Coupling-based convergence assessment of some Gibbs samplers for high-dimensional Bayesian regression with shrinkage priors

Niloy Biswas | Anirban Bhattacharya | Pierre E. Jacob | James E. Johndrow

1Harvard University, Cambridge, USA
2Texas A&M University, College Station, USA
3ESSEC Business School, Cergy-Pontoise, France
4The Wharton School, University of Pennsylvania, Philadelphia, USA

Correspondence
Niloy Biswas, Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA.
Email: niloy_biswas@g.harvard.edu

Abstract
We consider Markov chain Monte Carlo (MCMC) algorithms for Bayesian high-dimensional regression with continuous shrinkage priors. A common challenge with these algorithms is the choice of the number of iterations to perform. This is critical when each iteration is expensive, as is the case when dealing with modern data sets, such as genome-wide association studies with thousands of rows and up to hundreds of thousands of columns. We develop coupling techniques tailored to the setting of high-dimensional regression with shrinkage priors, which enable practical, non-asymptotic diagnostics of convergence without relying on traceplots or long-run asymptotics. By establishing geometric drift and minorization conditions for the algorithm under consideration, we prove that the proposed couplings have finite expected meeting time. Focusing on a class of shrinkage priors which includes the ‘Horseshoe’, we empirically demonstrate the scalability of the proposed couplings. A highlight of our findings is that less than 1000 iterations can be enough for a Gibbs sampler to reach stationarity in a regression on 100,000 covariates. The numerical results also illustrate the impact of the prior on the computational efficiency of the coupling, and suggest the use of priors where the local precisions
are Half-$t$ distributed with degree of freedom larger than one.

**KEYWORDS**
Bayesian inference, couplings, Gibbs sampling, Horseshoe prior, parallel computation

## 1 | INTRODUCTION

### 1.1 | Iterative computation in high dimensions

We consider the setting of high-dimensional regression where the number of observations $n$ is smaller than the number of covariates $p$ and the true signal is sparse. This problem formulation is ubiquitous in modern applications ranging from genomics to the social sciences. Optimization-based methods, such as the LASSO (Tibshirani, 1994) or Elastic Net (Zou & Hastie, 2005), allow sparse point estimates to be obtained even when the number of covariates is on the order of hundreds of thousands. More specifically, iterative optimization procedures to obtain these estimates are practical because the following conditions are met: (1) the cost per iteration scales favourably with the size of the input ($n$ and $p$), (2) the number of iterations to convergence also scales favourably and (3) there are reliable stopping criteria to detect convergence.

In the Bayesian paradigm, Markov chain Monte Carlo (MCMC) methods are commonly used to sample from the posterior distribution. Default MCMC implementations, for example using general-purpose software such as BUGS (Lunn et al., 2000), JAGS (Plummer, 2003) and Stan (Carpenter et al., 2017), can lead to high computational costs in high-dimensional settings, both per iteration and in terms of the number of iterations required to reach convergence. In the setting of high-dimensional regressions, tailored algorithms can provide substantial improvements on both fronts (e.g. Bhattacharya & Johndrow, 2021; Bhattacharya et al., 2016; Johndrow et al., 2020; Nishimura & Suchard, 2018). Comparatively less attention has been put on the design of reliable stopping criteria. Stopping criteria for MCMC, such as effective sample size (ESS, e.g. Vats et al., 2019), or the $\hat{R}$ convergence diagnostic (e.g. Vats & Knudson, 2021), rely on the asymptotic behaviour of the chains as time goes to infinity, and thus effectively ignore the non-asymptotic ‘burn-in’ bias, which vanishes as time progresses. This is acceptable in situations where we have solid a priori estimates of the burn-in period; otherwise the lack of non-asymptotic stopping criteria poses an important practical problem. With Bayesian analysis of high-dimensional data sets, MCMC algorithms can require seconds or minutes per iteration, and thus any non-asymptotic insight on the number of iterations to perform is highly valuable.

Diagnostics of convergence are particularly useful when considering the number of factors that affect the performance of MCMC algorithms for Bayesian regressions. Beyond the size of the data set, the specification of the prior and the (much less explicit) signal-to-noise ratio in the data all have an impact on the convergence of MCMC algorithms. Across applications, the number of iterations required for the chain to converge varies by multiple orders of magnitude. This could lead users to either waste computational resources by running overly long chains, or worrisomely, to base their analysis on chains that have not converged. This manuscript proposes a concrete method to avoid these pitfalls in the case of a Gibbs sampler for Bayesian
regression with heavy-tailed shrinkage priors. Specifically, we follow the approach of Glynn and Rhee (2014), Jacob et al. (2020), Biswas et al. (2019) and use coupled lagged chains to monitor convergence.

Diagnostics based on Stein’s method (Gorham & Mackey, 2015, 2017) have also been employed to assess MCMC convergence. In particular, the diffusion Stein discrepancies proposed in Gorham et al. (2019) can accommodate heavy-tailed distributions. However, these diagnostics require the density of the target distribution to be continuously differentiable. This condition is not satisfied by the posterior distributions under consideration in this manuscript, which have infinite density about the origin (see section 1 of the online supplementary materials). Section 3.3 of the online supplementary materials contains some initial simulations which illustrate the challenges of applying Stein’s method in the present setting. Extending current Stein’s method based diagnostics to such target distributions is an exciting avenue for future research.

### 1.2 Bayesian shrinkage regression with Half-t(ν) priors

Consider Gaussian linear regression with \( n \) observations and \( p \) covariates, with a focus on the high-dimensional setting where \( n \ll p \). The likelihood is given by

\[
L(\beta, \sigma^2; X, y) = \frac{1}{(2\pi \sigma^2)^{n/2}} \exp\left( -\frac{1}{2\sigma^2} \| y - X\beta \|^2 \right),
\]

where \( \| \cdot \| \) denotes the \( L_2 \) norm, \( X \in \mathbb{R}^{n \times p} \) is the observed design matrix, \( y \in \mathbb{R}^n \) is the observed response vector, \( \sigma^2 > 0 \) is the unknown Gaussian noise variance, and \( \beta \in \mathbb{R}^p \) is the unknown signal vector that is assumed to be sparse. We study hierarchical Gaussian scale mixture priors on \((\beta, \sigma^2)\) given by

\[
\begin{align*}
\beta_j \mid \sigma^2, \xi, \eta & \sim \text{ind} \mathcal{N} \left( 0, \frac{\sigma^2}{\xi \eta_j} \right), \\
\xi & \sim \pi_{\xi}(\cdot), \\
\eta_j & \sim \text{i.i.d.} \pi_{\eta}(\cdot), \\
\sigma^2 & \sim \text{InvGamma} \left( \frac{a_0}{2}, \frac{b_0}{2} \right),
\end{align*}
\]

where \( a_0, b_0 > 0 \) and \( \pi_{\xi}(\cdot), \pi_{\eta}(\cdot) \) are continuous densities on \( \mathbb{R}_{>0} \). Such global-local mixture priors induce approximate sparsity, where the components of \( \beta \) can be arbitrarily close to zero, but never exactly zero. This is in contrast to point-mass mixture priors (e.g. Johnson & Rossell, 2012), where some components of \( \beta \) can be exactly zero a posteriori. The global precision parameter \( \xi \) relates to the number of signals, and we use \( \xi^{-1/2} \sim \text{Cauchy}_+(0, 1) \) throughout. The local precision parameters \( \eta_j \) determine which components of \( \beta \) are null.

We focus on the Half-t(ν) prior family for the local scale parameter, \( \eta_j^{-1/2} \sim t_+(\nu) \) for \( \nu \geq 1 \), where a \( t_+(\nu) \) distribution has density proportional to \( (1 + x^2 / \nu)^{-(\nu+1)/2} \mathbb{I}_{(0,\infty)}(x) \). The induced prior on the local precision \( \eta_j \) has density

\[
\pi_{\eta}(\eta_j) \propto \frac{1}{\eta_j^{2\nu} \left( 1 + \nu \eta_j \right)^{\nu+1/2}} \mathbb{I}_{(0,\infty)}(\eta_j).
\]

The case \( \nu = 1 \) corresponds to the Horseshoe prior (Carvalho et al., 2009, 2010) and has received overwhelming attention in the literature among this prior class; see Bhadra et al. (2019) for a recent overview. The use of Half-t(ν) priors on scale parameters in hierarchical Gaussian models
was popularized by Gelman (2006). However, the subsequent literature has largely gravitated towards the default choice (Polson & Scott, 2012) of $\nu = 1$, in part because a convincing argument for preferring a different value of $\nu$ has been absent to date. Here we give empirical evidence that Half-$t(\nu)$ priors yield statistical estimation properties similar to the Horseshoe prior, while leading to significant computational gains when $\nu > 1$. We note that Ghosh and Chakrabarti (2017) established optimal posterior concentration in the Normal means problem for a broad class of priors on the local scale that includes the Half-$t(\nu)$ priors, extending earlier work by van der Pas et al. (2014) for the Horseshoe, providing theoretical support for the comparable performance we observe with $\nu = 1$ and $\nu > 1$ in simulations.

Our stopping criterion relies on couplings. Couplings have long been used to analyse the convergence of MCMC algorithms, while methodological implementations have been rare. The work of Johnson (1996) pioneered the use of couplings for practical diagnostics of convergence, but relied on the assumption of some similarity between the initial distribution and the target, which can be hard to check. Here we instead follow the approach of Glynn and Rhee (2014) and Jacob et al. (2020) based on couplings of lagged chains, applied to the question of convergence in Biswas et al. (2019). That approach makes no assumptions on the closeness of the initial distribution to the target, and in our experiments we initialize all chains independently from the prior distribution.

The approach requires the ability to sample pairs of Markov chains such that (a) each chain marginally follows a prescribed MCMC algorithm and (b) the two chains ‘meet’ after a random—but almost surely finite—number of iterations, called the meeting time. When the chains meet, that is, when all of their components become identical, the user can stop the procedure. From the output, the user can obtain unbiased estimates of expectations with respect to the target distribution, which can be generated in parallel and averaged, as well as estimates of the finite-time bias of the underlying MCMC algorithm. Crucially, this information is retrieved from the distribution of the meeting time, which is an integer-valued random variable irrespective of the dimension of the problem, and thus provides a convenient summary of the performance of the algorithm. A difficulty inherent to the approach is the design of an effective coupling strategy for the algorithm under consideration. High-dimensional parameter spaces add a substantial layer of complication as some of the simpler coupling strategies, that are solely based on maximal couplings, lead to prohibitively long meeting times.

### 1.3 Our Contribution

In this paper, we argue that couplings offer a practical way of assessing the number of iterations required by MCMC algorithms in high-dimensional regression settings. Our specific contributions are summarized below.

We introduce blocked Gibbs samplers (Section 2) for Bayesian linear regression focusing on Half-$t(\nu)$ local shrinkage priors, extending the algorithm of Johndrow et al. (2020) for the Horseshoe. Our algorithm has an overall computation cost of $O(n^2p)$ per iteration, which is state-of-the-art for $n \ll p$. We then design a one-scale coupling strategy for these Gibbs samplers (Section 2.3) and show that it results in meeting times that have exponential tails, which in turn validates their use in convergence diagnostics (Biswas et al., 2019). Our proof of exponentially tailed meeting times is based on identifying a geometric drift condition, as a corollary to which we also establish geometric ergodicity of the marginal Gibbs sampler. Despite a decade
of widespread use, MCMC algorithms for the Horseshoe and Half-$t(\nu)$ priors generally have not previously been shown to be geometrically ergodic, in contrast to MCMC algorithms for global-local models with exponentially tailed local scale priors (Khare & Hobert, 2013; Pal & Khare, 2014). Our proofs utilize a uniform bound on generalized ridge regression estimates, which enables us to avoid a technical assumption of Johndrow et al. (2020), and auxiliary results on moments of distributions related to the confluent hypergeometric function of the second kind. Shortly after an earlier version of this manuscript was posted, Bhattacharya et al. (2022) have also established geometric ergodicity of related Gibbs samplers for the Horseshoe prior.

The meeting times resulting from the one-scale coupling increase exponentially as dimension increases. To address that issue we design improved coupling strategies (Section 3). We develop a two-scale strategy, based on a carefully chosen metric, which informs the choice between synchronous and maximal couplings. In numerical experiments in high dimensions this strategy leads to orders of magnitude shorter meeting times compared to the naïve approach. We establish some preliminary results in a stylized setting, providing a partial understanding of why the two-scale strategy leads to the observed drastic reduction in meeting times. We also describe other couplings that do not require explicit calculation of a metric between the chains, have exponentially tailed meeting times, and empirically result in similar meeting times as the two-scale coupling. For our two-scale coupling, we further note that priors with stronger shrinkage towards zero can give significantly shorter meeting times in high dimensions. This motivates usage of Half-$t(\nu)$ priors with greater degrees of freedom $\nu > 1$, which can give similar statistical performance to the Horseshoe ($\nu = 1$) while enjoying orders of magnitude computational improvements.

Finally, we demonstrate (Section 4) that the proposed coupled MCMC algorithm is applicable to big data settings. Figure 1 displays an application of our coupling strategy to monitor the convergence of the Gibbs sampler for Bayesian regression on a genome-wide association study (GWAS) data set with $(n, p) \approx (2000, 100,000)$. Our method suggests that a burn-in of just 700 iterations can suffice even for such a large problem, and confirms the applicability of coupled MCMC algorithms for practical high-dimensional inference.

Scripts in R (R Core Team, 2013) are available from https://github.com/niloyb/CoupledHalfT to reproduce the figures of the article.
2 | COUPLED MCMC FOR REGRESSION WITH HALF-\( t(\nu) \) PRIORS

We develop MCMC techniques for Bayesian shrinkage regression with Half-\( t(\nu) \) priors. The model is given by Equations (1)–(2) and \( \eta^{-1/2} \sim t(\nu) \) for some \( \nu \geq 1 \) as in Equation (3). Posterior distributions resulting from such heavy-tailed shrinkage priors have desirable statistical properties, but their features pose challenges to generic-purpose MCMC algorithms. Characterization of the marginal prior and posterior densities of \( \beta \) on \( \mathbb{R}^p \) given \( \xi \) and \( \sigma^2 \), and further comments on these challenges are presented in section 1 of the online supplementary materials. Specifically, the resulting posterior distributions present (a) multimodality, (b) heavy tails and (c) infinite density at zero. This hints at a trade-off between statistical accuracy and computational feasibility, since the very features that present computational difficulties are crucial for optimal statistical performance across sparsity levels and signal strengths.

2.1 | A blocked Gibbs sampler for Half-\( t(\nu) \) priors

Blocked Gibbs samplers are popularly used for Bayesian regression with global-local shrinkage priors, due to the analytical tractability of full conditional distributions (Bhattacharya et al., 2015; Carvalho et al., 2010; Makalic & Schmidt, 2016; Park & Casella, 2008; Polson et al., 2014). Several convenient blocking and marginalization strategies are possible, leading to conditionals that are easy to sample from. For the case of the Horseshoe prior (\( \nu = 1 \)), Johnedrow et al. (2020) have developed exact and approximate MCMC algorithms for high-dimensional settings, building on the algorithms of Polson et al. (2014) and Bhattacharya et al. (2016). We extend that sampler to general Half-\( t(\nu) \) priors, and summarize it in Equation (4). Given some initial state \((\eta_0, \beta_0, \sigma^2_0, \xi_0) \in \mathbb{R}_+^p \times \mathbb{R}^p \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \), it generates a Markov chain \((\eta_t, \beta_t, \sigma^2_t, \xi_t)_{t \geq 0} \) which targets the posterior corresponding to Half-\( t(\nu) \) priors.

\[
\pi(\eta_{t+1} | \beta_t, \sigma^2_t, \xi_t) \propto \prod_{j=1}^{p} \frac{e^{-m_{ij}y_{i+1,j}}}{\eta_{t+1,j}^{1/2} \left( 1 + \nu \eta_{t+1,j} \right)^{1/2}} \quad \text{for } m_{ij} = \frac{\xi_t \beta^2_{t,j}}{2 \sigma^2_t},
\]

\[
\pi(\xi_{t+1} | \eta_{t+1}) \propto L(y | \xi_{t+1}, \eta_{t+1}) \pi_t(\xi_{t+1}),
\]

\[
\sigma^2_{t+1} | \eta_{t+1}, \xi_{t+1} \sim \text{InvGamma} \left( \frac{a_0 + n}{2}, \frac{y^T M^{-1} y + b_0}{2} \right),
\]

\[
\beta_{t+1} | \eta_{t+1}, \xi_{t+1}, \sigma^2_{t+1} \sim \mathcal{N}(\Sigma^{-1} X^T y, \sigma^2_{t+1} \Sigma^{-1}).
\]

Above \( L(y | \xi_{t+1}, \eta_{t+1}) \) represents the marginal likelihood of \( y \) given \( \xi_{t+1} \) and \( \eta_{t+1} \), and we use the notation \( M = I_n + \xi_{t+1}^{-1} X \text{Diag}(\eta_{t+1}^{-1}) X^T \), and \( \Sigma = X^T X + \xi_{t+1} \text{Diag}(\eta_{t+1}) \). We sample \( \eta_{t+1} | \beta_t, \sigma^2_t, \xi_t \) component-wise independently using slice sampling and sample \( \xi_{t+1} | \eta_{t+1} \) using Metropolis–Rosenbluth–Teller–Hastings, with a Normal proposal on the logarithm of \( \xi \) with standard deviation \( \sigma_{\text{MRTH}} \) (Hastings, 1970). Full details of our sampler are in Algorithms 1 and 2 of the online supplementary materials, and derivations are in section 6 of the online supplementary materials.

Sampling \( \eta_{t+1} | \beta_t, \sigma^2_t, \xi_t \) component-wise independently is an \( O(p) \) cost operation. Sampling \( \xi_{t+1} | \eta_{t+1} \) requires evaluating \( L(y | \xi_{t+1}, \eta_{t+1}) \), which involves an \( O(n^3 p) \) cost operation (calculating the weighted matrix cross-product \( X \text{Diag}(\eta_{t+1}^{-1}) X^T \)) and an \( O(n^3) \) cost operation (calculating
Coupling the proposed Gibbs sampler

We develop couplings for the blocked Gibbs sampler presented in Section 2.1. We will use these couplings to generate coupled chains with a time lag $L \geq 1$, in order to implement the diagnostics of convergence proposed in Biswas et al. (2019). Consider an $L$-lag coupled chain $(C_t, \tilde{C}_{t-L})_{t \geq L}$ with random meeting time $\tau$ such that $C_t = \tilde{C}_{t-L}$ for all $t \geq \tau$. Here $C_t$ and $\tilde{C}_t$ denote the full states $(\eta_t, \beta_t, \sigma_t^2, \xi_t)$ and $(\tilde{\eta}_t, \tilde{\beta}_t, \tilde{\sigma}_t^2, \tilde{\xi}_t)$, respectively. For the coupling-based methods proposed in Biswas et al. (2019), Jacob et al. (2020) to be most practical, the meeting times should occur as early as possible, under the constraint that both chains $(C_{t})_{t \geq 0}$ and $(\tilde{C}_{t})_{t \geq 0}$ start from the same initial distribution and marginally evolve according to the blocked Gibbs sampler in Equation (4).

The total variation (TV) distance between two probability distributions $P$ and $Q$ is defined as $\text{TV}(P, Q) := \inf \mathbb{P}(X \neq Y)$, where the infimum is taken over all pairs $(X, Y)$ such that marginally $X \sim P$ and $Y \sim Q$. Briefly, we recall how $L$-lag coupled chains enable the estimation of upper bounds on $\text{TV}(\pi_t, \pi)$, which denotes the total variation distance between the distribution of the blocked Gibbs sampler in Equation (4) after $t$ iterations $(\pi_t)$ and its stationary distribution $(\pi)$. Since $C_t = \tilde{C}_{t-L}$ for all $t \geq \tau$, we obtain $\mathbb{I}(C_t \neq \tilde{C}_{t-L}) = \mathbb{I}(\tau > t)$. If the meeting times are exponentially tailed, Biswas et al. (2019) establishes

$$\text{TV}(\pi_t, \pi) \leq \sum_{j=1}^{\infty} \text{TV}(\pi_{t+jL-L}, \pi_{t+jL}) \leq \sum_{j=1}^{\infty} \mathbb{E}[\mathbb{I}(\tilde{C}_{t+jL-L} \neq C_{t+jL})] = \mathbb{E}\left[\max\left(0, \left|\frac{\tau - L - t}{L}\right|\right)\right],$$

where we first used the triangle inequality, the second inequality follows from the coupling representation of the TV distance, and the equality follows from interchanging the summation and expectation using the dominated convergence theorem, which can be applied when the meeting times are subexponential. Exponentially tailed meeting times also ensure that the expected computation time of unbiased estimators as in Jacob et al. (2020), generated independently in parallel, scales at a logarithmic rate in the number of processors. Thus, we will be particularly interested in couplings that result in exponentially tailed meeting times; see Middleton et al. (2020) for the use of coupled chains with polynomially tailed meeting times.

In Section 2.3, we consider an algorithm based on maximal couplings. We show that the ensuing meeting times have exponential tails, and discuss how our analysis directly implies that the marginal chain introduced in Section 2.1 is geometrically ergodic. We then illustrate difficulties encountered by such scheme in high dimensions, which will motivate alternative strategies described in Section 3. For simplicity, we mostly omit the lag $L$ from the notation.

2.3 One-scale coupling

For the blocked Gibbs sampler in Section 2.1, we first consider a coupled MCMC algorithm that attempts exact meetings at every step. We will apply a maximal coupling algorithm with
independent residuals (see Thorisson, 2000, chapter 1 section 4.4; Johnson, 1998). It is included in Algorithm 3 of the online supplementary materials, and has an expected computation cost of two units (Johnson, 1998).

Our initial coupled MCMC kernel is given in Algorithm 1, which we refer to as a one-scale coupling. That is, before the chains have met (when \( C_t \neq \tilde{C}_t \)), the coupled kernel on Steps (1) and (2)(a–c) does not explicitly depend on the distance between states \( C_t \) and \( \tilde{C}_t \). After meeting, the coupled chains remain together by construction, such that \( C_t = \tilde{C}_t \) implies \( C_{t+1} = \tilde{C}_{t+1} \). When \( C_t \neq \tilde{C}_t \), Step (1) uses the coupled slice sampler of Algorithm 2 component-wise for \((\eta_{t+1}, \tilde{\eta}_{t+1})\), which allows each pair of components \((\eta_{t+1,j}, \tilde{\eta}_{t+1,j})\) to meet exactly with positive probability. In Algorithm 2, we use a common random numbers or a ‘synchronous’ coupling of the auxiliary random variables \((U_j, \tilde{U}_j)\), and alternative couplings could be considered. Steps (2)(a–b) are maximal couplings of the conditional sampling steps for \((\xi, \tilde{\xi})\) and \((\sigma^2, \tilde{\sigma}^2)\), such that \((\eta_{t+1}, \sigma_{t+1}, \xi_{t+1}) = (\tilde{\eta}_{t+1}, \tilde{\sigma}_{t+1}, \tilde{\xi}_{t+1}) \) occurs with positive probability for all \( t \geq 0 \). Step (2)(c) uses common random numbers, such that \((\eta_{t+1}, \sigma_{t+1}, \xi_{t+1}) = (\tilde{\eta}_{t+1}, \tilde{\sigma}_{t+1}, \tilde{\xi}_{t+1}) \) implies \( \beta_{t+1} = \tilde{\beta}_{t+1} \). This allows the full chains to meet exactly with positive probability at every step.

**Algorithm 1**: A one-scale coupled MCMC kernel for Half-\( t(\nu) \) priors.

**Input**: \( C_t := (\eta_t, \beta_t, \sigma^2_t, \xi_t) \) and \( \tilde{C}_t := (\tilde{\eta}_t, \tilde{\beta}_t, \tilde{\sigma}^2_t, \tilde{\xi}_t) \).

**if** \( C_t = \tilde{C}_t \) **then** Sample \( C_{t+1} \mid C_t \) as in (4) and set \( \tilde{C}_{t+1} = C_{t+1} \).

**else**

1. Sample \((\eta_{t+1}, \tilde{\eta}_{t+1}) \mid C_t, \tilde{C}_t \) component-wise independently using Algorithm 2.

2. Sample \(((\xi_{t+1}, \sigma^2_{t+1}, \beta_{t+1}), (\tilde{\xi}_{t+1}, \tilde{\sigma}^2_{t+1}, \tilde{\beta}_{t+1})) \) given \( \eta_{t+1}, \tilde{\eta}_{t+1} \) as follows:

   (a) Sample \((\xi_{t+1}, \tilde{\xi}_{t+1}) \mid \eta_{t+1}, \xi_t, \tilde{\eta}_{t+1}, \tilde{\xi}_t \)
   using coupled
   Metropolis–Rosenbluth–Teller–Hastings (Algorithm 7 of the supplementary materials).

   (b) Sample \((\sigma^2_{t+1}, \tilde{\sigma}^2_{t+1}) \mid \xi_{t+1}, \eta_{t+1}, \xi_t, \tilde{\xi}_{t+1}, \tilde{\eta}_{t+1} \)
   from a maximal coupling of two Inverse Gamma distributions (Algorithm 3 of the supplementary materials).

   (c) Sample \((\beta_{t+1}, \tilde{\beta}_{t+1}) \mid \sigma^2_{t+1}, \xi_{t+1}, \eta_{t+1}, \sigma^2_t, \xi_t, \tilde{\sigma}^2_{t+1}, \tilde{\xi}_{t+1}, \tilde{\eta}_{t+1} \)
   using common random numbers for Gaussian scale mixture priors (Algorithm 6 of the supplementary materials).

**return** \((C_{t+1}, \tilde{C}_{t+1})\).

**Algorithm 2**: Coupled Slice Sampler for Half-\( t(\nu) \) priors using maximal couplings.

**Result**: A coupling of slice samplers marginally targeting \( p(\mid m_j) \) and \( p(\mid \tilde{m}_j) \), where

\[
p(\eta_j \mid m) \propto (\eta_j^{\nu_j} (1 + \nu m_j)^{\nu_j + 1})^{-1} e^{-m \eta_j} \text{ on } (0, \infty).
\]

**Input**: \( \eta_{t,j}, \tilde{\eta}_{t,j}, m_{t,j} := \xi_t \beta_{t,j}^2 / (2 \sigma^2_{t,j}), \tilde{m}_{t,j} := \tilde{\xi}_t \tilde{\beta}_{t,j}^2 / (2 \tilde{\sigma}^2_{t,j}) > 0. \)

1. Sample \( U_{j,*}^{\text{crn}} \sim \text{Uniform}(0,1) \), and set \( U_{j,*} := U_{j,*}^{\text{crn}} (1 + \nu \eta_{t,j})^{-\frac{\nu_j + 1}{2}} \) and
   \( \tilde{U}_{j,*} := U_{j,*}^{\text{crn}} (1 + \nu \tilde{\eta}_{t,j})^{-\frac{\nu_j + 1}{2}}. \)

2. Sample \((\eta_{t+1,j}, \tilde{\eta}_{t+1,j}) \mid U_{j,*}, \tilde{U}_{j,*} \)
from a maximal coupling of distributions \( P_j \) and \( \tilde{P}_j \) using Algorithm 3 of the supplementary materials, where \( P_j \) and \( \tilde{P}_j \) have unnormalized densities
\[
\eta \mapsto \eta^{s-1} e^{-m \eta} \text{ and } \eta \mapsto \eta^{s-1} e^{-\tilde{m} \eta} \text{ on } (0, T_j) \) and \((0, \tilde{T}_j)\) respectively, where
\[
T_j = (U_{j,*}^{-(1+\nu)^{-1}} - 1) / \nu, \tilde{T}_j = (\tilde{U}_{j,*}^{-(1+\nu)^{-1}} - 1) / \nu \text{ and } s = (1 + \nu) / 2.
\]

**return** \((\eta_{t+1,j}, \tilde{\eta}_{t+1,j})\).
Under the one-scale coupling, we prove that the meeting times have exponential tails and hence finite expectation. Our analysis is based on identifying a suitable drift function for a variant of the marginal chain in Section 2.1 and an application of Jacob et al. (2020 proposition 4). We assume that the global shrinkage prior \( \pi_\zeta(\cdot) \) has a compact support on \((0, \infty)\). Such compactly supported priors for \( \zeta \) have been recommended by van der Pas et al. (2017) to achieve optimal posterior concentration for the Horseshoe in the Normal means model. While the default half-Cauchy prior \( \xi^{-1/2} \sim \text{Cauchy}_\xi(0, 1) \) empirically used throughout the manuscript is not covered by our theoretical analysis, section 3 of the online supplementary materials reports essentially identical numerical results obtained with a compactly supported prior for \( \xi \). We also assume that each component \( \eta_{t+1,j} | \beta_{t,j}, \xi_t, \sigma_t^2 \) and \( \xi_{t+1} | \eta_{t+1} \) in Equation (4) are sampled perfectly, instead of slice sampling and Metropolis–Rosenbluth–Teller–Hastings steps, respectively. Perfect sampling algorithms for each component \( \eta_{t+1,j} | \beta_{t,j}, \xi_t, \sigma_t^2 \) and \( \xi_{t+1} | \eta_{t+1} \) are provided in Algorithms 8 and 9 of the online supplementary materials for completeness; see also Bhattacharya and Johndrow (2021) for perfect sampling of the components of \( \eta_{t+1} | \beta_{t,j}, \xi_t, \sigma_t^2 \) for the Horseshoe. Perfectly sampling \( \xi_{t+1} | \eta_{t+1} \) and \( \eta_{t+1,j} | \beta_{t,j}, \xi_t, \sigma_t^2 \) retains the \( O(n^2p) \) computational cost for the full blocked Gibbs sampler, though in practice this implementation is more expensive per iteration due to the computation of eigenvalue decompositions in lieu of Cholesky decompositions, and the computation of inverse of the confluent hypergeometric function of the second kind.

Proposition 1 gives a geometric drift condition for this variant of the blocked Gibbs sampler.

**Proposition 1** (Geometric drift). Consider the Markov chain \( (\beta_t, \xi_t, \sigma_t^2)_{t \geq 0} \) generated from the blocked Gibbs sampler in Equation (4), where each component \( \eta_{t+1,j} | \beta_{t,j}, \xi_t, \sigma_t^2 \) and \( \xi_{t+1} | \eta_{t+1} \) are sampled perfectly such that \( (\beta_t, \xi_t, \sigma_t^2) \rightarrow \eta_{t+1} \rightarrow (\beta_{t+1}, \xi_{t+1}, \sigma_{t+1}^2) \) with \( \eta_{t+1} \) intermediary for each \( t \geq 0 \). Assume the global shrinkage prior \( \pi_\zeta(\cdot) \) has a compact support. For \( m_j = \xi \beta_j^2 / (2\sigma_j^2) \), \( c \in (0, 1/2) \) and \( d \in (0, 1) \), define

\[
V(\beta, \xi, \sigma^2) = \sum_{j=1}^p m_j^c + m_j^d. \tag{5}
\]

Then for each \( v \geq 1 \), there exist some \( c \in (0, 1/2), d \in (0, 1) \) such that \( V \) is a drift function, that is, there exists \( \gamma \in (0, 1) \) and \( K \in (0, \infty) \) such that for all \( \beta_t \in \mathbb{R}^p, \xi_t > 0 \) and \( \sigma_t^2 > 0 \),

\[
\mathbb{E}[V(\beta_{t+1}, \xi_{t+1}, \sigma_{t+1}^2) | \beta_t, \xi_t, \sigma_t^2] \leq \gamma V(\beta_t, \xi_t, \sigma_t^2) + K.
\]

The drift (or ‘Lyapunov’) function in Equation (5), approaches infinity when any \( \beta_j \) approaches the origin or infinity. This ensures that the corresponding sublevel sets, defined by \( S(R) = \{ (\beta, \xi, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} : V(\beta, \xi, \sigma) < R \} \) for \( R > 0 \), exclude the pole at the origin and are bounded. Such geometric drift condition, in conjunction with Jacob et al. (2020 proposition 4), helps verify that the meeting times under the one-scale coupling have exponential tails and hence finite expectation.

**Proposition 2** Consider the blocked Gibbs sampler in Equation (4). In the setup of Proposition 1, assume that the global shrinkage prior \( \pi_\zeta(\cdot) \) has a compact support. Write \( Z_t = (\beta_t, \xi_t, \sigma_t^2) \) and \( Z_t = (\beta_t, \xi_t, \sigma_t^2) \). Consider the one-scale coupling given by \( (Z_t, \tilde{Z}_t) \rightarrow (\eta_{t+1}, \tilde{\eta}_{t+1}) \rightarrow (Z_{t+1}, \tilde{Z}_{t+1}) \) where \( (\eta_{t+1}, \tilde{\eta}_{t+1}) \) are maximally coupled component-wise, and \( (Z_{t+1}, \tilde{Z}_{t+1}) \) are coupled using common random numbers. Denote the meeting time by \( \tau_Z := \inf \{ t \geq 0 : Z_t = \tilde{Z}_t \} \). Then

\[
\mathbb{P}(\tau_Z > t) \leq A_0 k_0 t^c \quad \text{for some constants } A_0 \in (0, \infty) \text{ and } k_0 \in (0, 1), \text{ and for all } t \geq 0.
\]
Exponentially timed meeting times immediately imply that the marginal \((Z_t)_{t \geq 0}\) chain of the blocked Gibbs sampler in Section 2.1 is geometrically ergodic. This follows from noting that Jacob et al. (2020 proposition 4) is closely linked to a minorization condition (Rosenthal, 1995), and details are included in proposition 4.10 of the online supplementary materials. We emphasize that studying the marginal chain \((Z_t)_{t \geq 0} = (\beta_t, \xi_t, \sigma_t^2)_{t \geq 0}\), where \(\eta_t\) is viewed as an intermediate quantity, suffices to analyse the full chain of interest, \((C_t)_{t \geq 0} = (\eta_t, Z_t)_{t \geq 0}\). To see this, we briefly recall some facts about two-component Gibbs samplers (Diaconis et al., 2008, 2010). Let \(\pi_Z\), and \(\pi_\eta\) denote the distributions of \(Z_t\) and \(C_t\), respectively, and let \(\pi_Z\) and \(\pi_\eta\) denote the stationary distribution of the chains \((Z_t)_{t \geq 0}\) and \((C_t)_{t \geq 0}\), respectively. Then by Diaconis et al. (2008 lemma 2.4), TV\((\pi_{Z_t}, \pi_Z) \leq TV(\pi_{t_\eta}, \pi_\eta)\) for all \(t \geq 1\), and therefore the exponential rate of convergence of \((Z_t)\) is inherited by the full chain.

Furthermore, we note that the meeting time \(\tau_Z\) in Proposition 2 for the marginal chain is equal to the meeting time \(\tau := \inf\{t \geq 0 : C_t = \tilde{C}_t\}\) of the full chain. To see this, recall that under common random numbers, \(Z_t = \tilde{Z}_t\) if and only if \(\eta_t = \tilde{\eta}_t\). This implies \(Z_t = \tilde{Z}_t\) if and only if \(C_t = \tilde{C}_t\). Thus Proposition 2 applies to the meeting time \(\tau\) of the full chain.

For a class of full chains targeting shrinkage priors including the Bayesian LASSO, the normal-Gamma prior (Griffin & Brown, 2010), and Dirichlet-Laplace prior (Bhattacharya et al., 2015), Khare and Hobert (2013) and Pal and Khare (2014) have proven geometric ergodicity based on drift and minorization arguments (Meyn & Tweedie, 1993; Roberts & Rosenthal, 2004). For the Horseshoe prior, Johndrow et al. (2020) has recently established geometric ergodicity. In the high-dimensional setting with \(p > n\) for the Horseshoe, the proof of Johndrow et al. (2020) required truncating the prior on each \(\eta_j\) below by a small \(\epsilon > 0\) to guarantee the uniform bound \(\eta_j \geq \epsilon\). Our proof of geometric ergodicity for Half-t(\(\nu\)) priors including the Horseshoe (\(\nu = 1\)) works in both low- and high-dimensional settings, without any modification of the priors on the \(\eta_j\). In parallel work, Bhattacharya et al. (2022) have also established geometric ergodicity for the Horseshoe prior without requiring such truncation.

Remark 1 Our proofs of exponentially timed meeting times and geometric ergodicity generalize to a larger class of priors satisfying some moment conditions on the full conditionals of \(\eta_j\). Consider a compactly supported prior \(\pi_\xi\) on \(\xi\), and a prior \(\pi_\eta\) on each \(\eta_j\). Then the unnormalized density of the full conditional of each \(\eta_j\) is given by \(\pi(\eta_j|m_j = \xi \beta_j^2/(2\sigma_j^2)) \propto \eta_j^{1/2} e^{-m_j\eta} \pi_\eta(\eta_j)\). Consider the following assumptions on \(\pi_\eta\).

1. For some \(0 < c < 1/2\), there exist \(0 < e < \Gamma(\frac{1}{2} - c)^{-1} \sqrt{\pi}\) and \(K_{c,e}^{(1)} < \infty\) such that \(\mathbb{E}[\eta_j^e|m_j] \leq e m_j^{-c} + K_{c,e}^{(1)}\) for all \(m_j > 0\).

2. For some \(0 < d < 1\), there exist \(0 < e < 2^d\) and \(K_{d,e}^{(2)} < \infty\) such that \(\mathbb{E}[\eta_j^{-d}|m_j] \leq e m_j^d + K_{d,e}^{(2)}\) for all \(m_j > 0\).

Then the blocked Gibbs sampler in Equation (4) is geometrically ergodic.

We now consider the performance of Algorithm 1 on synthetic data sets. In particular, we empirically illustrate that the rate \(n_0\) in Proposition 2 can tend to 1 exponentially fast as dimension \(p\) increases. The design matrix \(X \in \mathbb{R}^{n \times p}\) is generated with \([X]_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)\), and a response vector as \(y \sim \mathcal{N}(X \beta_s, \sigma^2 I_n)\), where \(\beta_s \in \mathbb{R}^p\) is the true signal and \(\sigma \geq 0\) is the true error standard deviation. We choose \(\beta_s\) to be sparse such that given sparsity parameter \(s\), \(\beta_{s,j} = 2^{(9-s)/4}\) for \(1 \leq j \leq s\) and \(\beta_{s,j} = 0\) for all \(s > j\). Figure 2 shows the meeting times \(\tau\) of coupled Markov chains (with a lag \(L = 1\)) targeting the Horseshoe posterior (\(\nu = 1\)) with \(a_0 = b_0 = 1\) under Algorithm 1.
We consider $n = 100$, $s = 10$, $\sigma_s = 0.5$ and vary the dimension $p$. For each dimension $p$, we simulate 100 meeting times under the one-scale coupling for independently generated synthetic data sets. We initialize chains independently from the prior distribution. For the $\xi$ updates, we use a proposal step-size of $\sigma_{\text{MRTH}} = 0.8$ as in Johndrow et al. (2020). Figure 2, with a vertical axis on a logarithmic scale, shows that the meeting times increase exponentially and have heavier tails with increasing dimension.

3 | COUPLING STRATEGIES FOR HIGH-DIMENSIONAL SETTINGS

To address the scaling issues observed with the one-scale coupling in Section 2.3, we develop a two-scale coupling in Section 3.1 that delivers better empirical performance. We also offer some formal results supporting the empirical improvements. In Section 3.2 we develop other related coupling strategies which have exponentially tailed meeting times and have similar empirical performance as the two-scale coupling. In Section 3.3, we investigate the impact of the degree of freedom $\nu$ for Half-$t(\nu)$ priors on our two-scale coupling. We find that Half-$t(\nu)$ priors with higher degrees of freedom $\nu > 1$ can give similar statistical performance to the Horseshoe ($\nu = 1$), while enjoying orders of magnitude computational improvements through shorter meeting times.

3.1 | A two-scale coupling

We develop a two-scale coupling algorithm, which slightly differs from the terminology used in stochastic dynamics and in analysis of pre-conditioned HMC (Bou-Rabee & Eberle, 2020). It is also closely linked to the delayed one-shot coupling of Roberts and Rosenthal (2002), used for example in Beskos and Roberts (2005) in an implementation of perfect sampling for truncated Gaussian distributions, which refers to coupled kernels that attempt exact meetings when the chains are close, and aim for a contraction when the chains are far. The motivation for this...
construction is that in Algorithm 1, the components of \((\eta_t, \tilde{\eta}_t)\) that fail to meet instead evolve independently. As dimension grows, the number of components of \((\eta_t, \tilde{\eta}_t)\) that fail to meet also tends to grow, making it increasingly difficult for all components to exactly meet on subsequent iterations. In the two-scale coupling, we only attempt to obtain exact meetings when the associated probability is large enough. This is done by constructing a metric \(d\) and a corresponding threshold parameter \(d_0 \geq 0\) such that when the current states are \(d\)-close with \(d(C_t, \tilde{C}_t) \leq d_0\), we apply Algorithm 2 and try to obtain exact meetings. When \(d(C_t, \tilde{C}_t) > d_0\), we instead employ common random numbers to sample \((\eta_t, \tilde{\eta}_t)\).

Our chosen metric \(d\) on \(\mathbb{R}_{>0}^p \times \mathbb{R}^p \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}\) is

\[
d(C_t, \tilde{C}_t) = \mathbb{P}(\eta_{t+1} \neq \tilde{\eta}_{t+1} | C_t, \tilde{C}_t),
\]

(6)

where recall that \(C_t = (\eta_t, \beta_t, \sigma_t^2, \xi_t)\), \(\tilde{C}_t = (\tilde{\eta}_t, \tilde{\beta}_t, \tilde{\sigma}_t^2, \tilde{\xi}_t)\), and \((\eta_{t+1}, \tilde{\eta}_{t+1})|C_t, \tilde{C}_t\) is sampled using the coupled slice sampler in Algorithm 2. As the coupling in Algorithm 2 is independent component-wise, we can obtain a simplified expression involving only univariate probabilities as

\[
d(C_t, \tilde{C}_t) = 1 - \prod_{j=1}^p \mathbb{P}(\eta_{t+1,j} = \tilde{\eta}_{t+1,j} | \eta_{t,j}, \tilde{\eta}_{t,j}, m_{t,j}, \tilde{m}_{t,j}).
\]

(7)

where \(m_{t,j} = \xi_j \beta_{t,j}^2/(2\sigma_t^2)\) and \(\tilde{m}_{t,j} = \tilde{\xi}_j \tilde{\beta}_{t,j}^2/(2\tilde{\sigma}_t^2)\). Under this metric \(d\) and a threshold \(d_0 \in [0, 1]\), an exact meeting is attempted when \(\mathbb{P}(\eta_{t+1} = \tilde{\eta}_{t+1} | C_t, \tilde{C}_t) \geq 1 - d_0\). Once \(\eta_{t+1}\) and \(\tilde{\eta}_{t+1}\) coincide, and if the scalars \(\xi_{t+1}\) and \(\tilde{\xi}_{t+1}\) also subsequently coincide, then the entire chains meet. When \(d_0 = 1\), the two-scale coupling reduces to the one-scale coupling of Algorithm 1 since the maximal coupling slice sampler (Algorithm 2) is invoked at each step. On the other hand, when threshold \(d_0 = 0\), we always apply the common random numbers. Then the chains \(C_t, \tilde{C}_t\) may come arbitrarily close but will never exactly meet. Empirically we find that different values of the threshold \(d_0\) away from 0 and 1 give similar meeting times. Simulations concerning these choices can be found in section 2.2 of the online supplementary materials.

We now discuss computation of the metric \(d(C_t, \tilde{C}_t)\). Since the probability in Equation (6) is unavailable in closed form, we can resort to various approximations. One option is to evaluate the one-dimensional integrals in Equation (7) using numerical integration. Alternatively, one can form the Monte Carlo based estimate

\[
\widehat{d}_R^{(1)}(C_t, \tilde{C}_t) = \frac{1}{R} \sum_{r=1}^R \mathbb{1}\{\eta_{t+1}^{(r)} \neq \tilde{\eta}_{t+1}^{(r)}\},
\]

(8)

where \((\eta_{t+1}^{(r)}, \tilde{\eta}_{t+1}^{(r)})\) are sampled independently for \(r = 1, \ldots, R\) using Algorithm 2. We recommend a Rao–Blackwellized estimate by combining Monte Carlo with analytical calculations of integrals as

\[
\widehat{d}_R^{(2)}(C_t, \tilde{C}_t) = 1 - \prod_{j=1}^p \left( \frac{1}{R} \sum_{r=1}^R \mathbb{P}\left(\eta_{t+1,j} = \tilde{\eta}_{t+1,j} | U_j^{(r)}, \tilde{U}_j^{(r)}, m_{t,j}, \tilde{m}_{t,j}\right) \right).
\]

(9)

where \((U_j^{(r)}, \tilde{U}_j^{(r)})\) are sampled independently for \(r = 1, \ldots, R\) using Algorithm 2 and each \(\mathbb{P}(\eta_{t+1,j} = \tilde{\eta}_{t+1,j} | U_j^{(r)}, \tilde{U}_j^{(r)}, m_{t,j}, \tilde{m}_{t,j})\) is calculated analytically based on meeting probabilities from
Algorithm 2. Further metric calculation details are in section 2.3 of the online supplementary materials. For a number of samples \( R \), estimates Equations (8) and (9) both have computation cost of order \( pR \). Compared to the estimate in Equation (8), the estimate in Equation (9) has lower variance and faster run-time as it only involves sampling uniformly distributed random numbers. It suffices to choose a small number of samples \( R \), and often we take \( R = 1 \). Indeed it appears unnecessary to estimate \( d(C_t, \tilde{C}_t) \) accurately, as we are only interested in comparing that distance to a fixed threshold \( d_0 \in [0, 1] \). Often the trajectories \( (d(C_t, \tilde{C}_t))_{t \geq 0} \) initially take values close to 1, and then sharply drop to values close to 0. This leads to the estimates in Equations (8) and (9) having low variance. Henceforth, we will use the estimate in Equation (9) in our experiments with two-scale couplings, unless specified otherwise.

Overall the two-scale coupling kernel has twice the \( \Theta(n^2p) \) cost of the single chain kernel in Equation (4). When \( C_t \neq \tilde{C}_t \), we calculate a distance estimate and sample from a coupled slice sampling kernel. Calculating the distance estimate involves sampling \( 2pR \) uniforms and has \( \Theta(p) \) cost. Coupled slice sampling with maximal couplings (Algorithm 2) or common random numbers both have expected or deterministic cost \( \Theta(p) \). The remaining steps of the two-scale coupling match the one-step coupling in Algorithm 1.

We now consider the performance of the two-scale coupling on synthetic data sets. The setup is identical to that introduced in Section 2.3, where for each dimension \( p \) we simulate 100 meeting times based on independently generated synthetic data sets. We target the Horseshoe posterior \((\nu=1)\) with \( a_0 = b_0 = 1 \). Figure 3 shows the meeting times \( \tau \) of coupled Markov chains (with a lag \( L = 1 \)), for both the one-scale coupling and the two-scale coupling with \( R = 1 \) and threshold \( d_0 = 0.5 \). Figure 3 shows that the two-scale coupling can lead to orders of magnitude reductions in meeting times compared to the one-scale coupling.

We next present some preliminary results which hint as to why the two-scale coupling leads to the drastic reduction in meeting times. We focus on the coupling of the \((\eta, \tilde{\eta})\) full conditionals, as the one-scale and two-scale algorithms differ only at this step. First, we show that for any monotone function \( h : (0, \infty) \to \mathbb{R} \) and all components \( j = 1, \ldots, p \), the expected distance \( \mathbb{E}[|h(\eta_{t+j}) - h(\tilde{\eta}_{t+j})|] \) is the same under both common random numbers and maximal coupling.
Proposition 3 Consider the setup of Proposition 2, such that the $\eta$ full conditionals are sampled perfectly. Then for any monotone function $h : (0, \infty) \to \mathbb{R}$ and all components $j = 1, \ldots, p$,

$$\mathbb{E}_{\text{max}}[|h(\eta_{t+1,j}) - h(\tilde{\eta}_{t+1,j})| \mid m_t, \tilde{m}_t] = \mathbb{E}_{\text{CRN}}[|h(\eta_{t+1,j}) - h(\tilde{\eta}_{t+1,j})| \mid m_t, \tilde{m}_t],$$

where $\mathbb{E}_{\text{max}}$ and $\mathbb{E}_{\text{CRN}}$ correspond to expectations under the maximal coupling and CRN coupling, respectively, of $(\eta_{t+1}, \tilde{\eta}_{t+1})$ given $m_t$ and $\tilde{m}_t$.

Proposition 3 implies that such single-step expectations alone do not distinguish the behaviour of CRN from maximal couplings. This compels us to investigate other distributional features of $|h(\eta_{t+1,j}) - h(\tilde{\eta}_{t+1,j})|$ under either coupling to possibly tease apart differences in their behaviour on high probability events. Focusing on the Horseshoe prior and making the choice $h(x) = \log(1 + x)$, we uncover a distinction between the tail behaviour of the two couplings which can substantially accumulate over $p$ independent coordinates. We offer an illustration in a stylized setting below.

Proposition 4 For the Horseshoe prior ($v = 1$), consider when $m_t = \mu 1_p \in (0, \infty)^p$ and $\tilde{m}_t = \tilde{\mu} 1_p \in (0, \infty)^p$ for some positive scalars $\mu$ and $\tilde{\mu}$ with $\tilde{\mu} \neq \mu$. Then under a CRN coupling,

$$\mathbb{P}\left(\max_{j=1}^p \log(1 + \eta_{t+1,j}) - \log(1 + \tilde{\eta}_{t+1,j}) > |\log \mu - \log \tilde{\mu}| \mid m_t, \tilde{m}_t\right) = 0$$

for all $p \geq 1$. Define a positive constant $L = \log(1 + 1/\max\{\mu, \tilde{\mu}\})/2$. Under maximal coupling, for any $\alpha \in (0, 1)$ and any $\alpha' \in (\alpha, 1)$, there exists a constant $D > 0$ which does not depend on $p$ such that for all $p > 1$,

$$\mathbb{P}\left(\max_{j=1}^p \log(1 + \eta_{t+1,j}) - \log(1 + \tilde{\eta}_{t+1,j}) > \log(\alpha L) + \log \log p + \log \log |m_t, \tilde{m}_t|\right) > 1 - e^{-Dp^{1-\alpha'}}. $$

Extensions of Proposition 4 which allow different components of $m_t$ and $\tilde{m}_t$ to take different values under some limiting assumptions are omitted here for simplicity. This stylized setting already captures why the one-scale algorithm may result in much larger meeting times in high dimensions compared to the CRN-based two-scale algorithm. We note that Proposition 4 remains informative whenever $|\log \mu - \log \tilde{\mu}| \ll \log \log p$, even when $\log \log p$ is not large. For example, consider $\mu = \tilde{\mu} e^\delta$ for some small $\delta > 0$ with $\delta \ll \log \log p$. Under CRN, $\log(1 + \eta_{t+1,j})$ and $\log(1 + \tilde{\eta}_{t+1,j})$ remain $\delta$-close almost surely for all components $j$. In contrast, under the maximal coupling, at least one component will have larger than $\log(\alpha L) + \log \log p$ deviations with high probability. Since $\alpha$ and $L = \log(1 + 1/\mu)/2$ do not depend on $p$ or $\delta$, $\log(\alpha L) + \log \log p \gg \delta$ for fixed $\mu, \tilde{\mu} > 0$. These larger deviations between some components of $\eta_{t+1}$ and $\tilde{\eta}_{t+1}$ under maximal couplings push the two chains much further apart when sampling $(m_{t+1}, \tilde{m}_{t+1}) \mid \eta_{t+1}, \tilde{\eta}_{t+1}$, resulting in meeting probabilities that are much lower over multiple steps of the algorithm. Overall, Propositions 3 and 4 show that multistep kernel calculations taking into account such high-probability events may be necessary to further distinguish the relative performances of one-scale and two-scale couplings. A full analytic understanding of this difference in performance remains an open problem.

Another way of understanding the benefits of two-scale coupling would be to find a distance under which the coupled chains under common random numbers exhibit a contraction. The CRN coupling part of Proposition 4 hints that $\tilde{d}(m_t, \tilde{m}_t) := \sum_{j=1}^p |\log m_{t,j} - \log \tilde{m}_{t,j}|$ may be a natural metric. Section 2.1 of the online supplementary materials contains discussions and related simulations comparing this metric with alternatives. Proposition 5 verifies that such metric bounds the total variation distance between one-step transition kernels.
Proposition 5 Let $P$ denote the Markov transition kernel associated with the update from $(\beta_t, \xi_t, \sigma_t)$ to $(\beta_{t+1}, \xi_{t+1}, \sigma_{t+1})$ for the Gibbs sampler in Equation (4). Let $Z = (\beta, \xi, \sigma^2)$, $\tilde{Z} = (\tilde{\beta}, \tilde{\xi}, \tilde{\sigma}^2)$, $m_j = \xi\tilde{\beta}_j^2/(2\sigma^2)$, and $\tilde{m}_j = \tilde{\xi}\tilde{\beta}_j^2/(2\tilde{\sigma}^2)$ for $j = 1, \ldots, p$. Then,

$$TV(P(Z, \cdot), P(\tilde{Z}, \cdot))^2 \leq \frac{(1 + \nu)(e - 1) - d(Z, \tilde{Z})}{4}.$$  

(10)

By Proposition 5, it suffices to show that $(d(Z_t, \tilde{Z}_t))_{t \geq 0}$ contracts under CRN to a sufficiently small value such that the upper bound in Equation (10) is strictly less than 1. This would ensure that CRN coupling will bring the marginal chains close enough that the probability of one-step max coupling resulting in a meeting is high. Note that Equation (10) corresponds to the total variation distance of the full chain, so under this setup we would attempt to maximally couple the full vectors $\eta_{t+1}$ and $\tilde{\eta}_{t+1}$ (rather than maximally coupling component-wise) when close. Analytical calculations to establish such a contraction, and to understand how the contraction rate scales with dimension and other features of the problem such as the choice of prior, sparsity and signal to noise ratio requires further work.

3.2 Alternative coupling strategies

The rapid meeting of the two-scale coupling in simulations recommends its use. However, we were unable to establish finite expected meeting times with this coupling. We now describe several small modifications of the two-scale coupling that have similar empirical performance but for which we can prove that the meeting time distribution has exponential tails.

Our basic approach is to maximally couple each component $(\eta_{t,j}, \tilde{\eta}_{t,j})$ until a failed meeting $\{\eta_{t,j} \neq \tilde{\eta}_{t,j}\}$ occurs, after which we switch to CRN for the remaining components. The ordering of the components $j = 1, \ldots, p$ can be deterministic or randomized at each iteration. Under this construction, only one component of $(\eta_t, \tilde{\eta}_t)$, corresponding to the failed meeting, will evolve independently; the other components will either meet or evolve under CRN. Therefore, we expect to obtain benefits similar to the two-scale coupling in high dimensions. An apparent advantage is that we can bypass the choice of metric and associated threshold. While the switch-to-CRN coupling has the same $O(n^2p)$ cost per iteration as the two-scale coupling, it can give faster run-times in practice when the dimension $p$ is large compared to the number of observations $n$ as it avoids repeated calculation of a metric.

We report the performance of the switch-to-CRN coupling with the ordering of the components randomized at each iteration. The setup is identical to that in Section 3.1, where for each dimension $p$ we simulate 100 meeting times based on independently generated synthetic data sets. Figure 4 shows that the switch-to-CRN coupling leads to similar meeting times as the two-scale coupling.

We can establish that the meeting times under this switch-to-CRN coupling have exponential tails and hence finite expectation.

Proposition 6 Consider the blocked Gibbs sampler in Equation (4). We follow the setup in Proposition 1, and assume that the global shrinkage prior $\pi(\cdot)$ has compact support. Write $Z_t = (\beta_t, \xi_t, \sigma_t^2)$, $\tilde{Z}_t = (\tilde{\beta}_t, \tilde{\xi}_t, \tilde{\sigma}_t^2)$, $C_t = (\eta_t, \tilde{\beta}_t, \tilde{\xi}_t, \tilde{\sigma}_t^2)$ and $\tilde{C}_t = (\tilde{\eta}_t, \tilde{\beta}_t, \tilde{\xi}_t, \tilde{\sigma}_t^2)$. Consider the switch-to-CRN coupling given by $(Z_t, \tilde{Z}_t) \mapsto (\eta_{t+1}, \tilde{\eta}_{t+1}) \mapsto (Z_{t+1}, \tilde{Z}_{t+1})$ where $(\eta_{t+1}, \tilde{\eta}_{t+1})$ are maximally coupled component-wise (under any fixed or random ordering of the components) until the first failed meeting, after which common random numbers are employed,
and \((Z_{t+1}, \tilde{Z}_{t+1})\) are coupled using common random numbers. Denote the meeting time by 
\[
\tau := \inf\{t \geq 0 : C_t = \tilde{C}_t\} = \inf\{t \geq 0 : Z_t = \tilde{Z}_t\}. 
\]
Then \(P(\tau > t) \leq A_1 \kappa_1^t\) for some constants \(A_1 \in (0, \infty)\) and \(\kappa_1 \in (0, 1)\), and for all \(t \geq 0\).

The equality of the meeting times for the \((Z_t)\) and the \((C_t)\) chains follows from a similar argument as after Proposition 2. As in Proposition 2, our proof of Proposition 6 implies a rate \(\kappa_1\) which tends to 1 exponentially as dimension \(p\) increases. To obtain more favourable rates with respect to the dimension as hinted by Figure 4 would require a better understanding of the CRN coupling, for example establishing a contraction in some parts of the state space as discussed above.

Overall, Figures 3 and 4 show significant empirical improvements in using the two-scale coupling and related strategies, compared to the one-scale coupling of Section 2.3. Therefore, we recommend the use of two-scale couplings and variants such as the switch-to-CRN coupling for practical high-dimensional applications. Furthermore, in settings where estimating the metric \(d\) is computationally expensive or analytically intractable, variants such as the switch-to-CRN coupling may offer faster run-time per iteration. Henceforth, we will use the two-scale coupling of Section 3.1 in our experiments and GWAS data set applications, unless specified otherwise.

### 3.3 Computational and statistical impact of degree of freedom ν

We empirically investigate the impact of the degrees of freedom \(\nu \geq 1\) for Half-\(t(\nu)\) priors. Higher degrees of freedom \(\nu\) corresponds to priors on \(\beta\) which have stronger shrinkage towards zero (Proposition 1.1). We consider meeting times resulting from the two-scale coupling of Section 3.1, as well as the statistical performance of the posteriors. On the computation side, recent work on convergence analysis (Qin & Hobert, 2019a, b) has highlighted the impact of shrinkage on convergence of simpler MCMC algorithms. Under a common random numbers coupling, Qin and Hobert (2019b) have studied the convergence of the Gibbs sampler of Albert and Chib (1993) for Bayesian probit regression, proving dimension-free convergence rates for priors with sufficiently strong shrinkage towards zero. On the statistical estimation side, Half-\(t(\nu)\) priors have been long proposed (Carvalho et al., 2009, 2010; Gelman, 2006) for Bayesian hierarchical models. van der Pas et al. (2014, 2017) have established (near) minimax optimal estimation for the Horseshoe (\(\nu = 1\)) in the normal means model. Ghosh and Chakrabarti (2017) have extended the results of van der Pas et al. (2014) to show minimax optimality of a wider class of priors including Half-\(t(\nu)\) priors.
Recently, Song (2020) has established that the degree of freedom $\nu$ can impact the multiplicative constant in the posterior contraction rate for the Normal means model. Posterior contraction rates of continuous shrinkage priors in the regression setting beyond the Normal means model would deserve further investigation.

We consider the meeting times obtained with Half-$t(\nu)$ priors under the two-scale coupling (with $R = 1$ and threshold $d_0 = 0.5$) on synthetic data sets. The synthetic data sets are generated as per Figure 2 of Section 2.3. Figure 5a is then based on 20 synthetic data sets generated independently for different degrees of freedom $\nu \in [1, 2]$. Figure 5a plots meeting times $\tau$ of 1-lag coupled Markov chains against dimension. It indicates that even a small increase in the degree of freedom $\nu > 1$ compared to the Horseshoe ($\nu = 1$) can lead to orders of magnitude computational improvements through shorter meeting times.

We also consider the statistical performance of Bayesian shrinkage regression with Half-$t(\nu)$ priors on synthetic data sets in Figure 5b. The synthetic data sets are generated as per Figure 2 of Section 2.3. We use the blocked Gibbs sampler of Algorithm 1 to draw samples from the posteriors corresponding to different degrees of freedom $\nu$, by simulating chains of length 1000, with a burn-in period chosen using the coupling-based total variation distance upper bounds of Biswas et al. (2019). Specifically we employ a burn-in of 300 steps for $\nu \geq 1.2$ and of 600 steps for $\nu = 1$. Figure 5b shows mean squared error (MSE) for different values of $\nu \geq 1$. For a fixed synthetic data set, the MSE in Figure 5b is calculated as $||\beta_* - \hat{\beta}||_2^2/p$ where $\hat{\beta} := \sum_{t=1}^{1000} \beta_t/1000$, is the MCMC estimator obtained from the Markov chain after burn-in. The error bars in Figure 5b are generated by simulating synthetic data sets independently 100 times and each time calculating the corresponding MSE from a Markov chain for different values of $\nu \in [1, 8]$. The MSE of the LASSO is also included, with regularization parameter chosen with cross-validation using the glmnet package (Friedman et al., 2010). For $\nu \in [1, 2]$, we observe a lower MSE using Half-$t(\nu)$ priors compared to the LASSO. The grey plots in Figure 5b show that much larger $\nu \in [4, 6, 8]$ can lead to higher MSE, as the corresponding posteriors are strongly concentrated about zero and less able to identify non-null signals. Overall, Figure 5b suggests that Half-$t(\nu)$ priors with degrees of freedom $\nu \in [1, 2]$ can result in comparable statistical performance while much larger values of $\nu$ should be discouraged.
4 | RESULTS ON GWAS DATA SETS

Section 3.3 suggests that Half-$t(\nu)$ priors with higher degrees of freedom $\nu > 1$ can give similar statistical performance to the Horseshoe ($\nu = 1$), while allowing orders of magnitude computational improvements.

Motivated by this observation, we apply our algorithms to genome-wide association study (GWAS) data sets using $\nu = 2$. We consider a bacteria data set (Bühlmann et al., 2014) with $n = 71$ observations corresponding to production of the vitamin riboflavin, and $p = 4088$ covariates corresponding to single nucleotide polymorphisms (SNPs) in the genome. We also consider a maize data set (Johndrow et al., 2020; Liu et al., 2016; Romay et al., 2013; Zeng & Zhou, 2017) with $n = 2266$ observations corresponding to average number of days taken for silk emergence in different maize lines, and $p = 98,385$ covariates corresponding to SNPs. Bayesian methods are well-suited to such GWAS data sets, as they provide interpretable notions of uncertainty through marginal posterior probabilities which allow for multimodality and enable the use of prior information (e.g. Guan & Stephens, 2011; Zhou et al., 2013). Furthermore, heavy-tailed continuous shrinkage priors can be particularly effective in the GWAS setting, as the associations between SNPs and the phenotype are expected to be sparse, and such priors have shown competitive empirical performance in polygenic prediction (Ge et al., 2019).

For both data sets, we target the posterior induced by a Half-$t(2)$ prior with $a_0 = b_0 = 1$. We use the two-scale coupling with $R = 1$ and threshold $d_0 = 0.5$, and initialize chains independently from the prior. For the $\xi$ updates, we use a Metropolis–Rosenbluth–Teller–Hastings proposal step-size of $\sigma_{\text{MRTH}} = 0.8$. Based on 100 independent coupled chains, Figure 6 shows upper bounds on the total variation distance to stationarity, $TV(\pi_t, \pi)$ as a function of $t$, using $L$-lag couplings with $L = 750$ and $L = 200$ for the maize and riboflavin data sets, respectively. The results indicate that these Markov chains converge to the corresponding posterior distributions in less than 1000 and 500 iterations, respectively. The lags $L = 750$ and $L = 200$ were selected based on some preliminary runs of the coupled chains to obtain informative total variation bounds, which incurs an additional preliminary cost. Any choice of lag results in valid upper bounds, but only large enough lags lead to upper bounds that are in the interval $[0, 1]$ even for small iterations $t$.

(a) Riboflavin dataset with $n = 71$ and $p = 4,088$

(b) Maize dataset with $n = 2,266$ and $p = 98,385$

FIGURE 6 | $L$-Lag coupling-based upper bounds on $TV(\pi_t, \pi)$, the distance between the chain at time $t$ and its limiting distribution, in Bayesian shrinkage regression with Half-$t(2)$ prior on genome-wide association study data sets. Maize data set: $n = 2266$ and $p = 98,385$; Riboflavin data set: $n = 71$ and $p = 4088$
In these GWAS examples, the run-time per iteration may be significant. For the maize data set, on a 2015 MacBook Pro with a 2.7 GHz dual-core Intel Core i5 processor, each iteration of the coupled chain takes approximately 60 s, and running one coupled chain until meeting can take more than one day. This run-time is dominated by the calculation of the weighted matrix cross-product \( X \text{Diag}(\eta)^{-1}X^T \), with an \( \mathcal{O}(n^2p) \) cost. In this setting, a scientist wanting to run chains for \( 10^4 \) or \( 10^5 \) iterations may have to wait for weeks or months. Using the proposed couplings, a scientist with access to parallel computing can be confident that samples obtained after \( 10^3 \) iterations would be indistinguishable from perfect samples from the target. For the riboflavin data set, each iteration of the coupled chain takes approximately 0.1 s, and running one coupled chain until meeting takes only 10 to 20 s. The run-time there is dominated by the component-wise \( \eta \) updates which have \( \mathcal{O}(p) \) cost. In the riboflavin example, the coupling-based diagnostics enable the scientist to perform reliable Bayesian computation directly from personal computers. Here we have demonstrated the use of the proposed couplings for convergence diagnostics, but they can also deliver unbiased estimators of posterior expectations (Jacob et al., 2020); section 3 of the online supplementary materials presents some results on unbiased estimation for the riboflavin data set. Overall, the GWAS examples illustrate that couplings of MCMC algorithms can aid practitioners in high-dimensional settings.

5 | DISCUSSION

We have introduced couplings of Gibbs samplers for Bayesian shrinkage regression with Half-t(\( \nu \)) priors. The proposed two-scale coupling is operational in realistic settings, including a GWAS setting with \( n \approx 2000, p \approx 100,000 \).

First, our work participates in a wider effort to apply MCMC algorithms and coupling techniques to ever more challenging settings (Lee et al., 2020b; Middleton et al., 2020; Ruiz et al., 2021; Trippe et al., 2021; Xu et al., 2021). In particular, the short meeting times obtained here vindicate the use of coupling techniques in multimodal, high-dimensional sampling problems. This relates to questions raised in some comments of Lee et al. (2020a) and Paulin (2020) in the discussion of Jacob et al. (2020), and shows that the dimension of the state space may not necessarily be the most important driver of the performance of MCMC algorithms and couplings thereof.

Second, we have applied L-lag couplings (Biswas et al., 2019) to obtain upper bounds on the total variation distance between the Markov chain at a finite time and its stationary distribution. This allows empirical investigations of the mixing time and how it varies with the inputs of the problem, including number of observations, number of covariates, signal to noise ratio and sparsity. We find that coupling techniques constitute a convenient non-asymptotic tool to monitor the performance of MCMC algorithms.

Third, we observe that Half-t(\( \nu \)) priors with degrees of freedom \( \nu \) higher than one give similar statistical estimation performance to the Horseshoe while providing significant computational advantages. This contributes towards the discussion on the impact of the prior on the trade-off between statistical estimation and computational feasibility, in the setting of high-dimensional Bayesian regression.

The following questions arise from our work.

1. **Convergence of the blocked Gibbs sampler.** Short meeting times suggest that the blocked Gibbs sampler converges quickly, even in high dimensions. This motivates a more formal study of the convergence of that Markov chain, to understand how the convergence rate varies with
features of the data generating process and of the prior. The Lyapunov function in Proposition 1 and the two-scale coupling may prove useful for such analysis. Our initial work in this area suggests that finding a convenient metric that gives sharp bounds on the metric $d$ used here, while simultaneously being amenable to theoretical analysis, will be a key step.

2. **Alternative coupling algorithms.** Section 3.2 indicates that many coupling strategies are available in the present setting, and more generally for various Gibbs samplers. For example, we could combine the switch-to-CRN coupling with the two-scale coupling to form other coupling strategies. Under the two-scale coupling setup, we could apply the switch-to-CRN coupling only when the chains are close with respect to a chosen metric, and employ CRN couplings when far away. Alternatively, we could always apply the switch-to-CRN coupling, and estimate component-wise meeting probabilities to select the order of the updates, for example such that components more likely to meet are updated first. Some couplings may result in shorter meeting times for the Horseshoe prior. This may allow the Horseshoe prior to remain competitive with Half-$t(\nu)$ priors with higher degrees of freedom. In our experiments, we did not find ways to obtain shorter meetings for the Horseshoe, but we certainly cannot rule out that possibility. If one could obtain short meetings for the Horseshoe in high dimensions, the apparent trade-off between statistical performance and computing cost identified in the present work would disappear.

3. **Interplay between posterior concentration and MCMC convergence.** For Bayesian regression with spike-and-slab priors, Yang et al. (2016) and Atchadé (2021) have shown that posterior contraction can aid the convergence of MCMC in high dimensions. The performance of the proposed couplings motivates similar investigations for continuous shrinkage priors.

**ACKNOWLEDGEMENTS**

The authors thank Xiaolei Liu and Xiang Zhou for sharing the Maize GWAS data, and Yves F. Atchadé, Qian Qin and Neil Shephard for helpful comments. Niloy Biswas gratefully acknowledges support by a GSAS Merit Fellowship and a Two Sigma Fellowship Award. Anirban Bhattacharya gratefully acknowledges support by the National Science Foundation through an NSF CAREER award (DMS-1653404). Pierre E. Jacob gratefully acknowledges support by the National Science Foundation through grants DMS-1712872 and DMS-1844695. James E. Johndrow acknowledges conversations with Persi Diaconis and Jonathan Mattingly that provided early motivation for this project. The maize data set GWAS simulations in this article were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

**DATA AVAILABILITY STATEMENT**

Simulation details for all the synthetic datasets are mentioned in the main text. The corresponding code scripts in R are available from [https://urldefense.com/v3/__https://github.com/niloyb/CoupledHalfT__/!!N11eV2iwtfslVnb2AWti1MLJXFV0V_9trXnk7phdVZVOGTc_i6lBGDSKo3VulRkIoZvB4SOZRiwwMRMXjWRyc4ZVcb85Xw$](https://urldefense.com/v3/__https://github.com/niloyb/CoupledHalfT__/!!N11eV2iwtfslVnb2AWti1MLJXFV0V_9trXnk7phdVZVOGTc_i6lBGDSKo3VulRkIoZvB4SOZRiwwMRMXjWRyc4ZVcb85Xw$). The riboflavin GWAS dataset from Section 4 is publicly available. The maize GWAS dataset from Section 4 is not available publicly, but we have included the simulation scripts in the shared code repository for reference.

**ORCID**

Niloy Biswas [https://orcid.org/0000-0001-9081-5702](https://orcid.org/0000-0001-9081-5702)
REFERENCES

Albert, J.H. & Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422), 669–679. ISSN 01621459. Available from: http://www.jstor.org/stable/2290350

Atchadé, Y.F. (2021) Approximate spectral gaps for Markov chain mixing times in high dimensions. *SIAM Journal on Mathematics of Data Science*, 3(3), 854–872. Available from: https://doi.org/10.1137/19M1283082

Beskos, A. & Roberts, G.O. (2005) One-shot CFTP; application to a class of truncated Gaussian densities. *Methodology and Computing in Applied Probability*, 7(4), 407.

Bhadra, A., Datta, J., Polson, N.G. & Willard, B. (2019) Lasso meets Horseshoe: a survey. *Statistical Science*, 34(3), 405–427.

Bhattacharya, A. & Johndrow, J.E. (2021) Sampling local-scale parameters in high-dimensional regression models. *Wiley StatsRef: Statistics Reference Online*, 1, 1–14. Available from: https://doi.org/10.1002/9781118445112.stat08292

Bhattacharya, A., Pati, D., Pillai, N.S. & Dunson, D.B. (2015) Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512), 1479–1490. Available from: https://doi.org/10.1080/01621459.2014.960967

Bhattacharya, A., Chakraborty, A. & Mallick, B.K. (2016) Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, 103(4), 985–991. ISSN 0006-3444. Available from: https://doi.org/10.1093/biomet/asw042

Bhattacharya, S., Khare, K. & Pal, S. (2022) Geometric ergodicity of Gibbs samplers for the Horseshoe and its regularized variants. *Electronic Journal of Statistics*, 16(1), 1–57. Available from: https://doi.org/10.1214/21-EJS1932

Biswas, N., Jacob, P.E. & Vanetti, P. (2019) Estimating convergence of Markov chains with L-lag couplings. In: Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E. & Garnett, R. (Eds.) *Advances in neural information processing systems*, Red Hook, NY: Curran Associates, Inc. pp. 7389–7399. https://proceedings.neurips.cc/paper/2019/hash/aec851e56546f6835e915293381e20a-Abstract.html

Bou-Rabee, N. & Eberle, A. (2020) Two-scale coupling for preconditioned Hamiltonian Monte Carlo in infinite dimensions. *Stochastics and Partial Differential Equations: Analysis and Computations*. Available from: https://doi.org/10.1007/s40072-020-00175-6

Bühlmann, P., Kalisch, M. & Meier, L. (2014) High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and its Application*, 1(1), 255–278. Available from: https://doi.org/10.1146/annurev-statistics-022513-115545

Carpenter, B., Gelman, A., Hoffmann, M.D., Lee, D., Goodrich, B., Betancourt, M. et al. (2017) Stan: a probabilistic programming language. *Grantee Submission*, 76(1), 1–32.

Carvalho, C.M., Polson, N.G. & Scott, J.G. (2009) Handling sparsity via the Horseshoe. In: *Proceedings of machine learning research*, vol. 5. pp. 73–80. PMLR. Available from: http://proceedings.mlr.press/v5/carvalho09a.html.

Carvalho, C.M., Polson, N.G. & Scott, J.G. (2010) The horseshoe estimator for sparse signals. *Biometrika*, 97(2), 465–480. ISSN 00063444. Available from: http://www.jstor.org/stable/25734098

Diaconis, P., Khare, K. & Saloff-Coste, L. (2008) Gibbs sampling, exponential families and orthogonal polynomials. *Statistical Science*, 23(2), 151–178. Available from: https://doi.org/10.1214/07-STS252

Diaconis, P., Khare, K. & Saloff-Coste, L. (2010) Gibbs sampling, conjugate priors and coupling. *Sankhyā: The Indian Journal of Statistics, Series A* (2008-), 72(1), 136–169. Available from: http://www.jstor.org/stable/41941459.

Friedman, J.H., Hastie, T. & Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, Articles*, 33(1), 1–22. ISSN 1548-7660. Available from: https://www.jstatsoft.org/v033/i01

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C.A. & Smoller, J.W. (2019) Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, 10(1), 1776. Available from: https://doi.org/10.1038/s41467-019-09718-5

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3), 515–534. Available from: https://doi.org/10.1214/06-BA117A
Ghosh, P. & Chakrabarti, A. (2017) Asymptotic optimality of one-group shrinkage priors in sparse high-dimensional problems. *Bayesian Analysis*, 12(4), 1133–1161. Available from: https://doi.org/10.1214/16-BA1029

Glynn, P.W. & Rhee, C.-H. (2014) Exact estimation for Markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A), 377–389.

Gorham, J. & Mackey, L. (2015) Measuring sample quality with Stein’s method. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M. & Garnett, R. (Eds.) *Advances in neural information processing systems*, vol. 28. Red Hook, NY: Curran Associates, Inc. Available from: https://proceedings.neurips.cc/paper/2015/file/698d51a19d8a121ce581499d7b701668-Paper.pdf

Gorham, J. & Mackey, L. (2017) Measuring sample quality with kernels. In: Precup, D. & Teh, Y.W. (Eds.) *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*, pp. 1292–1301, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. Available from: http://proceedings.mlr.press/v70/gorham17a.html

Gorham, J., Duncan, A.B., Vollmer, S.J. & Mackey, L. (2019) Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5), 2884–2928. Available from: https://doi.org/10.1214/19-AAP1467

Griffin, J.E. & Brown, P.J. (2010) Inference with Normal-Gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1), 171–188. Available from: https://doi.org/10.1214/10-BA507

Guan, Y. & Stephens, M. (2011) Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, 5(3), 1780–1815. Available from: https://doi.org/10.1214/10-AOAS455

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

Jacob, P.E., O’Leary, J. & Atchadé, Y.F. (2020) Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3), 543–600. Available from: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12336

Johndrow, J., Orenstein, P. & Bhattacharya, A. (2020) Scalable approximate MCMC algorithms for the Horseshoe prior. *Journal of Machine Learning Research*, 21(73), 1–61. Available from: http://jmlr.org/papers/v21/19-536.html

Johnson, V.E. (1996) Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths. *Journal of the American Statistical Association*, 91(433), 154–166.

Johnson, V.E. (1998) A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms. *Journal of the American Statistical Association*, 93(441), 238–248.

Johnson, V.E. & Rossell, D. (2012) Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498), 649–660. Available from: https://doi.org/10.1080/01621459.2012.682536

Khare, K. & Hobert, J.P. (2013) Geometric ergodicity of the Bayesian lasso. *Electronic Journal of Statistics*, 7, 2150–2163. Available from: https://doi.org/10.1214/13-EJS841

Lee, A., Singh, S.S. & Vihola, M. (2020a) Comment on ‘Unbiased Markov chain Monte Carlo with couplings’ by Jacob, O’Leary and Atchadé. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3), 543–600.

Lee, A., Singh, S.S. & Vihola, M. (2020b) Coupled conditional backward sampling particle filter. *Annals of Statistics*, 48(5), 3066–3089.

Liu, X., Huang, M., Fan, B., Buckler, E.S. & Zhang, Z. (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLOS Genetics*, 12(2), 1–24. Available from: https://doi.org/10.1371/journal.pgen.1005767

Lunn, D.J., Thomas, A., Best, N. & Spiegelhalter, D. (2000) WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10(4), 325–337.

Makalic, E. & Schmidt, D.F. (2016) A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1), 179–182.

Meyn, S.P. & Tweedie, R.L. (1993) *Markov chains and stochastic stability*. New York, NY: Springer.
Nishimura, A. & Suchard, M.A. (2018) Prior-preconditioned conjugate gradient for accelerated Gibbs sampling in 'large n & large p' sparse Bayesian logistic regression models. *arXiv preprint arXiv:1810.12437*.

Pal, S. & Khare, K. (2014) Geometric ergodicity for Bayesian shrinkage models. *Electronic Journal of Statistics*, 8(1), 604–645. Available from: https://doi.org/10.1214/14-EJS896

Park, T. & Casella, G. (2008) The Bayesian lasso. *Journal of the American Statistical Association*, 103(482), 681–686. Available from: https://doi.org/10.1198/016214508000000337

van der Pas, S.L., Kleijn, B.J.K. & van der Vaart, A.W. (2014) The horseshoe estimator: posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8(2), 2585–2618. Available from: https://doi.org/10.1214/14-EJS962

van der Pas, S., Szabó, B. & and van der Vaart, A. (2017) Adaptive posterior contraction rates for the Horseshoe. *Electronic Journal of Statistics*, 11(2), 3196–3225. Available from: https://doi.org/10.1214/17-EJS1316

Paulin, D. (2020) Comment on ‘Unbiased Markov chain Monte Carlo with couplings’ by Jacob, O’leary and Atchadé. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3), 543–600.

Plummer, M. (2003) JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Available from: https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf

Polson, N.G. & Scott, J.G. (2012) On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4), 887–902.

Polson, N.G., Scott, J.G. & Windle, J. (2014) The Bayesian bridge. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4), 713–733. Available from: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12042

Qin, Q. & Hobert, J.P. (2019a) Convergence complexity analysis of Albert and Chib’s algorithm for Bayesian probit regression. *The Annals of Statistics*, 47(4), 2320–2347. Available from: https://doi.org/10.1214/18-AOS1749

Qin, Q. & Hobert, J.P. (2019b) Wasserstein-based methods for convergence complexity analysis of MCMC with applications. *arXiv preprint arXiv:1810.08826*.

R Core Team (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org/

Roberts, G.O. & Rosenthal, J.S. (2002) One-shot coupling for certain stochastic recursive sequences. *Stochastic Processes and their Applications*, 99(2), 195–208. ISSN 0304-4149. Available from: https://doi.org/10.1016/S0304-4149(02)00096-0

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

Romay, M.C., Millard, M.J., Glaubitz, J.C., Peiffer, J.A., Swarts, K.L., Casstevens, K.L. et al. (2013) Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biology*, 14(6), R55. Available from: https://doi.org/10.1186/gb-2013-14-6-r55

Rosenthal, J.S. (1995) Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90(430), 558–566. Available from: http://www.jstor.org/stable/2291067

Ruiz, F.J.R., Titsias, M.K., Cemgil, T. & Doucet, A. (2021) Unbiased gradient estimation for variational auto-encoders using coupled Markov chains. In: de Campos, C. & Maathuis, M.H. (Eds.) *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, vol. 161 of *Proceedings of Machine Learning Research*, pp. 707–717. PMLR, 27–30. Available from: https://proceedings.mlr.press/v161/ruiz21a.html

Song, Q. (2020) Bayesian shrinkage towards sharp minimaxity. *Electronic Journal of Statistics*, 14(2), 2714–2741. Available from: https://doi.org/10.1214/20-EJS1732

Thorisson, H. (2000) *Coupling, stationarity, and regeneration*. New York, NY: Springer.

Tibshirani, R. (1994) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267–288.
Trippe, B., Nguyen, T.D. & Broderick, T. (2021) Optimal transport couplings of Gibbs samplers on partitions for unbiased estimation. *3rd Symposium on Advances in Approximate Bayesian Inference.*

Vats, D. & Knudson, C. (2021) Revisiting the Gelman–Rubin diagnostic. *Statistical Science,* 36(4), 518–529. Available from: https://doi.org/10.1214/20-STS812

Vats, D., Flegal, J.M. & Jones, G.L. (2019) Multivariate output analysis for Markov chain Monte Carlo. *Biometrika,* 106(2), 321–337. ISSN 0006-3444. Available from: https://doi.org/10.1093/biomet/asz002

Xu, K., Fjelde, T.E., Sutton, C. & Ge, H. (2021) Couplings for multinomial Hamiltonian Monte Carlo. In: *International conference on artificial intelligence and statistics.* pp. 3646–3654. PMLR.

Yang, Y., Wainwright, M.J. & Jordan, M.I. (2016) On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics,* 44(6), 2497–2532. Available from: https://doi.org/10.1214/15-AOS1417

Zeng, P. & Zhou, X. (2017) Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nature Communications,* 8(1), 456. Available from: https://doi.org/10.1038/s41467-017-00470-2

Zhou, X., Carbonetto, P. & Stephens, M. (2013) Polygenic modeling with Bayesian sparse linear mixed models. *PLOS Genetics,* 9(2), 1–14. Available from: https://doi.org/10.1371/journal.pgen.1003264

Zou, H. & Hastie, T. (2005) Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B,* 67, 301–320.

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of the article at the publisher’s website.

**How to cite this article:** Biswas, N., Bhattacharya, A., Jacob, P.E. & Johndrow, J.E. (2022) Coupling-based convergence assessment of some Gibbs samplers for high-dimensional Bayesian regression with shrinkage priors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 84(3), 973–996. Available from: https://doi.org/10.1111/rssb.12495