If θ^* exists and, for any $\theta \in \mathbb{R}^d$, $f(\theta) \geq f(\theta^*) + \alpha(\theta - \theta^*)^\top C^{-1}(\theta - \theta^*)/2$, then*

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

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

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

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 assumption in Proposition 2 is sometimes referred to as a *quadratic growth condition* at θ^*. This condition does not require convexity of f, but is satisfied and the point θ^* is guaranteed to exist if the function f satisfies a strong convexity property. A function $h : \mathbb{R}^d \to \mathbb{R}$ is strongly convex with parameter μ if there is a $\mu \in (0, +\infty)$ such that $h(\cdot) - \mu \left\|\cdot\right\|^2 /2$ is convex [17, 32]. 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 [32]. For a point $\theta_0 \in \mathbb{R}^d$, we consider the proposal corresponding to a d-dimensional Gaussian distribution, $\mathrm{N}_d(\theta_0 - \alpha^{-1}Cg_{f(\theta_0)}, \alpha^{-1}C)$. When f is differentiable, this construction of the proposal uses the gradient of f in a similar way to MALA. The ratio of the proposal and target density evaluated at θ_0 is $\varepsilon_{\theta_0} = (2\pi)^{-d/2} \det(\alpha^{-1}C)^{-1/2} Z_\Pi \exp(f(\theta_0) - {g_{f(\theta_0)}}^\top Cg_{f(\theta_0)}/(2\alpha))$.

Choosing $\theta_0 \equiv \theta^*$ maximizes the convergence rate and yields the centered Gaussian proposal, but we also have an exact convergence expression in other cases.

Proposition 3. *If the function $\theta \mapsto f(\theta) - \alpha\theta^\top C^{-1}\theta/2$ is convex for all points on $\mathbb{R}^d$, then $\mathcal{W}_\rho(P^t(\theta_0, \cdot), \Pi) = (1 - \varepsilon_{\theta_0})^t \int \rho(\theta, \theta_0)\, \mathrm{d}\Pi(\theta)$.*

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

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

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

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

An application of Proposition 1 completes the proof. □

5.1. 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 are conditionally independent with $Y_i \mid X_i, \beta \sim \text{Poisson}(\exp(\beta^\top X_i))$. Similarly, for negative-binomial regression, if $\xi \in (0, +\infty)$, assume $Y_i \mid X_i, \beta \sim \text{Negative-Binomial}(\xi, (1 + \exp(-\beta^\top X_i))^{-1})$. For binary response regression, if $S : \mathbb{R} \to (0, 1)$, assume $Y_i \mid X_i, \beta \sim \text{Bernoulli}(S(\beta^\top X_i))$. 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 $\beta \sim \mathrm{N}_d(0, \alpha^{-1}C)$, 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 $Y = (Y_1, \ldots, Y_n)^\top$ and the matrix $X = (X_1, \ldots, X_n)^\top$. Let $\Pi(\cdot \mid X, Y)$ denote the posterior with density $\pi(\cdot \mid X, Y)$. If ℓ_n denotes the negative log-likelihood for each model, the posterior density is characterized by

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

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

Corollary 1. $\mathcal{W}_\rho(P^t(\beta^*, \cdot), \Pi(\cdot \mid X, Y)) = (1 - \varepsilon_{\beta^*})^t \int \rho(\beta, \beta^*)\, \mathrm{d}\Pi(\beta \mid X, Y)$, *where* $\varepsilon_{\beta^*} = \exp(\ell_n(\beta^*) + (\alpha/2){\beta^*}^\top C^{-1}\beta^*)Z_{\Pi(\cdot|X,Y)}((2\pi)^{d/2}\det(\alpha^{-1}C)^{1/2})^{-1}$.

5.2. Convergence complexity analysis in binary response regression

For the MHI algorithm, we continue to use a centered proposal $\mathrm{N}_d(\beta^*, \alpha^{-1}C)$ and first consider a more general posterior density of the form $\pi(\beta \mid X, Y) \propto \exp(-\ell_n(\beta) - \alpha\beta^\top C^{-1}\beta/2)$ depending on data X, Y of size n. We also assume the limit of the trace of the covariance matrix used in our prior to be finite, i.e. $\mathrm{tr}(C) \to s_0 \in (0, +\infty)$ as $d \to +\infty$. Note that the trace of the covariance being finite is a necessary condition for Gaussian distributions to exist in an infinite-dimensional Hilbert space [3].

Theorem 4. *Suppose that the following conditions hold for a sequence d_n, $n \to \infty$:*

(i) The negative log-likelihood ℓ_n is a twice continuously differentiable convex function.

(ii) There is a universal constant $H_0 \in (0, +\infty)$ such that the largest eigenvalue of the Hessian of the negative log-likelihood H_{ℓ_n} satisfies, for every $\beta \in \mathbb{R}^d$, $\limsup_{d_n, n\to\infty} \lambda_{\max}(H_{\ell_n}(\beta)) \le H_0$.

Then $\limsup_{d_n,n\to\infty} \mathcal{W}_\rho(P^t(\beta^*, \cdot), \Pi(\cdot \mid X, Y)) \le M_0(1 - \exp(-H_0 s_0/(2\alpha)))^t$, *where* $M_0 = \limsup_{d_n,n\to\infty} \int \rho(\beta, \beta^*)\, \mathrm{d}\Pi(\beta \mid X, Y)$.

Proof. Define the function $f : \mathbb{R}^d \to \mathbb{R}$ by $f(\beta) = \ell_n(\beta) + \alpha\beta^\top C^{-1}\beta/2$, where ℓ_n is the negative log-likelihood loss function, and define $Z_Q = (2\pi)^{d/2} \det(\alpha^{-1}C)^{1/2}$. We first lower bound the quantity $\exp(f(\beta^*))Z_{\Pi(\cdot|X,Y)}/Z_Q$. For $\varepsilon \in (0, 1)$ and sufficiently large d_n, n, we have, for any $\beta \in \mathbb{R}^d$ and any $v \in \mathbb{R}^d$, $v^\top H_{\ell_n}(\beta)v \le (1+\varepsilon)H_0 \, \|v\|_2^2$. 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 $v^\top H_f(\beta)v \le v^\top((1+\varepsilon)H_0 I_d + \alpha C^{-1})v$. Since the function ℓ_n is twice continuously differentiable, 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

$$\begin{aligned} f(\beta) &= f(\beta^*) + \int_0^1 \int_0^t (\beta - \beta^*)^\top H_f(\beta^* + s(\beta - \beta^*))(\beta - \beta^*)\, \mathrm{d}s\, \mathrm{d}t \\ &\le f(\beta^*) + \tfrac{1}{2}(\beta - \beta^*)^\top((1+\varepsilon)H_0 I_d + \alpha C^{-1})(\beta - \beta^*). \end{aligned}$$

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

$$Z_{\Pi(\cdot|X,Y)} = \int_{\mathbb{R}^d} \exp(-f(\beta))\, \mathrm{d}\beta \ge \frac{\exp(-f(\beta^*))(2\pi)^{d/2}}{\det((1+\varepsilon)H_0 I_d + \alpha C^{-1})^{1/2}}.$$

This in turn yields a lower bound on the ratio

$$\frac{Z_{\Pi(\cdot|X,Y)}}{Z_Q} \exp(f(\beta^*)) \ge \frac{\det(\alpha C^{-1})^{1/2}}{\det((1+\varepsilon)H_0 I_d + \alpha C^{-1})^{1/2}}.$$

The matrix C is symmetric and positive-definite, 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

https://doi.org/10.1017/jpr.2023.21 Published online by Cambridge University Press

$(1+\varepsilon)H_0)I_d+\alpha C^{-1}$ exist and are $((1+\varepsilon)H_0+\alpha/\lambda_i(C))_{i=1}^d$. Then

$$\begin{aligned}\frac{\det(\alpha C^{-1})}{\det((1+\varepsilon)H_0 I_d+\alpha C^{-1})} &= \frac{\prod_{i=1}^d \alpha/\lambda_i(C)}{\prod_{i=1}^d ((1+\varepsilon)H_0+\alpha/\lambda_i(C))}\\ &= \prod_{i=1}^d \frac{\alpha/\lambda_i(C)}{(1+\varepsilon)H_0+\alpha/\lambda_i(C)}\\ &= \prod_{i=1}^d \frac{1}{(1/\alpha)(1+\varepsilon)H_0\lambda_i(C)+1}\\ &= \exp\left(-\sum_{i=1}^d \log\left(\frac{1}{\alpha}(1+\varepsilon)H_0\lambda_i(C)+1\right)\right). \end{aligned} \tag{6}$$

We have the basic inequality $\log(x+1)\le x$ for any $x\in[0,+\infty)$. Since the eigenvalues of C are positive and H_0 is nonnegative, we have the upper bound

$$\sum_{i=1}^d \log\left(\frac{1}{\alpha}(1+\varepsilon)H_0\lambda_i(C)+1\right)\le \frac{1}{\alpha}(1+\varepsilon)H_0\sum_{i=1}^d \lambda_i(C)$$

This yields a lower bound on (6). Define the doubly-indexed sequence $(a_{d,n})$ by

$$a_{d,n}=\frac{1}{2\alpha}(1+\varepsilon)H_0\sum_{i=1}^d \lambda_i(C).$$

We have then shown that

$$\frac{Z_{\Pi(\cdot|X,Y)}}{Z_Q}\exp(f(\beta^*))\ge \exp(-a_{d,n}).$$

By our assumption, $\operatorname{tr}(C)\to s_0$ as $d\to\infty$. That is to say, $\lim_{d\to+\infty}\sum_{i=1}^d \lambda_i(C)=s_0$. This implies, by using continuity,

$$\lim_{d_n,n\to\infty}(1-\exp(-a_{n,d}))^t=(1-\exp(-(1+\varepsilon)H_0 s_0/(2\alpha)))^t.$$

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

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

https://doi.org/10.1017/jpr.2023.21 Published online by Cambridge University Press

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

$$\begin{aligned}
&\limsup_{d_n, n\to\infty} \mathcal{W}_\rho(P^t(\beta^*, \cdot), \Pi(\cdot \mid X, Y)) \\
&\qquad\qquad \le \limsup_{d_n, n\to\infty} (1-\exp(-a_{n,d}))^t \limsup_{d_n, n\to\infty} \int_{\mathbb{R}^d} \rho(\beta, \beta^*)\, \mathrm{d}\Pi(\beta \mid X, Y) \\
&\qquad\qquad = \lim_{\substack{d,n\to\infty \\ d/n\to\gamma}} (1-\exp(-a_{n,d}))^t \limsup_{d_n, n\to\infty} \int_{\mathbb{R}^d} \rho(\beta, \beta^*)\, \mathrm{d}\Pi(\beta \mid X, Y) \\
&\qquad\qquad = (1-\exp(-(1+\varepsilon)H_0 s_0/(2\alpha)))^t \limsup_{d_n, n\to\infty} \int_{\mathbb{R}^d} \rho(\beta, \beta^*)\, \mathrm{d}\Pi(\beta \mid X, Y).
\end{aligned}$$

This holds for every $\varepsilon \in (0, 1)$, so taking the limit completes the proof in this case. The other case, when $\limsup_{d_n, n\to\infty} \int_{\mathbb{R}^d} \rho(\beta, \beta^*)\, \mathrm{d}\Pi(\beta \mid X, Y) = +\infty$, is trivial. □

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 assume that $(Y_i, X_i)_{i=1}^n$ are independent with $Y_i \mid X_i, \beta \sim \text{Bernoulli}(S(\beta^\top X_i))$ and $X_i \sim N_d(0, \sigma^2 n^{-1} I_d)$ with $\sigma^2 \in (0, +\infty)$. This scaling ensures the variance of the columns of the random matrix of features X is of fixed order; it is often used to ensure nondegeneracy in large-system limits of d, n [12]. Similar scaling assumptions on the data are used for high-dimensional maximum-likelihood theory in logistic regression [47].

Corollary 2. *Suppose the following conditions hold:*

(i) The negative log-likelihood ℓ_n is a twice continuously differentiable convex function.

(ii) 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 $\beta \in \mathbb{R}^d$, $\lambda_{\max}(H_{\ell_n}(\beta)) \le r_0 \lambda_{\max}(X^\top X)$.

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, $\limsup_{d/n\to\gamma} \mathcal{W}_\rho(P^t(\beta^*, \cdot), \Pi(\cdot \mid X, Y)) \le M_0(1-\exp(-a_0))^t$, where $M_0 = \limsup_{d/n\to\gamma} \int \rho(\beta, \beta^*)\, \mathrm{d}\Pi(\beta \mid X, Y)$.

Proof. 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 $X^\top X$ by $\lambda_{\max}(X^\top X)$. Therefore, as $d, n \to \infty$ in such a way that $d/n \to \gamma \in (0, +\infty)$,

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

almost surely [12, Theorem 1]. We have, for any $\beta \in \mathbb{R}^d$ and any $v \in \mathbb{R}^d$, $v^\top H_{\ell_n}(\beta) v \le r_0 \lambda_{\max}(X^\top X) \|v\|_2^2$. The proof follows from Theorem 4. □

Corollary 2 applies to both Bayesian logistic and probit regression. For logistic regression, ℓ_n is a twice continuously differentiable convex function and we may choose $r_0 = 4^{-1}$.

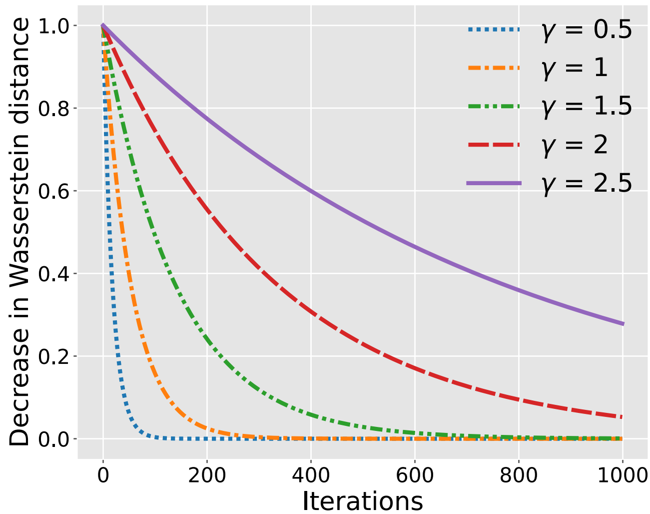

FIGURE 1. The limiting decrease in the Wasserstein distance using different values of γ, the limiting ratio of the dimension and sample size, versus the number of iterations.

Similarly, for probit regression, ℓ_n is also a twice continuously differentiable convex function and we may choose $r_0 = 1$ [6].

In Figure 1 we plot $(1 - \exp(-a_0))^t$, the limiting decrease in the Wasserstein distance according to our upper bound, with varying values of the limiting ratio γ and 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.

6. Final remarks

We have 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. We showed that the exact convergence rate at every initialization for Wasserstein distances weaker than the total variation distance is the same and matches that of the total variation convergence rate [52]. When starting at a certain point, we gave exact convergence expressions. 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. The centered MHI sampler 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 that this exact convergence rate may be upper bounded when sampling high-dimensional posteriors in Bayesian binary response regression.

Appendix A. Proof of Theorem 1

The proof will proceed by establishing the upper and lower bounds separately in Lemmas 1 and 2, respectively. This is done largely because the conditions for the upper bound are weaker than those for the lower bound.

https://doi.org/10.1017/jpr.2023.21 Published online by Cambridge University Press

The following definitions will be used in the proofs of Lemmas 1 and 2. First, for $\theta \in \Theta$, real-valued measurable functions f, and a Markov kernel K, we use the notation $K^t f(\theta) = \int f \, \mathrm{d}K^t(\theta, \cdot) = \int f(\theta') K^t(\theta, \mathrm{d}\theta')$ 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 1. *Let* $\varepsilon^* = \inf_{\theta\in\Theta}\{q(\theta)/\pi(\theta)\}$. *Then*

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

Proof. Let $\theta \in \Theta$ and let φ satisfy $\|\varphi\|_{\mathrm{Lip}(\rho)} \leq 1$. The existence of ε^* implies the minorization condition $P(\theta, \cdot) \geq \varepsilon^* \Pi(\cdot)$ [48, Corollary 4] which, in turn, ensures the residual kernel $R(\theta, \cdot) = [P(\theta, \cdot) - \varepsilon^* \Pi(\cdot)]/(1 - \varepsilon^*)$ is a Markov kernel with invariant distribution Π. It then follows that

$$\begin{aligned}\int \varphi \, \mathrm{d}P^t(\theta, \cdot) - \int \varphi \, \mathrm{d}\Pi &= (1 - \varepsilon^*)\left[\int R\varphi \, \mathrm{d}P^{t-1}(\theta, \cdot) - \int \varphi \, \mathrm{d}\Pi\right] \\ &= (1 - \varepsilon^*)\left[\int R\varphi \, \mathrm{d}P^{t-1}(\theta, \cdot) - \int R\varphi \, \mathrm{d}\Pi\right] \\ &\;\;\vdots \\ &= (1 - \varepsilon^*)^t\left[\int \varphi \, \mathrm{d}R^t(\theta, \cdot) - \int \varphi \, \mathrm{d}\Pi\right].\end{aligned}$$

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

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

Taking the supremum with respect to φ and using the Kantorovich–Rubinstein theorem [50, Theorem 1.14],

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

□

We now turn our attention to establishing the lower bound.

Lemma 2. *Let* $\varepsilon^* = \inf_{\theta\in\Theta}\{q(\theta)/\pi(\theta)\}$. *Suppose* q *is lower semicontinuous and* π *is upper semicontinuous on* Θ. *Suppose* Θ *can be expressed as a countable union of compact sets. If* $\rho(\cdot, \cdot) \leq 1$, *then* $\sup_{\theta\in\Theta} \mathcal{W}_\rho(P^t(\theta, \cdot), \Pi) \geq (1 - \varepsilon^*)^t \inf_{\theta\in\Theta} \int \rho(\cdot, \theta) \, \mathrm{d}\Pi$.

Proof. Since Θ can be expressed as a countable union of compact sets, there is a sequence of compact sets $B_n \subseteq B_{n+1} \subseteq \Theta$ increasing to $\Theta = \cup_{n=1}^{\infty} B_n$. We can assume $\Pi(B_n) > 0$, otherwise we can take n large enough so this holds. Since π, $q > 0$ and π is upper semicontinuous on Θ, then q/π is lower semicontinuous on Θ. By Lemma 3, $\inf_{\theta\in B_n}\{q(\theta)/\pi(\theta)\}$ is monotonically

https://doi.org/10.1017/jpr.2023.21 Published online by Cambridge University Press

nonincreasing to $\varepsilon^* = \inf_{\theta\in\Theta}\{q(\theta)/\pi(\theta)\}$. 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

$$\theta_n^* = \operatorname{argmin}_{\theta\in B_n}\{q(\theta)/\pi(\theta)\}. \tag{7}$$

We can then define the sequence $\varepsilon_{\theta_n^*} = \inf_{\theta\in B_n}\{q(\theta)/\pi(\theta)\} = q(\theta_n^*)/\pi(\theta_n^*)$, and this is monotonically nonincreasing to ε^*.

Define P_n to be the Metropolis–Hastings independence kernel with independent proposal Q with density q, and target distribution $\Pi(\cdot \mid B_n)$ with density $\pi(\cdot \mid B_n) = \pi(\cdot)I_{B_n}(\cdot)/\Pi(B_n)$. 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 $\Theta \equiv \operatorname{supp}(q)$ by assumption, $q(\theta_n) > 0$. Also, if $y \in B_n^{\mathrm{c}} \cap \Theta$, 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 $\sup_{x\in\mathbb{R}^d} |\varphi(x)| \le 1$. Therefore, for any $\theta_n \in B_n$ and any function $\varphi \in M_1(\mathbb{R}^d)$,

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

Let $\varepsilon \in (0, 1 - \varepsilon^*)$. Since Q and Π are probability measures, we may then choose n_ε sufficiently large that, for all $n \ge n_\varepsilon$, $2\max\{\Pi(B_n^{\mathrm{c}}), Q(B_n^{\mathrm{c}})\} \le \varepsilon/2$. We then have

$$\begin{aligned}&\sup_{\theta_n\in B_n}\sup_{\varphi\in M_1(\mathbb{R}^d)}\left|\int_{\mathbb{R}^d} \varphi\,\mathrm{d}P_n(\theta_n, \cdot) - \int_{\mathbb{R}^d} \varphi\,\mathrm{d}P(\theta_n, \cdot)\right| \\ &= \sup_{\theta_n\in B_n}\sup_{\varphi\in M_1(\mathbb{R}^d)}\left|\int_{B_n^{\mathrm{c}}\cap\Theta} \varphi(y)\min\left\{\frac{\pi(y)q(\theta_n)}{\pi(\theta_n)q(y)}, 1\right\} q(y)\,\mathrm{d}y\right. \\ &\qquad\qquad \left. + \varphi(\theta_n)\int_{B_n^{\mathrm{c}}\cap\Theta} \min\left\{\frac{\pi(y)q(\theta_n)}{\pi(\theta_n)q(y)}, 1\right\} q(y)\,\mathrm{d}y\right| \\ &\le 2\int_{B_n^{\mathrm{c}}} q(y)\,\mathrm{d}y \le \varepsilon/2. \end{aligned} \tag{8}$$

Similarly,

$$\begin{aligned}&\sup_{\varphi\in M_1(\mathbb{R}^d)}\left|\int_{\mathbb{R}^d}\varphi\,\mathrm{d}\Pi(\cdot \mid B_n) - \int_{\mathbb{R}^d}\varphi\,\mathrm{d}\Pi\right| \\ &= \sup_{\varphi\in M_1(\mathbb{R}^d)}\left|\int_{\mathbb{R}^d}\varphi(1 - \Pi(B_n))\,\mathrm{d}\Pi(\cdot \mid B_n) - \int_{\mathbb{R}^d}\varphi\,\mathrm{d}\Pi(\cdot \mid B_n^{\mathrm{c}})\Pi(B_n^{\mathrm{c}})\right| \\ &= \Pi(B_n^{\mathrm{c}})\sup_{\varphi\in M_1(\mathbb{R}^d)}\left|\int_{\mathbb{R}^d}\varphi\,\mathrm{d}\Pi(\cdot \mid B_n) - \int_{\mathbb{R}^d}\varphi\,\mathrm{d}\Pi(\cdot \mid B_n^{\mathrm{c}})\right| \\ &\le 2\Pi(B_n^{\mathrm{c}}) \le \varepsilon/2. \end{aligned} \tag{9}$$

https://doi.org/10.1017/jpr.2023.21 Published online by Cambridge University Press

With θ_n^* as in (7), let $\psi_n(\cdot) = -\rho(\cdot, \theta_n^*)$. Then, for any $x, y \in \mathbb{R}^d$,

$$|\psi_n(x) - \psi_n(y)| \leq \rho(x, y) \tag{10}$$

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

$$\int_{\mathbb{R}^d} \psi_n \, \mathrm{d}P^t(\theta_n^*, \cdot) - \int_{\mathbb{R}^d} \psi_n \, \mathrm{d}\Pi = \int_{\mathbb{R}^d} P^{t-1}\psi_n(\cdot) \, \mathrm{d}P(\theta_n^*, \cdot) - \int_{\mathbb{R}^d} P^{t-1}\psi_n(\cdot) \, \mathrm{d}\Pi(\cdot). \tag{11}$$

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 (8), (9), and (11),

$$\int_{\mathbb{R}^d} \psi_n \, \mathrm{d}P^t(\theta_n^*, \cdot) - \int_{\mathbb{R}^d} \psi_n \, \mathrm{d}\Pi \geq \int_{\mathbb{R}^d} P^{t-1}\psi_n \, \mathrm{d}P_n(\theta_n^*, \cdot) - \int_{\mathbb{R}^d} P^{t-1}\psi_n \, \mathrm{d}\Pi(\cdot \mid B_n) - \varepsilon. \tag{12}$$

By the construction of θ_n^* in (7), we have

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

For measurable $A \subset \mathbb{R}^d$ [52, Remark 1, Theorem 2], we then have the identity

$$P_n(\theta_n^*, A) = \varepsilon_{\theta_n^*} \Pi(B_n) \Pi(A \mid B_n) + (1 - \varepsilon_{\theta_n^*} \Pi(B_n)) \delta_{\theta_n^*}(A).$$

Since $P^{t-1}\psi_n$ is a bounded measurable function, this identity leads to the following one:

$$\begin{aligned}&\int_{\mathbb{R}^d} P^{t-1}\psi_n(\cdot) \, \mathrm{d}P_n(\theta_n^*, \cdot) - \int_{\mathbb{R}^d} P^{t-1}\psi_n(\cdot) \, \mathrm{d}\Pi(\cdot \mid B_n) \\ &\qquad = (1 - \varepsilon_{\theta_n^*} \Pi(B_n)) \left(P^{t-1}\psi_n(\theta_n^*) - \int_{\mathbb{R}^d} P^{t-1}\psi_n(\cdot) \, \mathrm{d}\Pi(\cdot \mid B_n) \right).\end{aligned} \tag{13}$$

Using (12) in the first inequality, (13) in the second inequality, (9) in the third inequality, and the invariance of Π for the Markov kernel P in the last inequality,

$$\begin{aligned}&\int_{\mathbb{R}^d} P^{t-1}\psi_n P(\theta_n^*, \cdot) - \int_{\mathbb{R}^d} P^{t-1}\psi_n \, \mathrm{d}\Pi \\ &\qquad \geq \int_{\mathbb{R}^d} P^{t-1}\psi_n \, \mathrm{d}P_n(\theta_n^*, \cdot) - \int_{\mathbb{R}^d} P^{t-1}\psi_n \, \mathrm{d}\Pi(\cdot \mid B_n) - \varepsilon \\ &\qquad \geq (1 - \varepsilon_{\theta_n^*} \Pi(B_n)) \left(P^{t-1}\psi_n(\theta_n^*) - \int_{\mathbb{R}^d} P^{t-1}\psi_n \, \mathrm{d}\Pi(\cdot \mid B_n) \right) - \varepsilon \\ &\qquad \geq (1 - \varepsilon_{\theta_n^*} \Pi(B_n)) \left(P^{t-1}\psi_n(\theta_n^*) - \int_{\mathbb{R}^d} P^{t-1}\psi_n \, \mathrm{d}\Pi \right) - 2\varepsilon \\ &\qquad \geq (1 - \varepsilon_{\theta_n^*} \Pi(B_n)) \left(\int_{\mathbb{R}^d} P^{t-2}\psi_n \, \mathrm{d}P(\theta_n^*, \cdot) - \int_{\mathbb{R}^d} P^{t-2}\psi_n \, \mathrm{d}\Pi \right) - 2\varepsilon.\end{aligned}$$

Applying this inequality recursively and using the definition of ψ_n,

$$\begin{aligned}
\int_{\mathbb{R}^d} \psi_n \, \mathrm{d}P^t(\theta_n^*, \cdot) - \int_{\mathbb{R}^d} \psi_n \, \mathrm{d}\Pi \\
&= \int_{\mathbb{R}^d} P^{t-1}\psi_n \, \mathrm{d}P(\theta_n^*, \cdot) - \int_{\mathbb{R}^d} P^{t-1}\psi_n \, \mathrm{d}\Pi \\
&\geq (1 - \varepsilon_{\theta_n^*}\Pi(B_n))^t \left(\psi_n(\theta_n^*) - \int_{\mathbb{R}^d} \psi_n \, \mathrm{d}\Pi \right) - 2\varepsilon \sum_{s=0}^{t-1} (1 - \varepsilon_{\theta_n^*}\Pi(B_n))^s \\
&= (1 - \varepsilon_{\theta_n^*}\Pi(B_n))^t \int_{\mathbb{R}^d} \rho(\theta, \theta_n^*) \, \mathrm{d}\Pi - 2\varepsilon \sum_{s=0}^{t-1} (1 - \varepsilon_{\theta_n^*}\Pi(B_n))^s. \qquad (14)
\end{aligned}$$

Since $\Pi(B_n) \to 1$ and $\varepsilon_{\theta_n^*} \to \varepsilon^*$, we may take n large enough that $|\varepsilon_{\theta_n^*}\Pi(B_n) - \varepsilon^*| \leq \varepsilon$. For all large enough n and since $\varepsilon < 1 - \varepsilon^*$, we lower bound (14) to get

$$\int_{\mathbb{R}^d} \psi_n \, \mathrm{d}P^t(\theta_n^*, \cdot) - \int_{\mathbb{R}^d} \psi_n \, \mathrm{d}\Pi \geq (1 - \varepsilon^* - \varepsilon)^t \inf_{\theta \in \Theta} \int \rho(\cdot, \theta) \, \mathrm{d}\Pi - 2\varepsilon \sum_{s=0}^{t-1} (1 - \varepsilon^* + \varepsilon)^s. \qquad (15)$$

Combining (10) and (15), we lower bound the Wasserstein distance with

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

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

Appendix B. Technical lemmas

Lemma 3. *Let* $\varepsilon^* = \inf_{\theta \in \Theta}\{q(\theta)/\pi(\theta)\}$. *Suppose there is a sequence of compact sets* $B_n \subseteq B_{n+1} \subseteq \Theta$ *increasing to* $\Theta = \cup_{n=1}^{\infty} B_n$. *Define* $\varepsilon_n = \inf_{\theta \in B_n} q(\theta)/\pi(\theta)$. *Then* ε_n *is monotonically nonincreasing to its limit* ε^*.

Proof. By the definition of infimum, $\varepsilon_n \geq \varepsilon_{n+1}$ and $\varepsilon_n \geq \varepsilon^*$. Hence, the sequence ε_n converges. Let $\delta \in (0, \infty)$. By the definition of the infimum, we can choose $\theta_\delta \in \Theta$ with $\pi(\theta_\delta) > 0$ such that $q(\theta_\delta)/\pi(\theta_\delta) - \delta \leq \varepsilon^*$. We can choose B_{n_δ} such that $\theta_\delta \in B_{n_\delta}$. Then $\varepsilon_{n_\delta} - \delta \leq q(\theta_\delta)/\pi(\theta_\delta) - \delta \leq \varepsilon^*$. It follows that, for any $n \geq n_\delta$, $|\varepsilon_n - \varepsilon^*| = \varepsilon_n - \varepsilon^* \leq \varepsilon_{n_\delta} - \varepsilon^* \leq \delta$. Therefore, $\lim_n \varepsilon_n = \varepsilon^*$. $\square$

Lemma 4. *Let* $C \in \mathbb{R}^{d \times d}$ *be a positive-definite, symmetric matrix, and let* $\alpha \in (0, \infty)$. *Let* $f : \mathbb{R}^d \to \mathbb{R}$ *and suppose* $\theta \mapsto f(\theta) - \alpha\theta^\top C^{-1}\theta/2$ *is convex for all points on* $\mathbb{R}^d$. *Then there exists* $\theta^* = \mathrm{argmin}_{\theta \in \mathbb{R}^d} f(\theta)$ *and*

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

Proof. Since the function $f(\theta) - \alpha\theta^\top C^{-1}\theta/2$ is convex for all points on $\mathbb{R}^d$, it follows that, for any $\lambda \in [0, 1]$ and any $(\theta, \theta') \in \mathbb{R}^d \times \mathbb{R}^d$,

$$f(\lambda\theta + (1-\lambda)\theta') \leq \lambda f(\theta) + (1-\lambda)f(\theta') - \frac{\alpha}{2}\lambda(1-\lambda)(\theta' - \theta)^\top C^{-1}(\theta' - \theta).$$

Since C^{-1} is positive-definite, $\alpha\lambda(1-\lambda)(\theta' - \theta)^\top C^{-1}(\theta' - \theta)/2$ 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, f attains its minimum $\theta^* \in \mathbb{R}^d$. The right directional derivative $f'(\theta^*;\theta) = \lim_{t \downarrow 0} t^{-1}[f(\theta^* + t\theta) - f(\theta^*)]$ exists for all points $\theta \in \mathbb{R}^d$ [32, Theorem 3.1.12]. For $\lambda \in (0, 1)$,

$$\frac{1}{(1-\lambda)}\frac{1}{\lambda}[f(\theta^* + \lambda(\theta - \theta^*)) - f(\theta^*)] - \frac{1}{(1-\lambda)}(f(\theta) - f(\theta^*)) \leq -\frac{\alpha}{2}(\theta - \theta^*)^\top C^{-1}(\theta - \theta^*).$$

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

$$f'(\theta^*;\theta - \theta^*) - f(\theta) + f(\theta^*) \leq -\frac{\alpha}{2}(\theta - \theta^*)^\top C^{-1}(\theta - \theta^*).$$

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 $\theta \in \mathbb{R}^d$,

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

Acknowledgements

We would like to thank the associate editor and referees for their helpful comments that improved the presentation of this article. We would also like to thank Qian Qin and Dootika Vats for their helpful comments, which improved an earlier draft of this article.

Funding information

Jones was partially supported by NSF grant DMS-2152746.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88**, 669–679.

[2] Belloni, A. and Chernozhukov, V. (2009). On the computational complexity of MCMC-based estimators in large samples. *Ann. Statist.* **37**, 2011–2055.

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

[4] Brooks, S., Gelman, A., Jones, G. L. and Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. Chapman and Hall/CRC, New York.

[5] Brown, A. and Jones, G. L. (2023). Lower bounds on the rate of convergence for accept–reject-based Markov chains. Preprint, arXiv:2212.05955.

[6] Demidenko, E. (2001). Computational aspects of probit model. *Math. Commun.* **6**, 233–247.

[7] 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. *Statist. Comput.* **25**, 5–19.

[8] DURMUS, A. AND MOULINES, É. (2019). High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli* **25**, 2854–2882.

[9] DWIVEDI, R., CHEN, Y., WAINWRIGHT, M. J. AND YU, B. (2018). Log-concave sampling: Metropolis–Hastings algorithms are fast! *Proc. Mach. Learn. Res.* **75**, 793–797.

[10] EBERLE, A. (2014). Error bounds for Metropolis–Hastings algorithms applied to perturbations of Gaussian measures in high dimensions. *Ann. Appl. Prob.* **24**, 337–377.

[11] EKVALL, K. O. AND JONES, G. L. (2021). Convergence analysis of a collapsed Gibbs sampler for Bayesian vector autoregressions. *Electron. J. Statist.* **15**, 691–721.

[12] GEMAN, S. (1980). A limit theorem for the norm of random matrices. *Ann. Prob.* **8**, 252–261.

[13] GIBBS, A. L. (2004). Convergence in the Wasserstein metric for Markov chain Monte Carlo algorithms with applications to image restoration. *Stoch. Models* **20**, 473–492.

[14] GIRAUDO, D. (2014). Product measure with a Dirac delta marginal. Mathematics Stack Exchange. Available at: https://math.stackexchange.com/questions/794299/product-measure-with-a-dirac-delta-marginal.

[15] HAIRER, M., STUART, A. M. AND VOLLMER, S. J. (2014). Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *Ann. Appl. Prob.* **24**, 2455–2490.

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

[17] HIRIART-URRUTY, J.-B. AND LEMARÉCHAL, C. (2001). *Fundamentals of Convex Analysis*. Springer, Berlin.

[18] JARNER, S. F. AND HANSEN, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stoch. Process. Appl.* **85**, 341–361.

[19] JIN, R. AND TAN, A. (2020). Central limit theorems for Markov chains based on their convergence rates in Wasserstein distance. Preprint, arXiv:2002.09427.

[20] JOHNDROW, J. E., SMITH, A., PILLAI, N. AND DUNSON, D. B. (2019). MCMC for imbalanced categorical data. *J. Amer. Statist. Assoc.* **114**, 1394–1403.

[21] JOHNSON, L. T. AND GEYER, C. J. (2012). Variable transformation to obtain geometric ergodicity in the random-walk Metropolis algorithm. *Ann. Statist.* **40**, 3050–3076.

[22] JONES, G. L. (2004). On the Markov chain central limit theorem. *Prob. Surv.* **1**, 299–320.

[23] JOULIN, A. AND OLLIVIER, Y. (2010). Curvature, concentration and error estimates for Markov chain Monte Carlo. *Ann. Prob.* **38**, 2418–2442.

[24] KANTOROVICH, L. V. AND RUBINSTEIN, G. S. (1957). On a function space in certain extremal problems. *Dokl. Akad. Nauk USSR* **115**, 1058–1061.

[25] KHARE, K. AND HOBERT, J. P. (2012). Geometric ergodicity of the Gibbs sampler for Bayesian quantile regression. *J. Multivar. Anal*, **112**, 108–116.

[26] KOMOROWSKI, T. AND WALCZUK, A. (2011). Central limit theorem for Markov processes with spectral gap in the Wasserstein metric. *Stoch. Process. Appl.* **122**, 2155–2184.

[27] LIU, J. S. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statist. Comput.* **6**, 113–119.

[28] MADRAS, N. AND SEZER, D. (2010). Quantitative bounds for Markov chain convergence: Wasserstein and total variation distances. *Bernoulli* **16**, 882–908.

[29] MENGERSEN, K. L. AND TWEEDIE, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* **24**, 101–121.

[30] METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. AND TELLER, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092.

[31] MEYN, S. P. AND TWEEDIE, R. L. (2009). *Markov Chains and Stochastic Stability*, 2nd edn. Cambridge University Press.

[32] NESTEROV, Y. (2018). *Lectures on Convex Optimization*, 2nd edn. Springer, Cham.

[33] PAPASPILIOPOULOS, O., ROBERTS, G. O. AND ZANELLA, G. (2019). Scalable inference for crossed random effects models. *Biometrika* **107**, 25–40.

[34] PAPASPILIOPOULOS, O., STUMPF-FÉTIZON, T. AND ZANELLA, G. (2021). Scalable computation for Bayesian hierarchical models. Preprint, arXiv:2103.10875.

[35] PIERRE, J., ROBERT, C. P. AND SMITH, M. H. (2011). Using parallel computation to improve independent Metropolis–Hastings based estimation. *J. Comput. Graph. Statist.* **20**, 616–635.

[36] POLSON, N. G., SCOTT, J. G. AND WINDLE, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *J. Amer. Statist. Assoc.* **108**, 1339–1349.

[37] QIN, Q. AND HOBERT, J. P. (2019). Convergence complexity analysis of Albert and Chib's algorithm for Bayesian probit regression. *Ann. Statist.* **47**, 2320–2347.

[38] QIN, Q. AND HOBERT, J. P. (2021). On the limitations of single-step drift and minorization in Markov chain convergence analysis. *Ann. Appl. Prob.* **31**, 1633–1659
[39] QIN, Q. AND HOBERT, J. P. (2022). Geometric convergence bounds for Markov chains in Wasserstein distance based on generalized drift and contraction conditions. *Ann. Inst. H. Poincaré* **58**, 872–889.
[40] QIN, Q. AND HOBERT, J. P. (2022). Wasserstein-based methods for convergence complexity analysis of MCMC with applications. *Ann. Appl. Prob.* **32**, 124–166.
[41] 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.
[42] ROBERTS, G. O. AND TWEEDIE, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83**, 95–110.
[43] ROBERTSON, N., FLEGAL, J. M., VATS, D. AND JONES, G. L. (2021). Assessing and visualizing simultaneous simulation error. *J. Comput. Graph. Statist.* **30**, 324–334.
[44] ROSENTHAL, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **90**, 558–566.
[45] SHEPHARD, N. AND PITT, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika* **84**, 653–667.
[46] SMITH, R. L. AND TIERNEY, L. (1996). Exact transition probabilities for the independence Metropolis sampler. Technical report, Department of Statistics, University of Cambridge.
[47] SUR, P. AND CANDÈS, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Nat. Acad. Sci.* **116**, 14516–14525.
[48] TIERNEY, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22**, 1701–1728.
[49] VATS, D., FLEGAL, J. M. AND JONES, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika* **106**, 321–337.
[50] VILLANI, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society, Providence, RI.
[51] VILLANI, C. (2009). *Optimal Transport: Old and New*. Springer, Berlin.
[52] WANG, G. (2022). Exact convergence rate analysis of the independent Metropolis–Hastings algorithms. *Bernoulli* **28**, 2012–2033.
[53] YANG, Y., WAINWRIGHT, M. J. AND JORDAN, M. I. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *Ann. Statist.* **44**, 2497–2532.