The Annals of Statistics
2016, Vol. 44, No. 6, 2497–2532
DOI: 10.1214/15-AOS1417

ON THE COMPUTATIONAL COMPLEXITY OF HIGH-DIMENSIONAL BAYESIAN VARIABLE SELECTION

BY YUN YANG[1,2], MARTIN J. WAINWRIGHT[1,2] AND MICHAEL I. JORDAN[1]

University of California, Berkeley

We study the computational complexity of Markov chain Monte Carlo (MCMC) methods for high-dimensional Bayesian linear regression under sparsity constraints. We first show that a Bayesian approach can achieve variable-selection consistency under relatively mild conditions on the design matrix. We then demonstrate that the statistical criterion of posterior concentration need not imply the computational desideratum of rapid mixing of the MCMC algorithm. By introducing a truncated sparsity prior for variable selection, we provide a set of conditions that guarantee both variable-selection consistency and rapid mixing of a particular Metropolis–Hastings algorithm. The mixing time is linear in the number of covariates up to a logarithmic factor. Our proof controls the spectral gap of the Markov chain by constructing a canonical path ensemble that is inspired by the steps taken by greedy algorithms for variable selection.

1. Introduction. In many areas of science and engineering, it is common to collect a very large number of covariates $X_1, \dots, X_p$ in order to predict a response variable Y. We are thus led to instances of high-dimensional regression, in which the number of covariates p exceed the sample size n. A large literature has emerged to address problems in the regime $p \gg n$, where the ill-posed nature of the problem is addressed by imposing sparsity conditions—namely, that the response Y depends only on a small subset of the covariates. Much of this literature is based on optimization methods, where penalty terms are incorporated that yield both convex [32] and nonconvex [8, 38] optimization problems. Theoretical analysis is based on general properties of the design matrix and the penalty function.

Alternatively, one can take a Bayesian point of view on high-dimensional regression, placing a prior on the model space and performing the necessary integration so as to obtain a posterior distribution [4, 11, 15]. Obtaining such a posterior allows one to report a subset of possible models along with their posterior probabilities as opposed to a single model. One can also report the marginal posterior

Received May 2015; revised September 2015.

[1] Supported in part by Office of Naval Research MURI Grant N00014-11-1-0688.

[2] Supported in part by National Science Foundation Grants CIF-31712-23800 and DMS-11-07000.

MSC2010 subject classifications. Primary 62F15; secondary 60J10.

Key words and phrases. High-dimensional inference, Bayesian variable selection, Markov chain, spectral gap, rapid mixing.

probability of including each covariate. Recent work has provided some theoretical understanding of the performance of Bayesian approaches to variable selection. In the moderate-dimension scenario (in which p is allowed to grow with n but $p \leq n$), Shang and Clayton [27] establish posterior consistency for variable selection in a Bayesian linear model, meaning that the posterior probability of the true model that contains all influential covariates tends to one as n grows. Narisetty and He [25] consider a high-dimensional scenario in which p can grow nearly exponentially with n; in this setting, they show the Bayesian spike-and-slab variable-selection method achieves variable-selection consistency. Since this particular Bayesian method resembles a randomized version of ℓ_0-penalized methods, it could have better performance than ℓ_1-penalized methods for variable selection under high-dimensional settings [25, 28]. Empirical evidence for this conjecture is provided by Guan et al. [12] for SNP selection in genome-wide association studies, but it has not been confirmed theoretically.

The most widely used tool for fitting Bayesian models are sampling techniques based on Markov chain Monte Carlo (MCMC), in which a Markov chain is designed over the parameter space so that its stationary distribution matches the posterior distribution. Despite its popularity, the theoretical analysis of the computational efficiency of MCMC algorithms lags that of optimization-based methods. The central object of interest in such analyses is the *mixing time* of the Markov chain, which characterizes the number of iterations required to converge to an ε-distance of the stationary distribution from any initial configuration. In order for MCMC algorithms to be controlled approximations, one must provide meaningful bounds on the mixing time as a function of problem parameters such as the number of observations and the dimensionality. Of particular interest is determining whether the chain is *rapidly mixing*, meaning that the mixing time grows at most polynomially in the problem parameters, or *slowly mixing*, meaning that the mixing time grows exponentially in the problem parameters. In the latter case, one cannot hope to obtain approximate samples from the posterior in any reasonable amount of time for large models.

Unfortunately, theoretical analysis of mixing time is comparatively rare in the Bayesian literature and is dominated by negative results. On the positive side, Jones and Hobert [16] consider a Bayesian hierarchical version of the one-way random effects model, and obtain upper bounds on the mixing time of Gibbs and block Gibbs samplers as a function of the initial values, data and hyperparameters. Belloni and Chernozhukov [3] show that a Metropolis random walk is rapidly mixing in the dimension for regular parametric models in which the posterior converges to a normal limit. Schreck et al. [26] introduce a new MCMC method for Bayesian variable selection in high dimensions, and show that the Markov chain is geometrically ergodic. However, they do not provide an explicit characterization of the geometric rate at which the Markov chain converges, and the rate may be arbitrary close to one. It is more common to find negative results in the literature. Examples include Mossel and Vigoda [24], who show that the MCMC algorithm for Bayesian phylogenetics takes exponentially long to reach the stationary distribution as data accumulates and Woodard and Rosenthal [35], who analyze a Gibbs sampler used for genomic motif discovery and show that the mixing time increases exponentially as a function of the length of the DNA sequence.

The goal of the current paper is to study the computational complexity of Metropolis–Hastings procedures for high-dimensional Bayesian variable selection. For concreteness, we focus our analysis on a specific hierarchical Bayesian model for sparse linear regression, and an associated Metropolis–Hastings random walk, but these choices should be viewed as representative of a broader family of methods. In particular, we study the well-known Zellner g-prior for linear regression [37]. The main advantage of this prior is the simple expression that it yields for the marginal likelihood, which is convenient in our theoretical investigations. As in past analyses [25], we consider the marginal probability of including each covariate into the model as being on the order of $p^{-\mathcal{O}(1)}$. Moreover, we restrict the support of the prior to rule out unrealistically large models. As a specific computational methodology, we focus on an iterative, local-move and neighborhood-based procedure for sampling from the model space, which is motivated by shotgun stochastic search [13].

It has been suggested by some statisticians that the mixing time for high-dimensional Bayesian variable selection should be exponential, because the Markov chain must eventually visit all possible models. Interesting, our work shows that this plausible argument is misleading. First, although the state space contains exponentially many models, each of them can be reached from one another in at most p transition steps. Second, different models may share common covariates and are not independently structured. Therefore, it is possible for the Markov chain to quickly identify a highest posterior region without visiting all the models. Our main contribution in this paper is to provide conditions under which Bayesian posterior consistency holds, and moreover, when the mixing time grows linearly (up to a logarithmic factor) in the dimension p, implying that the chain is rapidly mixing. As a by-product, we provide conditions on the hyperparameter g to achieve model-selection consistency. We also provide a counterexample to illustrate that although ruling out unrealistically large models is not necessary for achieving variable-selection consistency, it is necessary in order that the Metropolis–Hastings random walk is rapidly mixing. To be clear, while our analysis applies to a fully Bayesian procedure for variable selection, it is based on a frequentist point of view in assuming that the data are generated according to a true model.

There are a number of challenges associated with characterizing the computational complexity of Markov chain methods for Bayesian models. First, the posterior distribution of a Bayesian model is usually a much more complex object than the highly structured distributions of statistical physics for which meaningful bounds on the Markov chain mixing times are often obtained (e.g., [5, 20, 22]).

Second, the transition probabilities of the Markov chain are themselves stochastic, since they depend on the underlying data-generating process. In order to address these challenges, our analysis exploits asymptotic properties of the Bayesian model to characterize the typical behavior of the Markov chain. We show that under conditions leading to Bayesian variable-selection consistency, the Markov chain over the model space has a global tendency of moving toward the true data-generating model, even though the posterior distribution can be highly irregular. In order to bound the mixing time, we make use of the canonical path technique developed by Sinclair [29, 30] and Diaconis and Stroock [7]. More precisely, the particular canonical path construction used in our proof is motivated by examining the solution path of stepwise regression procedures for linear model selection (e.g., [1, 39]), where a greedy criterion is used to decide at each step whether a covariate is to be included or deleted from the curent model.

Overall, our results reveal that there is a delicate interplay between the statistical and computational properties of Bayesian models for variable selection. On the one hand, we show that concentration of the posterior is not only useful in guaranteeing desirable statistical properties such as model-selection consistency, but they also have algorithmic benefits in certifying the rapid mixing of the Markov chain methods designed to draw samples from the posterior. On the other hand, we show that posterior consistency on its own is *not* sufficient for rapid mixing, so that algorithmic efficiency requires somewhat stronger conditions.

The remainder of this paper is organized as follows. Section 2 provides background on the Bayesian approach to variable selection, as well as Markov chain algorithms for sampling and techniques for analysis of mixing times. In Section 3, we state our two main results (Theorems 1 and 2) for a class of Bayesian models for variable selection, along with simulations that illustrate the predictions of our theory. Section 4 is devoted to the proofs of our results, with many of the technical details deferred to the appendices in the Supplement ([36]). We conclude in Section 5 with a discussion.

2. Background and problem formulation.

In this section, we introduce some background on the Bayesian approach to variable selection, as well some background on Markov chain algorithms for sampling and techniques for analyzing their mixing times.

2.1. *Variable selection in the Bayesian setting.*

Consider a response vector $Y \in \mathbb{R}^n$ and a design matrix $X \in \mathbb{R}^{n \times p}$ that are linked by the standard linear model

$$Y = X\beta^* + w \qquad \text{where } w \sim \mathcal{N}(0, \sigma_0^2 I_n), \tag{1}$$

and $\beta^* \in \mathbb{R}^p$ is the unknown regression vector. Based on observing the pair (Y, X), our goal is to recover the support set of β^*—that is, to select the subset of covariates with nonzero regression weights, or more generally, a subset of covariates with absolute regression weights above some threshold.

In generic terms, a Bayesian approach to variable selection is based on first imposing a prior over the set of binary indicator vectors, and then using the induced posterior [denoted by $\pi(\gamma|Y)$] to perform variable selection. Here, each binary vector $\gamma \in \{0,1\}^p$ should be thought of as indexing the model which involves only the covariates indexed by γ. We make use of the shorthand $|\gamma| = \sum_{j=1}^p \gamma_j$ corresponding to the number of nonzero entries in γ, or the number of active covariates in the associated model. It will be convenient to adopt a dualistic view of γ as both a binary indicator vector, and as a subset of $\{1,\dots,p\}$. Under this identification, the expression $\gamma \subset \gamma'$ for a pair of inclusion vectors (γ,γ') can be understood as that the subset of variables selected by γ is contained in the subset of variables selected by γ'. Similarly, it will be legitimate to use set operators on those indicator vectors, such as $\gamma \cap \gamma'$, $\gamma \cup \gamma'$ and $\gamma \setminus \gamma'$. Using this interpretation, we let $X_\gamma \in \mathbb{R}^{n\times|\gamma|}$ denote the submatrix formed of the columns indexed by γ, and we define the subvector $\beta_\gamma \in \mathbb{R}^{|\gamma|}$ in an analogous manner. We make use of this notation in defining the specific hierarchical Bayesian model analyzed in this paper, defined precisely in Section 3.1 to follow.

2.2. *MCMC algorithms for Bayesian variable selection.* Past work on MCMC algorithms for Bayesian variable selection can be divided into two main classes—Gibbs samplers (e.g., [11, 15, 25]) and Metropolis–Hastings random walks (e.g., [12, 13]). In this paper, we focus on a particular form of Metropolis–Hastings updates.

In general terms, a Metropolis–Hastings random walk is an iterative and local-move-based procedure involving three steps:

Step 1. Use the current state γ to define a neighborhood $\mathcal{N}(\gamma)$ of proposal states.

Step 2. Choose a proposal state γ' in $\mathcal{N}(\gamma)$ according to some probability distribution $\mathbf{S}(\gamma,\cdot)$ over the neighborhood; for example, the uniform distribution.

Step 3. Move to the new state γ' with probability $\mathbf{R}(\gamma,\gamma')$, and stay in the original state γ with probability $1-\mathbf{R}(\gamma,\gamma')$, where the acceptance ratio is given by

$$\mathbf{R}(\gamma,\gamma') := \min\left\{1, \frac{\pi_n(\gamma'|Y)\mathbf{S}(\gamma',\gamma)}{\pi_n(\gamma|Y)\mathbf{S}(\gamma,\gamma')}\right\}. \tag{2}$$

In this way, for any fixed choice of the neighborhood structure $\mathcal{N}(\gamma)$, we obtain a Markov chain with transition probability given by

$$\mathbf{P}_{\mathrm{MH}}(\gamma,\gamma') = \begin{cases} \mathbf{S}(\gamma,\gamma')\mathbf{R}(\gamma,\gamma'), & \text{if } \gamma' \in \mathcal{N}(\gamma), \\ 0, & \text{if } \gamma' \notin \mathcal{N}(\gamma)\cup\{\gamma\} \quad \text{and} \\ 1 - \sum_{\tilde\gamma\neq\gamma} \mathbf{P}_{\mathrm{MH}}(\gamma,\tilde\gamma), & \text{if } \gamma' = \gamma. \end{cases}$$

The specific form of Metropolis–Hastings update analyzed in this paper is obtained by randomly selecting one of the following two schemes to update γ, each with probability 0.5.

Single flip update: Choose an index $j \in [p]$ uniformly at random, and form the new state γ' by setting $\gamma'_j = 1 - \gamma_j$.

Double flip update: Define the subsets $S(\gamma) = \{j \in [p] | \gamma_j = 1\}$ and let $S^c(\gamma) = \{j \in [p] | \gamma_j = 0\}$. Choose an index pair $(k, \ell) \in S(\gamma) \times S^c(\gamma)$ uniformly at random, and form the new state γ' by flipping γ_k from 1 to 0 and γ_ℓ from 0 to 1. [If the set $S(\gamma)$ is empty, then we do nothing.]

This scheme can be understood as a particular of the general Metropolis–Hastings scheme in terms of a neighborhood $\mathcal{N}(\gamma)$ to be all models γ' that can be obtained from γ by either changing one component to its opposite (i.e., from 0 to 1, or from 1 to 0) or switching the values of two components with different values.

Let $d_H(\gamma, \gamma') = \sum_{j=1}^p \mathbb{I}(\gamma_j \neq \gamma'_j)$ denote the Hamming distance between γ and γ'. With this notation, the overall neighborhood is given by the union $\mathcal{N}(\gamma) := \mathcal{N}_1(\gamma) \cup \mathcal{N}_2(\gamma)$, where

$$\begin{aligned}\mathcal{N}_1(\gamma) &:= \{\gamma' | d_H(\gamma', \gamma) = 1\}, \quad \text{and} \\ \mathcal{N}_2(\gamma) &:= \{\gamma | d_H(\gamma', \gamma) = 2, \text{ and} \\ &\qquad \exists (k, \ell) \in S(\gamma) \times S^c(\gamma) \text{ s.t. } \gamma'_k = 1 - \gamma_k \text{ and } \gamma'_\ell = 1 - \gamma_\ell\}.\end{aligned}$$

The transition matrix of the previously described Metropolis–Hastings scheme takes the form

$$\begin{aligned}&\mathbf{P}_{\mathrm{MH}}(\gamma, \gamma') \\ &= \begin{cases} \dfrac{1}{2p} \min\left\{1, \dfrac{\pi_n(\gamma'|Y)}{\pi_n(\gamma|Y)}\right\}, & \text{if } \gamma' \in \mathcal{N}_1(\gamma), \\ \dfrac{1}{2|S(\gamma)||S^c(\gamma)|} \min\left\{1, \dfrac{\pi_n(\gamma'|Y)}{\pi_n(\gamma|Y)}\right\}, & \text{if } \gamma' \in \mathcal{N}_2(\gamma), \\ 0, & \text{if } d_H(\gamma', \gamma) > 2 \quad \text{and} \\ 1 - \sum_{\tilde{\gamma} \neq \gamma} \mathbf{P}_{\mathrm{MH}}(\gamma, \tilde{\gamma}), & \text{if } \gamma' = \gamma. \end{cases}\end{aligned} \tag{3}$$

2.3. *Background on mixing times.* Let $\mathcal{C}$ be an irreducible, aperiodic Markov chain on the discrete state space $\mathscr{M}$, and described by the transition probability matrix $\mathbf{P} \in \mathbb{R}^{|\mathscr{M}| \times |\mathscr{M}|}$ with stationary distribution π. We assume throughout that $\mathcal{C}$ is reversible; that is, it satisfies the detailed balance condition $\pi(\gamma)\mathbf{P}(\gamma, \gamma') = \pi(\gamma')\mathbf{P}(\gamma', \gamma)$ for all $\gamma, \gamma' \in \mathscr{M}$. It is easy to see that the previously described Metropolis–Hastings matrix $\mathbf{P}_{\mathrm{MH}}$ satisfies this reversibility condition. It is convenient to identify a reversible chain with a weighted, undirected graph G on the vertex set $\mathscr{M}$, where two vertices γ and γ' are connected if and only if the edge weight $\mathbf{Q}(\gamma,\gamma') := \pi(\gamma)\mathbf{P}(\gamma,\gamma')$ is strictly positive.

For $\gamma \in \mathscr{M}$ and any subset $S \subseteq \mathscr{M}$, we write $\mathbf{P}(\gamma, S) = \sum_{\gamma' \in S} \mathbf{P}(\gamma,\gamma')$. If γ is the initial state of the chain, then the total variation distance to the stationary distribution after t iterations is

$$\Delta_\gamma(t) = \left\|\mathbf{P}^t(\gamma,\cdot) - \pi(\cdot)\right\|_{\mathrm{TV}} := \max_{S \subset \mathscr{M}} \left|\mathbf{P}^t(\gamma, S) - \pi(S)\right|.$$

The ε-mixing time is given by

$$\tau_\varepsilon := \max_{\gamma \in \mathscr{M}} \min\left\{t \in \mathbb{N} \,|\, \Delta_\gamma(t') \le \varepsilon \text{ for all } t' \ge t\right\}, \tag{4}$$

which measures the number of iterations required for the chain to be within distance $\varepsilon \in (0, 1)$ of stationarity. The efficiency of the Markov chain can be measured by the dependence of τ_ε on the difficulty of the problem, for example, the dimension of the parameter space and the sample size. In our case, we are interested in the dependence of τ_ε on the covariate dimension p and the sample size n. Of particular interest is whether the chain is *rapidly mixing*, meaning that the mixing time grows at most polynomially in the pair (p, n), or *slowly mixing*, meaning that the mixing time grows exponentially.

3. Main results and their consequences.

The analysis of this paper applies to a particular family of hierarchical Bayes models for variable selection. Accordingly, we begin by giving a precise description of this family of models, before turning to statements of our main results and a discussion of their consequences. Our first result (Theorem 1) provides sufficient conditions for posterior concentration, whereas our second result (Theorem 2) provides sufficient conditions for rapid mixing of the Metropolis–Hastings updates.

3.1. *Bayesian hierarchical model for variable selection.*

In addition to the standard linear model (1), the Bayesian hierarchical model analyzed in this paper involves three other ingredients: a prior over the precision parameter ϕ (or inverse noise variance) in the linear observation model, a prior on the regression coefficients and a prior over the binary indicator vectors. More precisely, it is given by

$$\mathbb{M}_\gamma: \quad \text{Linear model:} \quad Y = X_\gamma\beta_\gamma + w, \qquad w \sim \mathcal{N}(0, \phi^{-1} I_n), \tag{5a}$$

$$\text{Precision prior:} \quad \pi(\phi) \propto \frac{1}{\phi}, \tag{5b}$$

$$\text{Regression prior:} \quad (\beta_\gamma|\gamma) \sim \mathcal{N}(0, g\phi^{-1}(X_\gamma^T X_\gamma)^{-1}), \tag{5c}$$

$$\text{Sparsity prior:} \quad \pi(\gamma) \propto \left(\frac{1}{p}\right)^{\kappa|\gamma|} \mathbb{I}[|\gamma| \le s_0]. \tag{5d}$$

For each model $\mathbb{M}_\gamma$, there are three parameters to be specified: the integer $s_0 < n$ is a prespecified upper bound on the maximum number of important covariates, the hyperparameter $g > 0$ controls the degree of dispersion in the regression prior, and the hyperparameter $\kappa > 0$ penalizes models with large size. For a given integer $s_0 \in \{1, \dots, p\}$, we let $\mathscr{M}(s_0) = \{\mathbb{M}_\gamma \| \gamma| \le s_0\}$ the class of all models involving at most s_0 covariates.

Let us make a few remarks on our choice of Bayesian model. First, the choice of covariance matrix in the regression prior—namely, involving $X_\gamma^T X_\gamma$—is made for analytical convenience, in particular to simplify the posterior. A more realistic choice would be the independence prior

$$\beta_\gamma | \gamma \sim \mathcal{N}(0, g\phi^{-1} I_{|\gamma|}).$$

However, the difference between the impacts on the posterior of these choices will be negligible when $g \gg n$, which, as shown by our theoretical analysis, is the regime under which the posterior is well-behaved. Another popular choice for the prior of β_γ is the spike-and-slab prior [15], where for each covariate X_j, one specifies the marginal prior for β_j as a mixture of two normal distributions, one with a substantially larger variance than the other, and γ_j can be viewed as the latent class indicator for this mixture prior. Our primary motivation in imposing Zellner's g-prior is in order to streamline the theoretical analysis: it leads to an especially simple form of the marginal likelihood function. However, we note that our conclusions remain valid under essentially the same conditions when the independence prior or the spike-and-slab prior is used, but with much longer proofs. The sparsity prior on γ is similar to the prior considered by Narisetty and He [25] and Castillo et al. [6]. The $p^{-\kappa}$ decay rate for the marginal probability of including each covariate imposes a vanishing prior probability on the models of diverging sizes. The only difference is that we impose a constraint $|\gamma| \le s_0$ to rule out models with too many covariates. As will be clarified in the sequel, while this additional constraint is not needed for Bayesian variable-selection consistency, it is necessary for rapid mixing of the MCMC algorithm that we analyze.

Recall from our earlier set-up that the response vector $Y \in \mathbb{R}^n$ is generated from the standard linear model $Y = X\beta^* + w$, where $w \sim \mathcal{N}(0, \sigma_0^2 I_n)$, $\beta^* \in \mathbb{R}^p$ is the unknown regression vector, and σ_0 the unknown noise standard deviation. In rough terms, the goal of variable selection is to determine the subset S of "influential" covariates. In order to formalize this notion, let us fix a constant $C_\beta > 0$ depending on (σ_0, n, p) that quantifies the minimal signal size requirement for a covariate to be "influential." We then define $S = S(C_\beta)$ to consist of the indices with relatively large signal—namely

$$S := \{j \in [p] \| \beta_j^*| \ge C_\beta\}, \tag{6}$$

and our goal is to recover this subset. Thus, the "noninfluential" coefficients $\beta_{S^c}^*$ are allowed to be nonzero, but their magnitudes are constrained.

We let γ^* be the indicator vector that selects the influential covariates, and let $s^* := |\gamma^*|$ be the size of the corresponding "true" model $\mathbb{M}_{\gamma^*}$. Without loss of generality, we may assume that the first s^* components of γ^* are ones, and the rest are zeros. We assume throughout this section that we are in the high-dimensional regime where $p \geq n$, since the low-dimensional regime where $n < p$ is easier to analyze. For any symmetric matrix $\mathbf{Q}$, let $\lambda_{\min}(\mathbf{Q})$ and $\lambda_{\max}(\mathbf{Q})$ denote its smallest and largest eigenvalues. Our analysis involves the following assumptions:

ASSUMPTION A (Conditions on β^*). The true regression vector has components $\beta^* = (\beta^*_S, \beta^*_{S^c})$ that satisfy the bounds

$$\text{(7a)} \qquad \text{Full } \beta^* \text{ condition:} \qquad \left\| \frac{1}{\sqrt{n}} X\beta^* \right\|_2^2 \leq g\sigma_0^2 \frac{\log p}{n},$$

$$\text{(7b)} \qquad \text{Off-support } S^c \text{ condition:} \qquad \left\| \frac{1}{\sqrt{n}} X_{S^c}\beta^*_{S^c} \right\|_2^2 \leq \tilde{L}\sigma_0^2 \frac{\log p}{n},$$

for some $\tilde{L} \geq 0$.

In the simplest case, the true regression vector β^* is S-sparse (meaning that $\beta^*_{S^c} = 0$), so that the off-support condition holds trivially. As for the full β^* condition, it is known [27] that some form of upper bound on the norm $\|\beta^*\|_2$ in terms of the g-hyperparameter is required in order to prove Bayesian model selection consistency. The necessity of such a condition is a manifestation of the so-called information paradox of g-priors [21].

Our next assumption involves an integer parameter s, which is set either to a multiple of the true sparsity s^* (in order to prove posterior concentration) or the truncated sparsity s_0 (in order to prove rapid mixing).

ASSUMPTION B (Conditions on the design matrix). The design matrix has been normalized so that $\|X_j\|_2^2 = n$ for all $j = 1, \dots, p$; moreover, letting $Z \sim N(0, I_n)$, there exist constants $\nu \in (0, 1]$ and $L < \infty$ such that $L\nu \geq 4$ and

(7c)

Lower restricted eigenvalue $(\mathrm{RE}(s))$:

$$\min_{|\gamma| \leq s} \lambda_{\min}\left(\frac{1}{n} X_\gamma^T X_\gamma\right) \geq \nu \quad \text{and}$$

Sparse projection condition $(\mathrm{SI}(s))$:

$$\mathbb{E}_Z\left[\max_{|\gamma| \leq s} \max_{k \in [p] \backslash \gamma} \frac{1}{\sqrt{n}} |\langle (I - \Phi_\gamma) X_k, Z\rangle|\right] \leq \frac{1}{2}\sqrt{L\nu \log p},$$

where Φ_γ denotes projection onto the span of $\{X_j, j \in \gamma\}$.

The lower restricted eigenvalue condition is a mild requirement, one that plays a role in the information-theoretic analysis of variable selection [33]. On the other hand, the sparse projection condition can always be satisfied by choosing $L = \mathcal{O}(s_0)$. To see this, notice that $\frac{1}{\sqrt{n}}\|(I - \Phi_\gamma)X_k\| \leq 1$ and there are at most p^{s_0} different choice of distinct pair (γ, k). Therefore, by the Gaussianity of g_G, the sparse projection condition always holds with $L = 4\nu^{-1}s_0$. On the other extreme, if the design matrix X has orthogonal columns, then $(I - \Phi_\gamma)X_k = X_k$. As a consequence, due to the same argument, the sparse projection condition holds with $L = 4\nu^{-1}$, which depends neither on s^* nor on s_0.

ASSUMPTION C (Choices of prior hyperparameters). The noise hyperparameter g and sparsity penalty hyperparameter $\kappa \geq 2$ are chosen such that

$$g \asymp p^{2\alpha} \qquad \text{for some } \alpha \geq 1/2 \quad \text{and} \quad \kappa + \alpha \geq 4(L + \tilde{L}) + 2. \tag{7d}$$

In the low-dimensional regime, $p = o(n)$, the g-prior with either the unit information prior $g = n$, or the choice $g = \max\{n, p^2\}$ have been recommended [9, 17, 31]. In the intermediate regime where $p = \mathcal{O}(n)$, Sparks et al. [31] show that g must grow faster than $p \log p/n$ for the Bayesian linear model without variable selection to achieve posterior consistency. These considerations motivate us to choose the hyperparameter for the high-dimensional setting as $g \asymp p^{2\alpha}$ for some $\alpha > 0$, and our theory establishes the utility of this choice.

ASSUMPTION D (Sparsity control). One of the two following conditions holds:

Version $\mathrm{D}(s^*)$: We set $s_0 := p$ in the sparsity prior (5d), and the true sparsity s^* is bounded as $\max\{1, s^*\} \leq \frac{1}{32}\{\frac{n}{\log p} - 8\tilde{L}\}$.

Version $\mathrm{D}(s_0)$: The sparsity parameter s_0 in the prior (5d) satisfies the sandwich relation

$$\max\{1, (2\nu^{-2}\omega(X) + 1)s^*\} \leq s_0 \leq \frac{1}{32}\left\{\frac{n}{\log p} - 8\tilde{L}\right\}, \tag{7e}$$

where $\omega(X) := \max_{\gamma \in \mathscr{M}} |||(X_\gamma^T X_\gamma)^{-1} X_\gamma^T X_{\gamma^* \backslash \gamma}|||_{\mathrm{op}}^2$.

Assumptions A, B, C and D are a common set of conditions assumed in the existing literature (e.g., [25, 27]) for establishing Bayesian variable-selection consistency; that is, the posterior probability of the true model $\pi_n(\gamma^*|Y) \to 1$ as $n \to \infty$.

3.2. *Sufficient conditions for posterior consistency.* Our first result characterizes the behavior of the (random) posterior $\pi_n(\cdot|Y)$. As we mentioned in Section 2.1, Bayesian variable-selection consistency does not require that the sparsity prior (5d) be truncated at some sparsity level much less than p, so that we analyze the hierarchical model with $s_0 = p$, and use the milder Assumption $\mathrm{D}(s^*)$. The reader should recall from equation (6) the threshold parameter C_β that defines the subset $S = S(C_\beta)$ of influential covariates. In the rest of the paper, we use c and c_j $(j = 0, 1, \ldots)$ to denote universal constants.

THEOREM 1 (Posterior concentration). *Suppose that Assumptions A, B with $s = 2(\kappa + \alpha + \tilde{L} + 1)\max\{1, s^*\}$, Assumption C and Assumption D(s^*) hold. If the threshold C_β satisfies*

$$C_\beta^2 \geq 128\nu^{-2}(L + \tilde{L} + \alpha + \kappa)\sigma_0^2 \frac{\log p}{n}, \tag{8}$$

then we have $\pi_n(\gamma^*|Y) \geq 1 - c_1 p^{-1}$ with probability at least $1 - c_2 p^{-c_3}$. Here, the probability is with respect to the data-generating process.

The threshold condition (8) requires the set of influential covariates to have reasonably large magnitudes; this type of signal-to-noise condition is needed for establishing variable-selection consistency of any procedure [33]. We refer to it as the $\beta_{\min}$-condition in the rest of the paper. Due to the mildness of Assumption A (conditions on β^*), the claim in the theorem holds even when the true model is not exactly sparse: Assumption A allows the residual $\beta^*_{S^c}$ to be nonzero as long as it has small magnitude.

It is worth noting that the result of Theorem 1 covers two regimes, corresponding to different levels of signal-to-noise ratio. More precisely, it is useful to isolate the following two mutually exclusive possibilities:

$$\text{High SNR:} \quad S = \{j \in [p] | \beta_j^* \neq 0\} \quad \text{and} \quad \min_{j \in S} |\beta_j^*|^2 \geq 128\nu^{-2}(\alpha + \kappa + L)\sigma_0^2 \frac{\log p}{n}, \tag{9a}$$

$$\text{Low SNR:} \quad S = \varnothing \quad \text{and} \quad \left\| \frac{1}{\sqrt{n}} X\beta^* \right\|_2^2 \leq \left(\frac{\alpha + \kappa - 2}{4} - L \right) \sigma_0^2 \frac{\log p}{n}. \tag{9b}$$

In terms of the parameter $\tilde{L}$ in Assumption A, the high SNR regime corresponds to $\tilde{L} = 0$, whereas the low SNR regime corresponds to $\tilde{L} = \frac{\alpha+\kappa-2}{4} - L$. The intuition for the low SNR setting is that the signal in every component is so weak that the "penalty" induced by hyperparameters (g, κ) completely overwhelms it. Theorem 1 guarantees that the posterior concentrates around the model $\mathbb{M}_{\gamma^*}$ under the high SNR condition, and around the null model $\mathbb{M}_{\gamma_0}$ under the low SNR condition. More precisely, we have the following.

COROLLARY 1. *Under the conditions of Theorem 1, with probability at least $1 - c_2 p^{-c_3}$:*

(a) *Under the high SNR condition (9a), we have $\pi_n(\gamma^*|Y) \geq 1 - c_1 p^{-1}$.*

(b) *Conversely, under the low SNR condition (9b), the posterior probability of the null model is lower bounded as $\pi_n(\gamma_0|Y) \geq 1 - c_1 p^{-1}$.*

Corollary 1 provides a complete characterization of the high or low SNR regimes, but it does not cover the intermediate regime in which some component β_j^* of β^* is sandwiched as

$$\left(\frac{\alpha+\kappa-2}{4}-L\right)\sigma_0^2\frac{\log p}{n} \le |\beta_j^*|^2 \le 128\nu^{-2}(\alpha+\kappa+L)\sigma_0^2\frac{\log p}{n}. \tag{10}$$

On the one hand, Theorem 1 still guarantees a form of Bayesian variable selection consistency in this regime. However, the MCMC algorithm for sampling from the posterior can exhibit slow mixing due to multimodality in the posterior. In Appendix A.2 of the Supplement, we provide a simple example that satisfies the conditions of Theorem 1, so that posterior consistency holds, but the Metropolis–Hastings updates have mixing time growing exponentially in p. Our proof shows that multimodality mostly occurs in the region of large models, for example, those completely overfitted models that contain n covariates. This fact motivates us to modify the prior by assigning zero prior probabilities to large models that usually receive exponentially small posterior probabilities. This example reveals a phenomenon that might seem counter-intuitive at first sight: sharp concentration of the posterior distribution need not lead to rapid mixing of the MCMC algorithm.

3.3. *Sufficient conditions for rapid mixing.* With this distinction in mind, we now turn to developing sufficient conditions for Metropolis–Hastings scheme (3) to be rapidly mixing. As discussed in Section 2, this rapid mixing ensures that the number of iterations required to converge to an ε-ball of the stationary distribution grows only polynomially in the problem parameters. The main difference in the conditions is that we now require Assumption B—the RE and sparse projection conditions—to hold with parameter $s = s_0$, as opposed to with the possibly much smaller parameter $s = 2(\kappa + \alpha + \tilde{L} + 1)\max\{1, s^*\}$ involved in Theorem 1.

THEOREM 2 (Rapid mixing guarantee). *Suppose that Assumptions A, B with $s = s_0$, Assumption C, and Assumption D(s_0) all hold. Then under either the high SNR condition (9a) or the low SNR condition (9b), for any $\varepsilon \in (0, 1)$, the ε-mixing time of the Metropolis–Hastings chain (3) is upper bounded as*

$$\tau_\varepsilon \le 12ps_0^2\big((\alpha n + \alpha s_0 + 2\kappa s_0)\log p + \log(1/\varepsilon) + \log 2\big) \tag{11}$$

with probability at least $1 - c_3 p^{-c_4}$.

According to our previous definition (4) of the mixing time, Theorem 2 characterizes the worst case mixing time, meaning the number of iterations when starting from the worst possible initialization. If we start with a good initial state, for example, the true model γ^* would be an appealing though impractical choice, then we can remove the $\log p$ term in the upper bound (11). In this way, the term $12ps_0^2(\alpha n + \alpha s_0 + 2\kappa s_0)\log p$ can be understood as the worst-case number of iterations required in the burn-in period of the MCMC algorithm. In practice, we may choose a good frequentist point estimator such as the LASSO [32] as the initial state.

Theorems 1 and 2 lead to the following corollary, stating that after $\mathcal{O}(\alpha n s_0^2 p \log p)$ iterations, the MCMC algorithm will output γ^* with high probability.

COROLLARY 2. *Under the conditions of Theorem 2, for any fixed iteration number t such that*

$$t \geq 12ps_0^2\big((\alpha n + \alpha s_0 + 2\kappa s_0 + 1)\log p + \log 2\big),$$

the iterate γ_t from the MCMC algorithm matches γ^* with probability at least $1 - c_3 p^{-c_4}$.

As with Corollary 1, Theorem 2 does not characterize the intermediate regime in which some component β_j^* of β^* satisfies the sandwich inequality (10). Based on our simulations, we suspect that the Markov chain might be slowly mixing in this regime, but we do not have a proof of this statement. The following heuristic argument provides some support for this conjecture: without the $\beta_{\min}$ condition, it is possible that all nonzero β_j^* have equally small magnitudes but as a whole exhibits a large signal. Then since the sparse penalty keeps dominating the incremental signal from adding one of the nonzero β_j^* into the model until a large portion of nonzero β_j^* have been added, there is an exponential posterior probability gap between the null model γ_0 and the true model $\gamma^* = \{j : \beta_j^* \neq 0\}$. As a consequence, unless the Markov chain is allowed to add a large portion of nonzero β_j^* at a time, it takes exponentially long to move across this gap. However, even adding a large portion at a time is allowed, it may take exponentially many steps to find a specific subset without any prior knowledge by random sampling.

3.4. *Illustrative simulations.*

In order to illustrate the predictions of Theorem 2, we conducted some simulations. We also provide an example for which a frequentist method such as the Lasso fails to perform correct variable selection while our Bayesian method succeeds.

3.4.1. *Comparison of mixing times.*

In order to study mixing times and their dependence on the model structure, we performed simulations for linear models with random design matrices, formed by choosing row $x_i \in \mathbb{R}^p$ i.i.d. from a multivariate Gaussian distribution. In detail, setting the noise variance $\sigma_0^2 = 1$, we considered two classes of linear models with random design matrices $X \in \mathbb{R}^{n\times p}$, in each case formed with i.i.d. rows $x_i \in \mathbb{R}^p$:

$$\text{Independent design:} \qquad Y \sim \mathcal{N}(X\beta^*, \sigma_0^2 I_n) \qquad \text{with } x_i \sim \mathcal{N}(0, I_p) \text{ i.i.d.;}$$

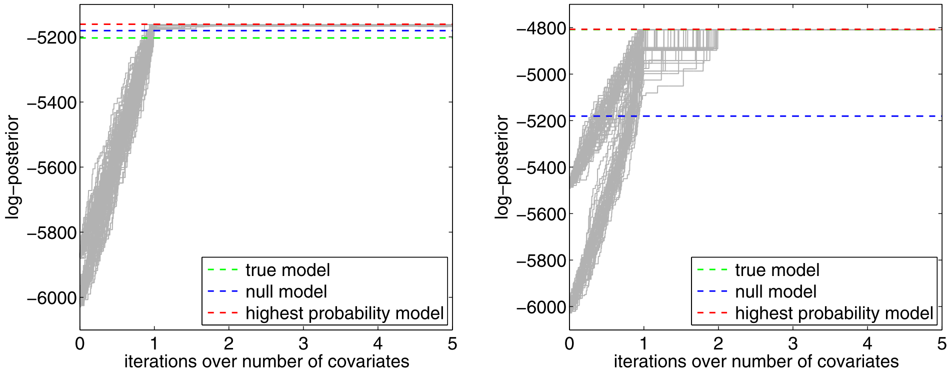
$$\text{Correlated design:} \qquad Y \sim \mathcal{N}(X\beta^*, \sigma_0^2 I_n) \qquad \text{with } x_i \sim \mathcal{N}(0, \Sigma) \text{ i.i.d.}$$
$$\text{and} \quad \Sigma_{jk} = e^{-|j-k|}.$$

In all cases, we choose a design vector $\beta^* \in \mathbb{R}^p$ with true sparsity $s^* = 10$, taking the form

$$\beta^* = \mathrm{SNR}\sqrt{\frac{\sigma_0^2 \log p}{n}}(2, -3, 2, 2, -3, 3, -2, 3, -2, 3, 0, \dots, 0)^T \in \mathbb{R}^p,$$

where $\mathrm{SNR} > 0$ is a signal-to-noise parameter. Varying the parameter SNR allows us to explore the behavior of the chains when the model lies on the boundary of the $\beta_{\min}$-condition. We performed simulations for the SNR parameter $\mathrm{SNR} \in \{0.5, 1, 2, 3\}$, sample sizes $n \in \{300, 900\}$, and number of covariates $p \in \{500, 5000\}$. In all cases, we specify our prior model by setting the dispersion hyperparameter $g = p^3$ and the expected maximum model size $s_0 = 100$.

Figure 1 plots the typical trajectories of log-posterior probability versus the number of iterations of the Markov chain under the independent design. In the strong signal regime ($\mathrm{SNR} = 3$), the true model receives the highest posterior probability, and moreover the Metropolis–Hastings chain converges rapidly to stationarity, typically within $3p$ iterations. This observation is confirmation of our theoretical prediction of the behavior when all nonzero components in β^* have relative high signal-to-noise ratio ($S = \{j : \beta_j \neq 0\}$). In the intermediate signal regime ($\mathrm{SNR} = 1$), Bayesian variable-selection consistency typically fails to hold, and here, we find that the chain converges even more quickly to stationarity, typically within $1.5p$ iterations. This observation cannot be fully explained by our theory. A simulation to follow using a correlated design shows that it is not a robust



(a) (b)

FIG. 1. *Log-posterior probability versus the number of iterations* (*divided by the number of covariates p*) *of* 100 *randomly initialized Markov chains with $n = 500$, $p = 1000$ and SNR $\in \{1, 3\}$ in the independent design. In all cases*, *each grey curve corresponds to one trajectory of the chain* (100 *chains in total*). *Half of the chains are initialized at perturbations of the null model and half the true model*. (a) *Weak signal case*: *SNR* $= 1$. (b) *Strong signal case*: *SNR* $= 3$ (*the posterior probability of the true model coincides with that of the highest probability model*).

phenomenon: the chain can have poor mixing performance in this intermediate signal regime when the design is sufficiently correlated.

In order to gain further insight into the algorithm's performance, for each pair $\{X, Y\}$ we ran the Metropolis–Hastings random walk based on six initializations: the first three of them are random perturbations of the null model, whereas the remaining three are the true model. We made these choices of initialization because our empirical observations suggest that the null model and the true model tend to be near local modes of the posterior distribution. We run the Markov chain for $20p$ iterations and use the Gelman–Rubin (GR) scale factor [10] to detect whether the chains have reached stationarity. More precisely, we calculate the GR scale factor for the coefficient of determination summary statistics

$$R_\gamma^2 = \frac{Y^T \Phi_\gamma Y}{\|Y\|_2^2} \qquad \text{for } \gamma \in \{0, 1\}^p,$$

where Φ_γ denotes the projection matrix onto the span of $\{X_j, j \in \gamma\}$. Since the typical failing of convergence to stationarity is due to the multimodality of the posterior distribution, the GR scale factor can effectively detect the problem. If the chains fail to converge, then the GR scale factor will be much larger than 2; otherwise, the scale factor should be close to 1. Convergence of the chain within at most $20p$ iterations provides empirical confirmation of our theoretical prediction that the mixing time grows at most linearly in the covariate dimension p. (As will be seen in our empirical studies, the sample size n and s_0 have little impact on the mixing time, as long as s_0 remains small compared to n.)

We report the percentage of simulated datasets for which the GR scale factor from six Markov chains is less than 1.5 (success). Moreover, to see whether the variable-selection procedure based on the posterior is consistent, we also compute the difference between the highest posterior probability found during the Markov chain iterations and the posterior probability of the true model (H-T) and the difference in posterior probabilities between the null model and the true model (N-T). If the true model receives the highest posterior probability, then H-T would be 0; if the null model receives the highest posterior probability, then N-T would be the same as H-T.

Table 1 shows the results for design matrices drawn from the independent ensemble. In this case, the Markov chain method has fast convergence in all settings (it converges within $20p$ iterations). From the table, the setting $\mathrm{SNR} = 0.5$ (resp., $\mathrm{SNR} \geq 2$) corresponds to the weak (resp., strong) signal regime, while $\mathrm{SNR} = 1$ is in the intermediate regime where neither the null model nor the true model receives the highest posterior probability. Table 2 shows the results for design matrices drawn from the correlated ensemble. Now the Markov chain method exhibits poor convergence behavior in the intermediate regime $\mathrm{SNR} = 1$ with $n = 500$, but still has fast convergence in the weak and strong signal regimes. However, with larger sample size $n = 1000$, the Markov chain has fast convergence for all settings of p

TABLE 1

Convergence behavior of the Markov chain methods with sample sizes $n \in \{500, 1000\}$, ambient dimensions $p \in \{1000, 5000\}$, and SNR $\in \{0.5, 1, 2, 3\}$ in the independent design. SP: proportion of successful trials (in which GR ≤ 1.5); H-T: log posterior probability difference between the highest probability model and the true model; N-T: log posterior probability difference between the null model and the true model. Each quantity is computed based on 20 *simulated datasets*

(*n*, *p*)		**SNR = 0.5**	**SNR = 1**	**SNR = 2**	**SNR = 3**
(500, 1000)	SP	**100**	**100**	**100**	**100**
	H-T	113.4	24.6	0	0
	N-T	113.4	11.4	−210.9	−383.6
(500, 5000)	SP	**100**	**100**	**100**	**100**
	H-T	148.7	33.2	0	0
	N-T	148.7	17.4	−216.6	−395.9
(1000, 1000)	SP	**100**	**100**	**100**	**100**
	H-T	117.1	34.8	0	0
	N-T	117.1	−6.9	−342.4	−649.5
(1000, 5000)	SP	**100**	**100**	**100**	**100**
	H-T	160.4	32.8	0	0
	N-T	160.4	−4.2	−377.6	−743.4

TABLE 2

Convergence behavior of the Markov chain methods with sample size $n \in \{500, 1000\}$, ambient dimension $p \in \{1000, 5000\}$, and parameter SNR $\in \{0.5, 1, 2, 3\}$ for the case of correlated design. SP: proportion of successful trials (in which GR ≤ 1.5); H-T: log posterior probability difference between the highest probability model and the true model; N-T: log posterior probability difference between the null model and the true model. Each quantity is computed based on 20 *simulated datasets*

(*n*, *p*)		**SNR = 0.5**	**SNR = 1**	**SNR = 2**	**SNR = 3**
(500, 1000)	SP	**100**	**95**	**80**	**100**
	H-T	123.4	75.2	0	0
	N-T	123.4	71.2	−107.3	−275.8
(500, 5000)	SP	**100**	**15**	**100**	**100**
	H-T	170.0	81.0	0	0
	N-T	170.0	78.7	−102.1	−288.9
(1000, 1000)	SP	**100**	**100**	**100**	**100**
	H-T	138.7	75.1	0	0
	N-T	138.7	−67.0	−180.8	−431.7
(1000, 5000)	SP	**100**	**100**	**100**	**100**
	H-T	161.8	61.9	0	0
	N-T	161.8	−58.8	−204.2	−445.4

and SNR. Comparing the results under the two different designs, we find that correlations among the covariates increases the difficulty of variable-selection tasks when Markov chain methods are used. Moreover, the results under the correlated design suggest that there exists a regime, characterized by n, p and SNR, in which the Markov chain is slowly mixing. It would be interesting to see whether or not this regime characterizes some type of fundamental limit on computationally efficient procedures for variable selection. We leave this question open as a possible future direction.

3.4.2. *Bayesian methods versus the Lasso.* Our analysis reveals one possible benefit of a Bayesian approach as opposed to ℓ_1-based approaches such as the Lasso. It is well known that the performance of the Lasso and related ℓ_1-relaxations depends critically on fairly restrictive incoherence conditions on the design matrix. Here, we provide an example of an ensemble of linear regression problems for which the Lasso fails to perform correct variable selection whereas the Bayesian approach succeeds with high probability.

For Lasso-based methods, the irrepresentable condition

$$\max_{|\gamma|=s^*} \max_{k \notin \gamma} \left\| X_k^T X_\gamma \left(X_\gamma^T X_\gamma\right)^{-1} \right\|_1 < 1 \tag{12}$$

is both sufficient and necessary for variable-selection consistency [23, 34, 40]. In our theory for the Bayesian approach, the analogous conditions are the upper bound in Assumption D(s_0) on the maximum model size, namely

$$s_0 \geq \max\{1, (2\nu^{-2}\omega(X) + 1)s^*\}, \tag{13}$$

as well as the sparse projection condition in Assumption B. Roughly speaking, the first condition is needed to ensure that saturated models, that is, models with size s_0, receive negligible posterior probability, such that if too many unimportant covariates are included the removal of some of them does not hurt the goodness of fit (see Lemma 8 in the Supplement). This condition is weaker than the irrepresentable condition since we can always choose s_0 large enough so that $s_0 \geq \max\{1, (2\nu^{-2}\omega(X) + 1)s^*\}$ holds, as long as Assumption B is not violated.

As an example, consider a design matrix $X \in \mathbb{R}^{n\times p}$ that satisfies

$$\frac{1}{n} X^T X = \Sigma_{\text{bad}} := \begin{bmatrix} 1 & \mu & \mu & \cdots & \cdots & \mu \\ \mu & 1 & 0 & \cdots & \cdots & 0 \\ \mu & 0 & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu & 0 & 0 & \cdots & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{p\times p},$$

with $\mu = (2\sqrt{p})^{-1}$. [When $p > n$, we may consider instead a random design X where the rows of X are generated i.i.d. from the p-variate normal distribution $\mathcal{N}(0, \Sigma_{\text{bad}})$.] This example was previously analyzed by Wainwright [33], who shows that it is an interesting case in which there is a gap between the performance of ℓ_1-based variable-selection recovery and that of an optimal (but computationally intractable) method based on searching over all subsets. For a design matrix of this form, we have $\max_{|\gamma|=s^*,k\notin\gamma} \|X_k^T X_\gamma (X_\gamma^T X_\gamma)^{-1}\|_1 \geq s^*\mu$, so that the irrepresentable condition fails if $s^* > 2\sqrt{p}$. Consequently, by known results on the necessity of the irrepresentable condition for Lasso [34, 40], it will fail in performing variable selection for this ensemble.

On the other hand, for this example, it can be verified that Assumption D(s_0) is satisfied with $s_0 \geq 13s^*$, and moreover, that the RE(s) condition in Assumption B holds with $\nu = 1/2$, whereas the sparse projection condition is satisfied with $L = 16(1 + s_0^2\mu^2) = 16 + \frac{4s_0^2}{p}$. The only consequence for taking larger values of L is in the $\beta_{\min}$-condition: in particular, the threshold C_β is always lower bounded by $(128\nu^{-2}L\sigma_0^2\frac{\log p}{n})^{1/2}$. Consequently, our theory shows that the Bayesian procedure will perform correct variable selection with high probability for this ensemble.

To compare the performance of the Bayesian approach and the Lasso experimentally under this setup, we generate our design matrix from a Gaussian version of this ensemble; that is, the rows of X are generated i.i.d. from the p-variate normal distribution $\mathcal{N}(0, \Sigma_{\text{bad}})$. We choose $(n, p, s^*) = (300, 80, 20)$ so that $s^*\mu = 10/\sqrt{80} \approx 1.1 > 1$; that is, the irrepresentable condition fails. Figure 2 shows the variable-selection performance for the Bayesian approach and the Lasso over 100 replicates. We report the logarithm of the ratio between the posterior probability [see equation (A.2) in the Supplement] of the selected model and the true model, where we use the median probability model [2] as the selected model of the Bayesian approach. If a variable-selection approach has good performance, then we will expect this logarithm to be close to zero. Figure 2 shows that

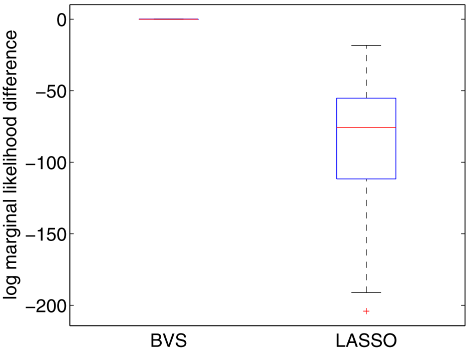


FIG. 2. *Boxplots indicating variable-selection performance of the Bayesian approach* (*BVS*) *and the Lasso. The boxplots are based on the logarithms of the ratio between the posterior probability of the selected model and the true model over* 100 *replicates. The model selected by the Bayesian approach is the median probability model* [2] *and the regularization parameter of the Lasso is chosen by cross-validation.*

the Bayesian approach almost always selects the true model while the Lasso fails most of the time, which is consistent with the theory.

4. Proofs. We now turn to the proofs of our main results, beginning with the rapid mixing guarantee in Theorem 2, which is the most involved technically. We then use some of the machinery developed in Theorem 2 to prove the posterior consistency guarantee in Theorem 1. Finally, by combining these two theorems we prove Corollary 2. In order to promote readability, we defer the proofs of certain more technical results to the appendices in the Supplement. In the proofs, we use c and c_j $(j = 0, 1, 2, \ldots)$ to denote universal constants whose magnitude may change from line to line.

4.1. *Proof of Theorem* 2. For the purposes of this proof, let $\tilde{\mathbf{P}}$ denote the transition matrix of the original Metropolis–Hastings sampler (3). Now consider instead the transition matrix $\mathbf{P} := \tilde{\mathbf{P}}/2 + \mathbf{I}/2$, corresponding to a lazy random walk that has a probability of at least 1/2 in staying in its current state. The difference $\operatorname{Gap}(\mathbf{P}) := 1 - \max\{|\lambda_2|, |\lambda_{\min}(\mathbf{P})|\}$ is known as the spectral gap. For any lazy Markov chain, the smallest eigenvalue of $\mathbf{P}$ will always be nonnegative, and as a consequence, the spectral gap of the Markov chain $\mathcal{C}$ is completely determined by the second largest eigenvalue λ_2 of $\mathbf{P}$. Then we have the sandwich relation

$$\frac{1}{2}\frac{(1-\operatorname{Gap}(\mathbf{P}))}{\operatorname{Gap}(\mathbf{P})}\log\bigl[1/(2\varepsilon)\bigr] \le \tau_\varepsilon \le \frac{(\log[1/\min_{\gamma\in\mathscr{M}}\pi(\gamma)] + \log(1/\varepsilon))}{\operatorname{Gap}(\mathbf{P})}. \tag{14}$$

See the papers [30, 35] for bounds of this form.

Using this sandwich relation, we claim that it suffices to show that there are universal constants (c_3, c_4) such that with probability at least $1 - c_3 p^{-c_4}$, the spectral gap of the lazy transition matrix $\mathbf{P}$ is lower bounded as

$$\operatorname{Gap}(\mathbf{P}) \ge \frac{1}{24 p s_0^2}. \tag{15}$$

To establish the sufficiency of this intermediate claim, we apply Theorem 1 and make use of the expression for the posterior distribution [equation (A.2) in the Supplement], thereby obtaining that for $\gamma \in \mathscr{M}$, the posterior probability is lower bounded as

$$\begin{aligned}
\pi_n(\gamma|Y) &= \pi_n(\gamma^*|Y)\cdot\frac{\pi_n(\gamma|Y)}{\pi_n(\gamma^*|Y)}\\
&\ge \bigl(1 - c_1 p^{-1}\bigr)\cdot\bigl(p^\kappa\sqrt{1+g}\bigr)^{-(|\gamma|-|\gamma^*|)}\cdot\frac{(1+g(1-R_{\gamma^*}^2))^{n/2}}{(1+g(1-R_\gamma^2))^{n/2}}\\
&\ge \frac{1}{2}\cdot p^{-(\kappa+\alpha/2)s_0}\cdot p^{-\alpha n/2}
\end{aligned}$$

with probability at least $1-c_3p^{-c_4}$ for $p \geq 2c_1$. Combining the above two displays with the sandwich relation (14), we obtain that for $\varepsilon \in (0,1)$,

$$\tau_\varepsilon \leq 12ps_0^2\big((\alpha n + \alpha s_0 + 2\kappa s_0)\log p + \log(1/\varepsilon) + \log 2\big)$$

with probability at least $1 - c_3p^{-c_4}$.

Accordingly, the remainder of our proof is devoted to establishing the spectral gap bound (15), and we do so via a version of the canonical path argument [30]. Let us begin by describing the idea of a canonical path ensemble associated with a Markov chain. Given a Markov chain $\mathcal{C}$ with state space $\mathscr{M}$, consider the weighted directed graph $G(\mathcal{C}) = (V, E)$ with vertex set $V = \mathscr{M}$ and edge set E in which an ordered pair $e = (\gamma, \gamma')$ is included as an edge with weight $\mathbf{Q}(e) = \mathbf{Q}(\gamma, \gamma') = \pi(\gamma)\mathbf{P}(\gamma, \gamma')$ if and only if $\mathbf{P}(\gamma, \gamma') > 0$. A *canonical path ensemble* $\mathcal{T}$ for $\mathcal{C}$ is a collection of paths that contains, for each ordered pair (γ, γ') of distinct vertices, a unique simple path $T_{\gamma,\gamma'}$ in the graph that connects γ and γ'. We refer to any path in the ensemble $\mathcal{T}$ as a canonical path.

In terms of this notation, Sinclair [30] shows that for any reversible Markov chain and any choice of canonical path ensemble $\mathcal{T}$, the spectral gap of $\mathbf{P}$ is lower bounded as

$$\underbrace{\mathrm{Gap}(\mathbf{P})}_{1-\lambda_2} \geq \frac{1}{\rho(\mathcal{T})\ell(\mathcal{T})}, \tag{16}$$

where $\ell(\mathcal{T})$ corresponds to the length of a longest path in the ensemble $\mathcal{T}$, and the quantity $\rho(\mathcal{T}) := \max_{e\in E} \frac{1}{\mathbf{Q}(e)} \sum_{T_{\gamma,\gamma'}\ni e} \pi(\gamma)\pi(\gamma')$ is known as the *path congestion parameter.*

In order to apply this approach to our problem, we need to construct a suitable canonical path ensemble $\mathcal{T}$. To begin with, let us introduce some notation for operations on simple paths. For two given simple paths T_1 and T_2:

- When the subset of overlapping edges of T_1 and T_2 is a connected subset of E, define the intersection path $T_1 \cap T_2$ as this subset. [For instance, if $T_1 = (1,1,1) \to (0,1,1) \to (0,0,1) \to (0,0,0)$ and $T_2 = (0,0,1) \to (0,0,0)$, then $T_1 \cap T_2 = (0,0,1) \to (0,0,0)$.]
- If $T_2 \subset T_1$, then $T_1 \setminus T_2$ denotes the path obtained by removing all edges in T_2 from T_1. [With the same specific choices of (T_1, T_2) as above, we have $T_1 \setminus T_2 = (1,1,1) \to (0,1,1) \to (0,0,1)$.]
- We use $\bar{T}_1$ to denote the reverse of T_1. [With the choice of T_1 as above, we have $\bar{T}_1 = (0,0,0) \to (0,0,1) \to (0,1,1) \to (1,1,1)$.]
- If the endpoint of T_1 and the starting point of T_2 are the same, then we define the union $T_1 \cup T_2$ as the path that connects T_1 and T_2 together. [If $T_1 = (0,0,0) \to (0,0,1)$ and $T_2 = (0,0,1) \to (0,1,1)$, then their union is given by $T_1 \cup T_2 = (0,0,0) \to (0,0,1) \to (0,1,1)$.]

We now turn to the construction of our canonical path ensemble. At a high level, our construction is inspired by the variable-selection paths carved out by greedy stepwise variable-selection procedures (e.g., [1, 39]).

Canonical path ensemble construction for $\mathscr{M}$. First, we construct the canonical path T_{γ,γ^*} from any $\gamma \in \mathscr{M}$ to the true model γ^*. The following construction will prove helpful. We call a set $\mathcal{R}$ of canonical paths *memoryless* with respect to the central state γ^* if: (1) for any state $\gamma \in \mathscr{M}$ satisfying $\gamma \neq \gamma^*$, there exists a unique simple path T_{γ,γ^*} in $\mathcal{R}$ that connects γ and γ^*; (2) for any intermediate state $\tilde{\gamma} \in \mathscr{M}$ on any path T_{γ,γ^*} in $\mathcal{R}$, the unique path $T_{\tilde{\gamma},\gamma^*}$ in $\mathcal{R}$ that connects $\tilde{\gamma}$ and γ^* is the sub-path of T_{γ,γ^*} starting from $\tilde{\gamma}$ and ending at γ^*. Intuitively, this memoryless property means that for any intermediate state on any canonical path toward the central state, the next move from this intermediate state toward the central state does not depend on the history. A memoryless canonical path ensemble has the property that in order to specify the canonical path connecting any state $\gamma \in \mathscr{M}$ and the central state γ^*, we only need to specify which state to move to from any $\gamma \neq \gamma^*$ in $\mathscr{M}$; that is, we need a transition function $\mathcal{G} : \mathscr{M} \setminus \{\gamma^*\} \to \mathscr{M}$ that maps the current state $\gamma \in \mathscr{M}$ to a next state $\mathcal{G}(\gamma) \in \mathscr{M}$. For simplicity, we define $\mathcal{G}(\gamma^*) = \gamma^*$ to make $\mathscr{M}$ as the domain of $\mathcal{G}$. Clearly, each memoryless canonical path ensemble with respect to a central state γ^* corresponds to a transition function $\mathcal{G}$ with $\mathcal{G}(\gamma^*) = \gamma^*$, but the converse is not true. For example, if there exist two states γ and γ' so that $\mathcal{G}(\gamma) = \gamma'$ and $\mathcal{G}(\gamma') = \gamma$, then $\mathcal{G}$ is not the transition function corresponding to any memoryless canonical path ensemble. However, every valid transition function $\mathcal{G}$ gives rise to a unique memoryless canonical path set consisting of paths connecting any $\gamma \in \mathscr{M}$ to γ^*, with γ^* corresponding to the fixed point of $\mathcal{G}$. We call function $\mathcal{G}$ a valid transition function if there exists a memoryless canonical path set for which $\mathcal{G}$ is the corresponding transition function. The next lemma provides a sufficient condition for a function $\mathcal{G} : \mathscr{M} \setminus \{\gamma^*\} \to \mathscr{M}$ to be valid, which motivates our construction to follow. Recall that d_H denotes the Hamming metric between a pair of binary strings.

LEMMA 1. *If a function $\mathcal{G} : \mathscr{M} \setminus \{\gamma^*\} \to \mathscr{M}$ satisfies that for any state $\gamma \in \mathscr{M} \setminus \gamma^*$, the Hamming distance between $\mathcal{G}(\gamma)$ and γ^* is strictly less than the Hamming distance between γ and γ^*, then $\mathcal{G}$ is a valid transition function.*

PROOF. Based on this function $\mathcal{G}$, we can construct the canonical path T_{γ,γ^*} from any state $\gamma \in \mathscr{M}$ to γ^* by defining T_{γ,γ^*} as $\gamma \to \mathcal{G}(\gamma) \to \mathcal{G}^2(\gamma) \to \cdots \to \mathcal{G}^{k_\gamma}(\gamma)$, where $\mathcal{G}^k := \mathcal{G} \circ \cdots \circ \mathcal{G}$ denotes the k-fold self-composition of $\mathcal{G}$ for any $k \in \mathbb{N}$ and $k_\gamma := \min_k\{\mathcal{G}^k(\gamma) = \gamma^*\}$. In order to show that the set $\{T_{\gamma,\gamma^*} : \gamma \in \mathscr{M}, \gamma \neq \gamma^*\}$ is a memoryless canonical path set, we only need to verify two things:

(a) for any $\gamma \neq \gamma^*$, T_{γ,γ^*} is a well-defined path; that is, it has finite length and ends at γ^*, and

(b) for any $\gamma \neq \gamma^*$, T_{γ,γ^*} is a simple path.

By our assumption, the function $F : \mathscr{M} \to \mathbb{R}$ defined by $F(\gamma) = d_H(\gamma, \gamma^*)$ is strictly decreasing along the path T_{γ,γ^*} for $\gamma \neq \gamma^*$. Because F only attains a finite number of values, there exists a smallest k_γ such that $\mathcal{G}^{k+1}(\gamma) = \mathcal{G}^k(\gamma)$ for each $k \geq k_\gamma$, implying that $\mathcal{G}^{k_\gamma}(\gamma)$ is a fixed point of $\mathcal{G}$. Since γ^* is the unique fixed point of $\mathcal{G}$, we must have $\mathcal{G}^{k_\gamma}(\gamma) = \gamma^*$, which proves the first claim. The second claim is obvious since the function F defined above is strictly decreasing along the path T_{γ,γ^*}, which means that the states on the path T_{γ,γ^*} are all distinct. □

Equipped with this lemma, we start constructing a memoryless set of canonical paths from any state $\gamma \in \mathscr{M}$ to γ^* by specifying a valid $\mathcal{G}$ function. First, we introduce some definitions on the states. A state $\gamma \neq \gamma^*$ is called *saturated* if $|\gamma| = s_0$ and *unsaturated* if $|\gamma| < s_0$. We call a state $\gamma \neq \gamma^*$ *overfitted* if it contains all influential covariates, that is, $\gamma^* \subset \gamma$, and *underfitted* if it does not contain at least one influential covariate. Recall the two updating schemes in our Metropolis–Hastings (MH) sampler: single flip and double flips. We accordingly construct the transition function $\mathcal{G}$ as follows:

(i) If $\gamma \neq \gamma^*$ is overfitted, then we define $\mathcal{G}(\gamma)$ to be γ', which is formed by deleting the least influential covariate from γ, that is, $\gamma'_j = \gamma_j$ for any $j \neq \ell_\gamma$ and $\gamma'_{\ell_\gamma} = 0$, where ℓ_γ is the index from the set $\gamma \setminus \gamma^*$ of uninfluential covariates that minimizes the difference

$$\|\Phi_\gamma X_{\gamma^*}\beta^*_{\gamma^*}\|_2^2 - \|\Phi_{\gamma\setminus\{\ell\}} X_{\gamma^*}\beta^*_{\gamma^*}\|_2^2,$$

where Φ_γ denotes the projection onto the span of $\{X_j, j \in \gamma\}$. This transition resembles the backward deletion step in the stepwise variable-selection procedure and involves the single flip updating scheme of the MH algorithm. By construction, if $\gamma \neq \gamma^*$ is overfitted, then $d_H(\mathcal{G}(\gamma), \gamma^*) = d_H(\gamma, \gamma^*) - 1$.

(ii) If $\gamma \neq \gamma^*$ is underfitted and unsaturated, then we define $\mathcal{G}(\gamma)$ to be γ', which is formed by adding the influential covariate from $\gamma^* \setminus \gamma$ that explains the most signal variation, that is, $\gamma'_j = \gamma_j$ for any $j \neq j_\gamma$ and $\gamma'_{j_\gamma} = 1$, where j_γ is defined as the $j \in \gamma^* \setminus \gamma$ that maximizes the quantity $\|\Phi_{\gamma\cup\{j\}} X_{\gamma^*}\beta^*_{\gamma^*}\|_2^2$. This transition remsembles the forward selection step in the stepwise variable selection procedure and involves the single flip updating scheme of the MH algorithm. By construction, if $\gamma \neq \gamma^*$ is underfitted and unsaturated, then $d_H(\mathcal{G}(\gamma), \gamma^*) = d_H(\gamma, \gamma^*) - 1$.

(iii) If $\gamma \neq \gamma^*$ is underfitted and saturated, then we define $\mathcal{G}(\gamma)$ to be γ', which is formed by replacing the least influential unimportant covariate in γ with the most influential covariate from $\gamma^* \setminus \gamma$, that is, $\gamma'_j = \gamma_j$ for any $j \notin \{j_\gamma, k_\gamma\}$, $\gamma'_{j_\gamma} = 1$ and $\gamma'_{k_\gamma} = 0$, where j_γ is defined in case 2 and $k_\gamma \in \gamma \setminus \gamma^*$ minimizes $\|\Phi_{\gamma\cup\{j\}} X_{\gamma^*}\beta^*_{\gamma^*}\|_2^2 - \|\Phi_{\gamma\cup\{j\}\setminus\{k\}} X_{\gamma^*}\beta^*_{\gamma^*}\|_2^2$. This transition step involves the double-flip updating scheme of the MH algorithm. By construction, if $\gamma \neq \gamma^*$ is underfitted and saturated, then $d_H(\mathcal{G}(\gamma), \gamma^*) = d_H(\gamma, \gamma^*) - 2$.

By Lemma 1, this transition function $\mathcal{G}$ is valid and gives rise to a unique memoryless set of canonical paths from any state $\gamma \in \mathscr{M}$ to γ^*. For example, Figure 3

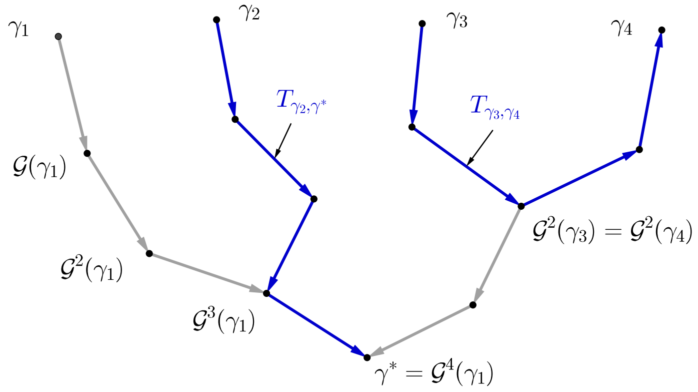

γ_1
γ_2
γ_3
γ_4
T_{γ_2,γ^*}
T_{γ_3,γ_4}
$\mathcal{G}(\gamma_1)$
$\mathcal{G}^2(\gamma_1)$
$\mathcal{G}^3(\gamma_1)$
$\mathcal{G}^2(\gamma_3) = \mathcal{G}^2(\gamma_4)$
$\gamma^* = \mathcal{G}^4(\gamma_1)$

FIG. 3. *Illustration of the construction of the canonical path ensemble. In the plot, γ^* is the central state, $\mathcal{G}$ is the transition function and solid blue arrows indicate canonical paths T_{γ_2,γ^*} and T_{γ_3,γ_4}.*

shows such a memoryless set of canonical paths for $\mathscr{M}$ consisting of 14 states, where T_{γ_2,γ^*} corresponds to the canonical path from state γ_2 to the central state γ^*.

Based on this memoryless canonical path set, we can finish constructing the canonical path ensemble $\mathcal{T}$ by specifying the path $T_{\gamma,\gamma'}$ connecting any distinct pair $(\gamma, \gamma') \in \mathscr{M} \times \mathscr{M}$. More specifically, by the memoryless property, the two simple paths T_{γ,γ^*} and T_{γ',γ^*} share an identical subpath toward γ^* from their first common intermediate state and, therefore, have a valid intersection path according to our definition. Let $T_{\gamma\cap\gamma'}$ denote this common subpath $T_{\gamma,\gamma^*} \cap T_{\gamma',\gamma^*}$, and let $T_{\gamma\backslash\gamma'} := T_{\gamma,\gamma^*} \setminus T_{\gamma\cap\gamma'}$ denote the remaining path of T_{γ,γ^*} after removing the segment $T_{\gamma\cap\gamma'}$. We define $T_{\gamma'\backslash\gamma}$ in a similar way as $T_{\gamma',\gamma^*} \setminus T_{\gamma\cap\gamma'}$. Then it is easy to see that the two remaining paths $T_{\gamma\backslash\gamma'}$ and $T_{\gamma'\backslash\gamma}$ share the same endpoint. Therefore, it is valid to define the path $T_{\gamma,\gamma'}$ as $T_{\gamma\backslash\gamma'} \cup \bar{T}_{\gamma'\backslash\gamma}$. To understand this construction, let us consider an example where $T_{\gamma,\gamma^*} = (0, 1, 1, 1) \to (1, 1, 0, 1) \to (1, 1, 0, 0)$ and $T_{\gamma',\gamma^*} = (1, 0, 0, 1) \to (1, 1, 0, 1) \to (1, 1, 0, 0)$. Their intersection is $T_{\gamma\cap\gamma'} = (1, 1, 0, 1) \to (1, 1, 0, 0)$ and the two remaining paths are $T_{\gamma\backslash\gamma'} = (0, 1, 1, 1) \to (1, 1, 0, 1)$ and $T_{\gamma'\backslash\gamma} = (1, 0, 0, 1) \to (1, 1, 0, 1)$. Consequently, the path $T_{\gamma,\gamma'}$ from γ to γ' is $(0, 1, 1, 1) \to (1, 1, 0, 1) \to (1, 0, 0, 1)$ by our construction. For example, path T_{γ_3,γ_4} in Figure 3 illustrates the construction of the path connecting (γ_3, γ_4) when $\mathscr{M}$ is composed of 14 states.

We call γ a *precedent* of γ' if γ' is on the canonical path $T_{\gamma,\gamma^*} \in \mathcal{T}$, and a pair of states γ, γ' *adjacent* if the canonical path $T_{\gamma,\gamma'}$ is $e_{\gamma,\gamma'}$, the edge in E connecting γ and γ'. For $\gamma \in \mathscr{M}$, let

$$\Lambda(\gamma) := \{\bar{\gamma} | \gamma \in T_{\bar{\gamma},\gamma^*}\} \tag{17}$$

denote the set of all its precedents. Use the notation $|T|$ to denote the length of a path T. The following lemma provides some important properties of the constructed canonical path ensemble that will be used later.

LEMMA 2. *For any distinct pair* $(\gamma, \gamma') \in \mathscr{M} \times \mathscr{M}$:

(a) *We have*

$$|T_{\gamma,\gamma^*}| \leq d_H(\gamma, \gamma^*) \leq s_0 \quad \textit{and} \tag{18a}$$

$$|T_{\gamma,\gamma'}| \leq d_H(\gamma, \gamma^*) + d_H(\gamma', \gamma^*) \leq 2s_0. \tag{18b}$$

(b) *If* γ *and* γ' *are adjacent* (*joined by edge* $e_{\gamma,\gamma'}$) *and* γ *is a precedent of* γ', *then*

$$\{(\bar{\gamma}, \bar{\gamma}') | T_{\bar{\gamma},\bar{\gamma}'} \ni e_{\gamma,\gamma'}\} \subset \Lambda(\gamma) \times \mathscr{M}.$$

PROOF. The first claim follows since the function $F : \mathscr{M} \to \mathbb{R}$ defined by $F(\gamma) = d_H(\gamma, \gamma^*)$ is strictly decreasing along the path T_{γ,γ^*} for $\gamma \neq \gamma^*$. Now we prove the second claim. For any pair $(\bar{\gamma}, \bar{\gamma}')$ such that $T_{\bar{\gamma},\bar{\gamma}'} \ni e_{\gamma,\gamma'}$, either $e_{\gamma,\gamma'} \in T_{\bar{\gamma}\backslash\bar{\gamma}'}$ or $e_{\gamma',\gamma} \in T_{\bar{\gamma}'\backslash\bar{\gamma}}$ should be satisfied since $T_{\bar{\gamma},\bar{\gamma}'} = T_{\bar{\gamma}\backslash\bar{\gamma}'} \cup \bar{T}_{\bar{\gamma}',\bar{\gamma}}$ by our construction. Because γ is a precedent of γ', we can only have $e_{\gamma,\gamma'} \in T_{\bar{\gamma}\backslash\bar{\gamma}'}$. This shows that γ is on the path $T_{\bar{\gamma},\gamma^*}$ and $\bar{\gamma} \in \Lambda(\gamma)$. □

According to Lemma 2(b), the path congestion parameter $\rho(\mathcal{T})$ of the canonical path ensemble $\mathcal{T}$ satisfies

$$\rho(\mathcal{T}) \leq \max_{(\gamma,\gamma')\in\Gamma^*} \frac{1}{\mathbf{Q}(\gamma, \gamma')} \sum_{\bar{\gamma}\in\Lambda(\gamma),\bar{\gamma}'\in\mathscr{M}} \pi(\bar{\gamma})\pi(\bar{\gamma}') = \max_{(\gamma,\gamma')\in\Gamma^*} \frac{\pi[\Lambda(\gamma)]}{\mathbf{Q}(\gamma, \gamma')}, \tag{19}$$

where the maximum is taken over the set

$$\Gamma^* := \{(\gamma, \gamma') \in \mathscr{M} \times \mathscr{M} | T_{\gamma,\gamma'} = e_{\gamma,\gamma'} \text{ and } \gamma \in \Lambda(\gamma')\}.$$

Here, we used the fact that the weight function $\mathbf{Q}$ of a reversible chain satisfies $\mathbf{Q}(\gamma, \gamma') = \mathbf{Q}(\gamma', \gamma)$ so as to be able to restrict the range of the maximum to pairs (γ, γ') where $\gamma \in \Lambda(\gamma')$.

For the lazy form of the Metropolis–Hastings walk (3), given any pair (γ, γ') such that $\mathbf{P}(\gamma, \gamma') > 0$, we have

$$\begin{aligned} \mathbf{Q}(\gamma, \gamma') &= \frac{1}{2}\pi_n(\gamma|Y)\mathbf{P}(\gamma, \gamma') \\ &\geq \frac{1}{2ps_0}\pi_n(\gamma|Y)\min\left\{1, \frac{\pi_n(\gamma'|Y)}{\pi_n(\gamma|Y)}\right\} = \frac{1}{2ps_0}\min\{\pi_n(\gamma'|Y), \pi_n(\gamma|Y)\}. \end{aligned}$$

Substituting this lower bound into our upper bound (19) on the path congestion parameter yields

$$\begin{aligned} \rho(\mathcal{T}) &\leq 2ps_0 \max_{(\gamma,\gamma')\in\Gamma^*} \frac{\pi_n[\Lambda(\gamma)|Y]}{\min\{\pi_n(\gamma|Y), \pi_n(\gamma'|Y)\}} \\ &= 2ps_0 \max_{(\gamma,\gamma')\in\Gamma^*} \left\{\max\left\{1, \frac{\pi_n(\gamma|Y)}{\pi_n(\gamma'|Y)}\right\} \cdot \frac{\pi_n[\Lambda(\gamma)|Y]}{\pi_n(\gamma|Y)}\right\}. \end{aligned} \tag{20}$$

In order to prove that $\rho(\mathcal{T}) = \mathcal{O}(ps_0)$ with high probability, it suffices to show that the two terms inside the maximum are $\mathcal{O}(1)$ with high probability. In order to do so, we make use of two auxiliary lemmas.

Given the noise vector $w \sim \mathcal{N}(0, \sigma_0^2 I_n)$, consider the following events:

$$\mathcal{A}_n := \left\{ \max_{\substack{(\gamma_1,\gamma_2)\in\mathscr{M}\times\mathscr{M} \\ \gamma_2\subset\gamma_1}} \frac{w^T(\Phi_{\gamma_1} - \Phi_{\gamma_2})w}{|\gamma_1| - |\gamma_2|} \le L\sigma_0^2 \log p \right\}, \tag{21a}$$

$$\mathcal{B}_n := \left\{ \max_{\gamma\in\mathscr{M}} \frac{w^T\Phi_\gamma w}{|\gamma|} \le 8\sigma_0^2 \log p \right\} \quad \text{and} \tag{21b}$$

$$\mathcal{C}_n := \left\{ \left| \frac{\|w\|_2^2}{n\sigma_0^2} - 1 \right| \le \frac{1}{2} \right\} \quad \text{and} \quad \mathcal{D}_n := \left\{ \frac{\|Y\|_2^2}{g} \le (2\log p + 3)\sigma_0^2 \right\}. \tag{21c}$$

Our first auxiliary lemma guarantees that, under the stated assumptions of our theorem, the intersection of these events holds with high probability.

LEMMA 3. *Under the conditions of Theorem 2, we have*

$$\mathbb{P}(\mathcal{A}_n \cap \mathcal{B}_n \cap \mathcal{C}_n \cap \mathcal{D}_n) \ge 1 - c_1 p^{-c_2}. \tag{22}$$

We prove this lemma in Section 4.2 to follow.

Our second auxiliary lemma ensures that when these four events hold, then the two terms on the right-hand side of the upper bound (20) are controlled.

LEMMA 4. *Suppose that, in addition to the conditions of Theorem 2, the compound event $\mathcal{A}_n \cap \mathcal{B}_n \cap \mathcal{C}_n \cap \mathcal{D}_n$ holds. Then for all $\gamma \neq \gamma^*$, we have*

$$\frac{\pi_n(\gamma|Y)}{\pi_n(\mathcal{G}(\gamma)|Y)} \le \begin{cases} p^{-2}, & \text{if } \gamma \text{ is overfitted,} \\ p^{-3}, & \text{if } \gamma \text{ is underfitted,} \end{cases} \tag{23a}$$

and moreover, for all γ,

$$\frac{\pi_n[\Lambda(\gamma)|Y]}{\pi_n(\gamma|Y)} \le 6. \tag{23b}$$

We prove this lemma in Section 4.3 to follow.

Combining Lemmas 3 and 4 with our earlier bound (20), we conclude that $\rho(\mathcal{T}) \le 12ps_0$. By Lemma 2(a), our path ensemble $\mathcal{T}$ has maximal length $\ell(\mathcal{T}) \le 2s_0$, and hence the canonical path lower bound (16) implies that $\text{Gap}(\mathbf{P}) \ge \frac{1}{24ps_0^2}$, as claimed in inequality (15). This completes the proof of the theorem.

The only remaining detail is to prove Lemmas 3 and 4, and we do so in the following two subsections.

4.2. *Proof of Lemma 3.*

We split the proof up into separate parts, one for each of the events $\mathcal{A}_n, \mathcal{B}_n, \mathcal{C}_n$ and $\mathcal{D}_n$.

Bound on $\mathbb{P}[\mathcal{C}_n]$. Since $\|w\|_2^2/\sigma_0^2 \sim \chi_n^2$, a standard tail bound for the χ_n^2 distribution (e.g., [18], Lemma 1) yields

$$\mathbb{P}[\mathcal{C}_n] \geq 1 - 2e^{-n/25} \geq 1 - 2p^{-1}, \tag{24}$$

where in the last step we used Assumption D which implies $n \geq 32 \log p$.

Bound on $\mathbb{P}[\mathcal{B}_n]$. For each state $\gamma \in \mathscr{M}$, the random variable $w^T \Phi_\gamma w/\sigma_0^2$ follows a chi-squared distribution with $|\gamma|$ degrees of freedom. For each integer $\ell \in \{1, \dots, s_0\}$, the model space $\mathscr{M}$ contains $\binom{p}{\ell}$ models of size ℓ. Therefore, by a union bound, we find that for $p > 1$,

$$\begin{aligned}\mathbb{P}[\mathcal{B}_n] &\geq 1 - \sum_{\ell=1}^{s_0} \binom{p}{\ell} \mathbb{P}(\chi_\ell^2 \geq 8\ell \log p) \geq 1 - \sum_{l=1}^{s_0} e^{-\ell \log p} \\ &\geq 1 - 2e^{-\log p} \\ &= 1 - 2p^{-1}.\end{aligned} \tag{25}$$

Bound on $\mathbb{P}[\mathcal{D}_n]$. Given the linear observation model, we have

$$\|Y\|_2^2 = \|X\beta^* + w\|_2^2 \leq 2\|X\beta^*\|^2 + 2\|w\|_2^2.$$

Combining this with inequality (24), we obtain

$$\mathbb{P}[\|Y\|_2^2 \geq 2\|X\beta^*\|_2^2 + 3n\sigma_0^2] \leq 2e^{-n/25} \leq 2p^{-1},$$

where we have used Assumption D that implies $n \geq 32 \log p$. By Assumption A, we have $\|X\beta^*\|_2^2 \leq g\sigma_0^2 \log p$, implying that for $g \geq n$ (which is the case under Assumption C),

$$\mathbb{P}[\mathcal{D}_n^c] \leq \mathbb{P}[\|Y\|_2^2 \geq 2\|X\beta^*\|_2^2 + 3n\sigma_0^2] \leq 2p^{-1}. \tag{26}$$

Bound on $\mathbb{P}[\mathcal{A}_n]$. To control this probability, we require two auxiliary lemmas.

LEMMA 5. *Under Assumption* B, *for any distinct pair* $(\gamma, \bar{\gamma}) \in \mathscr{M} \times \mathscr{M}$ *satisfying* $\gamma \subset \bar{\gamma}$, *we have*

$$\lambda_{\min}\left(\frac{1}{n} X_{\bar{\gamma}\backslash\gamma}^T (I_n - \Phi_\gamma) X_{\bar{\gamma}\backslash\gamma}\right) \geq \nu.$$

PROOF. By partitioning the matrix $X_{\bar{\gamma}}$ into a block form $(X_\gamma, X_{\bar{\gamma}\backslash\gamma})$ and using the formula for the inverse of block matrices, one can show that the lower right corner of $(n^{-1}X_{\bar{\gamma}}^T X_{\bar{\gamma}})^{-1}$ is $(n^{-1}X_{\bar{\gamma}\backslash\gamma}^T (I_n - \Phi_\gamma) X_{\bar{\gamma}\backslash\gamma})^{-1}$, which implies the claimed bound. □

LEMMA 6. *For $\gamma \in \mathscr{M}$ and $k \notin \gamma$, we have*

$$\Phi_{\gamma \cup \{k\}} - \Phi_\gamma = \frac{(I - \Phi_\gamma) X_k X_k^T (I - \Phi_\gamma)}{X_k^T (I - \Phi_\gamma) X_k}.$$

PROOF. By the block matrix inversion formula [14], we have

$$\begin{bmatrix} X_\gamma^T X_\gamma & X_\gamma^T X_k \\ X_k^T X_\gamma & X_k^T X_k \end{bmatrix}^{-1} = \begin{bmatrix} B + aBX_\gamma^T X_k X_k^T XB & -aBX_\gamma^T X_k \\ -aX_k^T X_\gamma B & a \end{bmatrix},$$

where $B = (X_\gamma^T X_\gamma)^{-1} \in \mathbb{R}^{|\gamma| \times |\gamma|}$ and $a = (X_k^T (I - \Phi_\gamma) X_k)^{-1} \in \mathbb{R}$. Then simple linear algebra yields

$$\begin{aligned} \Phi_{\gamma \cup \{k\}} - \Phi_\gamma &= [\, X_\gamma \quad X_k \,] \begin{bmatrix} X_\gamma^T X_\gamma & X_\gamma^T X_k \\ X_k^T X_\gamma & X_k^T X_k \end{bmatrix}^{-1} \begin{bmatrix} X_\gamma^T \\ X_k^T \end{bmatrix} - \Phi_\gamma \\ &= a(I - \Phi_\gamma) X_k X_k^T (I - \Phi_\gamma), \end{aligned}$$

which is the claimed decomposition. □

Returning to our main task, let us define the event

$$\mathcal{A}_n' := \Big\{ \max_{\substack{\gamma \in \mathscr{M}, k \in \{1,\dots,p\} \\ \text{s.t. } k \notin \gamma}} w^T (\Phi_{\gamma \cup \{k\}} - \Phi_\gamma) w \le L \sigma_0^2 \log p \Big\}.$$

By construction, we have $\mathcal{A}_n' \subseteq \mathcal{A}_n$ so that it suffices to lower bound $\mathbb{P}(\mathcal{A}_n')$. Lemma 6 implies that

$$w^T (\Phi_{\gamma \cup \{k\}} - \Phi_\gamma) w = \frac{|\langle (I - \Phi_\gamma) X_k, w \rangle|^2 / n}{X_k^T (I - \Phi_\gamma) X_k / n}. \tag{27}$$

Now we show that with probability at least $1 - p^{-c}$, the above quantity is uniformly bounded by $L\sigma_0^2 \log p$ over all $(\gamma, k) \in \mathscr{M} \times \{1, \dots, p\}$ satisfying $|\gamma| \le s_0$ and $k \notin \gamma$, which yields the intermediate result

$$\mathbb{P}(\mathcal{A}_n) \ge \mathbb{P}(\mathcal{A}_n') \ge 1 - p^{-c_4}. \tag{28}$$

Now Lemma 5 implies that $\frac{1}{n} X_k^T (I - \Phi_\gamma) X_k \ge \nu$ and, therefore, if we define the random variable

$$V(Z) := \max_{\substack{\gamma \in \mathscr{M}, k \in \{1,\dots,p\} \\ \text{s.t. } k \notin \gamma}} \frac{1}{\sqrt{n}} \big| \langle (I - \Phi_\gamma) X_k, Z \rangle \big| \qquad \text{where } Z \sim N(0, I_n),$$

then it suffices to show that $V(Z) \le \sqrt{L\nu \log p}$ with probability at least $1 - p^{-c}$. For any two vectors $Z, Z' \in \mathbb{R}^n$, we have

$$\begin{aligned} \big| V(Z) - V(Z') \big| &\le \max_{\substack{\gamma \in \mathscr{M}, k \in \{1,\dots,p\} \\ \text{s.t. } k \notin \gamma}} \frac{1}{\sqrt{n}} \big| \langle (I - \Phi_\gamma) X_k, Z - Z' \rangle \big| \\ &\le \frac{1}{\sqrt{n}} \| (I - \Phi_\gamma) X_k \|_2 \| Z - Z' \|_2 \le \| Z - Z' \|_2, \end{aligned}$$

where we have used the normalization condition of Assumption B in the last inequality. Consequently, by concentration of measure for Lipschitz functions of Gaussian random variables [19], we have

$$\mathbb{P}\big[V(Z) \ge \mathbb{E}\big[V(Z)\big] + t\big] \le e^{-t^2/2}. \tag{29}$$

By the sparse projection condition in Assumption B, the expectation satisfies $\mathbb{E}[V(Z)] \le \sqrt{L\nu \log p}/2$, which combined with (29) yields the claimed bound (28) with $c_4 = 1/2 \le L\nu/8$.

4.3. *Proof of Lemma* 4. We defer the proof of the claim (23a) to Appendix B in the Supplement as it is somewhat involved technically. It is worth mentioning that its proof uses some auxiliary results in Lemma 8 in Appendix B.4 in the Supplement, which characterizes some key properties of the state $\mathcal{G}(\gamma)$ selected by the transition function $\mathcal{G}$ via the greedy criterion.

It remains to prove the second bound (23b) in Lemma 4, and we split our analysis into two cases, depending on whether γ is underfitted or overfitted.

4.3.1. *Case γ is underfitted.* In this case, the bound (23a) implies that $\frac{\pi_n(\gamma|Y)}{\pi_n(\mathcal{G}(\gamma)|Y)} \le p^{-3}$. For each $\bar{\gamma} \in \Lambda(\gamma)$, where $\Lambda(\gamma)$ is defined in (17), we know $\gamma \in T_{\bar{\gamma},\gamma} \subset T_{\bar{\gamma},\gamma^*}$. Let the path $T_{\bar{\gamma},\gamma}$ be $\gamma_0 \to \gamma_1 \to \cdots \to \gamma_s$, where $s = |T_{\bar{\gamma},\gamma}|$ is the length of $T_{\bar{\gamma},\gamma}$, and $\gamma_0 = \bar{\gamma}$ and $\gamma_s = \gamma$ are the two endpoints. Since any intermediate state $\tilde{\gamma}$ on path $T_{\bar{\gamma},\gamma}$ is also underfitted, inequality (23a) ensures that

$$\frac{\pi_n(\bar{\gamma}|Y)}{\pi_n(\gamma|Y)} = \prod_{\ell=1}^{s} \frac{\pi_n(\gamma_{\ell-1}|Y)}{\pi_n(\gamma_l|Y)} \le p^{-3s} = p^{-3|T_{\bar{\gamma},\gamma}|}.$$

Now for each $s \in \{0, \dots, s^*\}$, we count the total number of states $\bar{\gamma}$ in $\Lambda(\gamma)$ that satisfies $|T_{\bar{\gamma},\gamma}| = s$. By construction, at each intermediate state in a canonical path, we either add a new influential covariate by the single flip updating scheme of the MH algorithm, or add a new influential covariate and delete an unimportant covariate by the double-flip updating scheme. As a consequence, any state in $\mathscr{M}$ has at most $(s^* + 1)p$ adjacent precedents, implying that the total number of states $\bar{\gamma}$ in $\Lambda(\gamma)$ with path length $|T_{\bar{\gamma},\gamma}| = s$ is upper bounded by $(s^* + 1)^s p^s$. Consequently, we have by the preceding display that under the event $\mathcal{A}_n \cap \mathcal{B}_n \cap \mathcal{C}_n \cap \mathcal{D}_n$

$$\begin{aligned} \frac{\pi_n[\Lambda(\gamma)|Y]}{\pi_n(\gamma|Y)} &= \sum_{\bar{\gamma} \in \int(\gamma)} \frac{\pi_n(\bar{\gamma}|Y)}{\pi_n(\gamma|Y)} \le \sum_{s=0}^{s^*} p^s (s^* + 1)^s p^{-3s} \\ &\le \sum_{s=0}^{\infty} p^{-s} \le \frac{1}{1 - 1/p}. \end{aligned} \tag{30}$$

The above argument is also valid for $\gamma = \gamma^*$.

4.3.2. *Case γ is overfitted.* In this case, we bound the ratio $\frac{\pi_n[\Lambda(\gamma)|Y]}{\pi_n(\gamma|Y)}$ by dividing the set $\Lambda(\gamma)$ into two subsets:

(a) Overfitted models: $\mathbb{M}_1 = \{\gamma' \in \Lambda(\gamma) : \gamma' \supset \gamma^*\}$, all models in $\Lambda(\gamma)$ that include all influential covariates.

(b) Underfitted models: $\mathbb{M}_2 = \{\gamma' \in \Lambda(\gamma) : \gamma' \not\supset \gamma^*\}$, all models in $\Lambda(\gamma)$ that miss at least one influential covariate.

First, we consider the ratio $\pi_n(\mathbb{M}_1|Y)/\pi_n(\gamma|Y)$. For each model $\bar{\gamma} \in \mathbb{M}_1$, according to our construction of the canonical path, all intermediate states on path $T_{\bar{\gamma},\gamma} = \gamma_0 \to \gamma_1 \to \cdots \to \gamma_k$ correspond to overfitted models (only involve the first flipping updating scheme of the MH algorithm), where endpoints $\gamma_0 = \bar{\gamma}$ and $\gamma_k = \gamma$, and k denotes the length of path $T_{\bar{\gamma},\gamma}$. As a consequence, inequality (23a) implies that

$$\frac{\pi_n(\bar{\gamma}|Y)}{\pi_n(\gamma|Y)} = \prod_{s=1}^{k} \frac{\pi_n(\gamma_{s-1}|Y)}{\pi_n(\gamma_s|Y)} \le p^{-2k}.$$

Since there are at most p^k states $\bar{\gamma}$ in $\mathbb{M}_1$ satisfying $|\bar{\gamma}| - |\gamma| = k$, we obtain that under the event $\mathcal{A}_n \cap \mathcal{B}_n \cap \mathcal{C}_n \cap \mathcal{D}_n$

$$\frac{\pi_n(\mathbb{M}_1|Y)}{\pi_n(\gamma|Y)} \le \sum_{k=0}^{p-|\gamma|} p^k p^{-2k} \le \sum_{k=0}^{\infty} p^{-k} \le \frac{1}{1-1/p} \le 2. \tag{31}$$

Second, we consider the ratio $\pi_n(\mathbb{M}_2|Y)/\pi_n(\gamma|Y)$. For fixed $\bar{\gamma} \in \mathbb{M}_2$, let $f(\bar{\gamma})$ be the first state along the path $T_{\bar{\gamma},\gamma}$ that contains all influential covariates. Since the overfitted state γ contains all influential covariates, $f(\bar{\gamma})$ exists and is well-defined. Moreover, this construction ensures that $f(\bar{\gamma}) \in \mathbb{M}_1$ and $\bar{\gamma} \subset \Lambda(f(\bar{\gamma})) \setminus \{f(\bar{\gamma})\}$. Applying inequality (23a) then yields

$$\begin{aligned}
\frac{\pi_n(\mathbb{M}_2|Y)}{\pi_n(\gamma|Y)} &= \sum_{\bar{\gamma}\in\mathbb{M}_2} \frac{\pi_n(\bar{\gamma}|Y)}{\pi_n(\gamma|Y)} \\
&= \sum_{\bar{\gamma}\in\mathbb{M}_2} \frac{\pi_n(f(\bar{\gamma})|Y)}{\pi_n(\gamma|Y)} \cdot \frac{\pi_n(\bar{\gamma}|Y)}{\pi_n(f(\bar{\gamma})|Y)} \\
&\le \sum_{\substack{\exists\bar{\gamma}\in\mathbb{M}_2 \\ \text{such that } \tilde{\gamma}=f(\bar{\gamma})}} \frac{\pi_n(\tilde{\gamma}|Y)}{\pi_n(\gamma|Y)} \sum_{\bar{\gamma}\in\Lambda(\tilde{\gamma})\setminus\{\tilde{\gamma}\}} \frac{\pi_n(\bar{\gamma}|Y)}{\pi_n(\tilde{\gamma}|Y)} \\
&= \sum_{\substack{\exists\bar{\gamma}\in\mathbb{M}_2 \\ \text{such that } \tilde{\gamma}=f(\bar{\gamma})}} \frac{\pi_n(\tilde{\gamma}|Y)}{\pi_n(\gamma|Y)} \cdot \left(\frac{\pi_n[\Lambda(\tilde{\gamma})|Y]}{\pi_n(\tilde{\gamma}|Y)} - 1\right).
\end{aligned}$$

Then, by treating $\tilde{\gamma} = f(\bar{\gamma}) \in \mathbb{M}_1$ as the γ in inequality (30) and inequality (31), we obtain that under the event $\mathcal{A}_n \cap \mathcal{B}_n \cap \mathcal{C}_n \cap \mathcal{D}_n$

$$\begin{aligned}\frac{\pi_n(\mathbb{M}_2|Y)}{\pi_n(\gamma|Y)} &\leq \sum_{\substack{\exists \bar{\gamma}\in\mathbb{M}_2 \\ \text{s.t. } \tilde{\gamma}=f(\bar{\gamma})}} \frac{\pi_n(\tilde{\gamma}|Y)}{\pi_n(\gamma|Y)} \cdot \left\{\frac{1}{1-1/p} - 1\right\} \\ &\leq \frac{2}{p} \sum_{\tilde{\gamma}\in\mathbb{M}_1} \frac{\pi_n(\tilde{\gamma}|Y)}{\pi_n(\gamma|Y)} \\ &= \frac{2}{p} \frac{\pi_n(\mathbb{M}_1|Y)}{\pi_n(\gamma|Y)} \\ &\leq \frac{4}{p}.\end{aligned} \tag{32}$$

Combining inequality (31) and inequality (32), we obtain that under the event $\mathcal{A}_n \cap \mathcal{B}_n \cap \mathcal{C}_n \cap \mathcal{D}_n$, the posterior ratio is upper bounded as

$$\frac{\pi_n[\Lambda(\gamma)|Y]}{\pi_n(\gamma|Y)} = \frac{\pi_n(\mathbb{M}_1|Y)}{\pi_n(\gamma|Y)} + \frac{\pi_n(\mathbb{M}_2|Y)}{\pi_n(\gamma|Y)} \leq 6. \tag{33}$$

The above argument is also valid for $\gamma = \gamma^*$, and this completes the proof of inequality (23b).

4.4. *Proof of Theorem* 1. We divide the analysis into two steps. In the first step, we show that the total posterior probability assigned to models with size $\mathcal{O}(\max\{1, s^*\})$ other than γ^* is small. In the second step, we use the fact that all large models receive small prior probabilities to show that the remaining models should also receive small posterior probability.

Step 1. Let $\mathbb{M}_S := \{\gamma \in \{0, 1\}^p : |\gamma| \leq K \max\{1, s^*\}, \gamma \neq \gamma^*\}$ denote the set of all models with moderate sizes, where $K \geq 1$ is some constant to be determined in step 2. Consider the quantity

$$\frac{\pi_n(\mathbb{M}_S|Y)}{\pi_n(\gamma^*|Y)} = \sum_{\gamma\in\mathbb{M}_S} \frac{\pi_n(\gamma|Y)}{\pi_n(\gamma^*|Y)}. \tag{34}$$

Similar to Lemma 3, we modify the definition of the four events $\mathcal{A}_n, \mathcal{B}_n, \mathcal{C}_n$ and $\mathcal{D}_n$ by replacing $\mathscr{M}$ with $\mathbb{M}_S$. Following the proof of Lemma 3, it is straightforward to show that these four events satisfy

$$\mathbb{P}[\mathcal{A}_n \cap \mathcal{B}_n \cap \mathcal{C}_n \cap \mathcal{D}_n] \geq 1 - c_1 p^{-c_2}. \tag{35}$$

The following auxiliary lemma ensures that when these four events hold, then the posterior ratios on the right-hand side of equation (34) are well controlled.

LEMMA 7. *Under Assumptions A–D and under the event $\mathcal{A}_n \cap \mathcal{B}_n \cap \mathcal{C}_n \cap \mathcal{D}_n$, the posterior ratio of any γ ($\neq \gamma^*$) in $\mathbb{M}_S$ is bounded as*

$$\frac{\pi_n(\gamma|Y)}{\pi_n(\gamma^*|Y)} \le \begin{cases} p^{-2|\gamma\setminus\gamma^*|}, & \text{if } \gamma \text{ is overfitted,} \\ p^{-2|\gamma\setminus\gamma^*|-2|\gamma^*\setminus\gamma|-2}, & \text{if } \gamma \text{ is underfitted.} \end{cases}$$

We prove this lemma in Appendix C in the Supplement.

Equipped with this lemma, a simple counting argument yields that under the event $\mathcal{A}_n \cap \mathcal{B}_n \cap \mathcal{C}_n \cap \mathcal{D}_n$,

$$\frac{\pi_n(\mathbb{M}_S|Y)}{\pi_n(\gamma^*|Y)} \overset{(i)}{\le} \sum_{k=1}^{\infty} p^k p^{-2k} + \sum_{\ell=0}^{\infty}\sum_{r=1}^{\infty} p^{l+r} p^{-2l-2r-2} \le 3p^{-1}$$

for $p \ge 2$, where in step *(i)*, we used the fact that there are at most p^k overfitted models γ with $|\gamma \setminus \gamma^*| = k$ and at most $p^\ell (s^*)^r \le p^{\ell+r}$ underfitted models γ with $|\gamma \setminus \gamma^*| = \ell$ and $|\gamma^* \setminus \gamma| = r$. Combining this with inequality (35), we obtain that with probability at least $1 - c_1 p^{-c_2}$,

$$\pi_n(\mathbb{M}_S|Y) \le 3p^{-1}\pi_n(\gamma^*|Y) \le 3p^{-1}. \tag{36}$$

Step 2. Let $\mathbb{M}_L := \{\gamma \in \{0,1\}^p : |\gamma| \ge K \max\{s^*, 1\} + 1\}$ denote the set of large models. By Bayes' theorem, we can express the posterior probability of $\mathbb{M}_L$ as

$$\pi_n(\mathbb{M}_L|Y) = \frac{\sum_{\gamma\in\mathbb{M}_L} \int_{\theta,\phi} d\mathbb{P}_{\beta,\phi,\gamma}/d\mathbb{P}_0(Y)\pi_n(d\theta, d\phi, \gamma)}{\sum_{\gamma\in\{0,1\}^p} \int_{\theta,\phi} d\mathbb{P}_{\beta,\phi,\gamma}/d\mathbb{P}_0(Y)\pi_n(d\theta, d\phi, \gamma)}, \tag{37}$$

where $\mathbb{P}_{\beta,\phi,\gamma}$ and $\mathbb{P}_0$ stand for the probability distribution of Y under parameters (β, ϕ, γ) and the true data generating model, respectively. We bound the numerator and denominator separately.

First consider the numerator. According to our specification of the sparsity prior (5d) for the binary indicator vector γ and Assumption C that $\kappa \ge 2$, the prior probability of $\mathbb{M}_L$ satisfies

$$\pi_n(\mathbb{M}_L) = \sum_{\gamma:|\gamma|\ge K\max\{1,s^*\}+1} \pi_n(\gamma) \le 2p^{-K\max\{1,s^*\}-1}.$$

By Fubini's theorem, we have the following bound for the expectation of the numerator:

$$\begin{aligned}
&\mathbb{E}_0\Big[\sum_{\gamma\in\mathbb{M}_L} \int_{\theta,\phi} \frac{d\mathbb{P}_{\beta,\phi,\gamma}}{d\mathbb{P}_0}(Y)\pi_n(d\theta, d\phi, \gamma)\Big] \\
&\quad = \sum_{\gamma\in\mathbb{M}_L} \int_{\theta,\phi} \mathbb{E}_0\Big[\frac{d\mathbb{P}_{\beta,\phi,\gamma}}{d\mathbb{P}_0}(Y)\Big]\pi_n(d\theta, d\phi, \gamma) \\
&\quad = \sum_{\gamma\in\mathbb{M}_L} \int_{\theta,\phi} \pi_n(d\theta, d\phi, \gamma) = \pi_n(\mathbb{M}_L) \le 2p^{-K\max\{1,s^*\}-1},
\end{aligned}$$

where we have used the fact that $\mathbb{E}_0[\frac{d\mathbb{P}_{\beta,\phi,\gamma}}{d\mathbb{P}_0}(Y)] = 1$. Therefore, by applying Markov's inequality we have

$$\begin{aligned}&\mathbb{P}_0\biggl[\sum_{\gamma\in\mathbb{M}_L}\int_{\theta,\phi}\frac{d\mathbb{P}_{\beta,\phi,\gamma}}{d\mathbb{P}_0}(Y)\pi_n(d\theta,d\phi,\gamma)\le 2p^{-K\max\{1,s^*\}/2-1}\biggr]\\ &\quad\ge 1-p^{-K\max\{1,s^*\}/2}.\end{aligned}\tag{38}$$

Using the expression for the marginal likelihood function [see equation (A.2) in the Supplement], we can bound the denominator from below by

$$\begin{aligned}&\int_{\theta,\phi}\frac{d\mathbb{P}_{\beta,\phi,\gamma^*}}{d\mathbb{P}_0}(Y)\pi_n\bigl(d\theta,d\phi,\gamma^*\bigr)\\ &\quad=\frac{\mathcal{L}_n(Y|\gamma^*)\pi_n(\gamma^*)}{d\mathbb{P}_0(Y)}\\ &\quad=\frac{\Gamma(n/2)(1+g)^{n/2}}{\pi^{n/2}}\frac{(1+g)^{-s^*/2}}{(\|Y\|_2^2+g\|(I-\Phi_{\gamma^*})\tilde{w}\|_2^2)^{n/2}}\cdot\frac{cp^{-\kappa s^*}}{d\mathbb{P}_0(Y)},\end{aligned}$$

where $\tilde{w} = w + X_{S^c}\beta^*_{S^c} \sim \mathcal{N}(X_{S^c}\beta^*_{S^c}, \sigma_0^2)$. Under the true data-generating model $\mathbb{P}_0$, the density for Y is $\sigma_0^{-n}(2\pi)^{-n/2}\exp\{-(2\sigma_0^2)^{-1}\|w\|_2^2\}$. By applying the lower bound $\Gamma(n/2) \ge (2\pi)^{1/2}(n/2-1)^{n/2-1/2}e^{-n/2+1}$ and using the fact that the projection operator $I - \Phi_{\gamma^*}$ is nonexpansive, we obtain

$$\begin{aligned}&\int_{\theta,\phi}\frac{d\mathbb{P}_{\beta,\phi,\gamma^*}}{d\mathbb{P}_0}(Y)\pi_n\bigl(d\theta,d\phi,\gamma^*\bigr)\\ &\quad\ge cp^{-\kappa s^*}(1+g)^{-s^*/2}\bigl(1+g^{-1}\bigr)^{n/2}\\ &\qquad\times\exp\bigl\{\bigl(2\sigma_0^2\bigr)^{-1}\bigl(\|w\|_2^2-\|\tilde{w}\|_2^2-\|Y\|_2^2/g\bigr)\bigr\}\underbrace{\bigl(u^{-n/2}e^{u/2}\bigr)}_{f(u)}\underbrace{\bigl(n^{n/2}e^{-n/2}\bigr)}_{1/f(n)},\end{aligned}$$

where $u = \sigma_0^{-2}(\|\tilde{w}\|_2^2 + \|Y\|_2^2/g)$. Since $g^{-1} \lesssim n^{-1}$ and the function $f(u) = u^{-n/2}e^{u/2}$ attains its minimum at $u = n$, we further obtain

$$\begin{aligned}&\int_{\theta,\phi}\frac{d\mathbb{P}_{\beta,\phi,\gamma^*}}{d\mathbb{P}_0}(Y)\pi_n\bigl(d\theta,d\phi,\gamma^*\bigr)\\ &\quad\ge cp^{-\kappa s^*}(1+g)^{-s^*/2}\exp\bigl\{\bigl(2\sigma_0^2\bigr)^{-1}\bigl(\|w\|_2^2-\|\tilde{w}\|_2^2-\|Y\|_2^2/g\bigr)\bigr\},\end{aligned}$$

with a different universal constant c.

The off-support S^c condition in Assumption A and the high probability bound for the event $\mathcal{C}_n \cap \mathcal{D}_n$ in Lemma 3 imply that the last exponential term is at least of order $p^{-(\tilde{L}+1)}$ with probability at least $1 - c_1p^{-c_2}$. Therefore, for $K \ge 2\kappa + \alpha + 2(\tilde{L}+1)$, we have

$$\int_{\theta,\phi}\frac{d\mathbb{P}_{\beta,\phi,\gamma^*}}{d\mathbb{P}_0}(Y)\pi_n\bigl(d\theta,d\phi,\gamma^*\bigr)\ge cp^{-K\max\{1,s^*\}/2}.\tag{39}$$

Combining equations (37), (38) and (39), we obtain that

$$\pi_n(\mathbb{M}_L|Y) \le cp^{-1}, \tag{40}$$

holds with probability at least $1 - c_1 p^{-c_2}$.

Finally, inequalities (36) and (40) in steps 1 and 2 together yield that

$$\pi_n(\gamma^*|Y) = 1 - \pi_n(\mathbb{M}_S|Y) - \pi_n(\mathbb{M}_L|Y) \ge 1 - c_3 p^{-1},$$

holds with probability at least $1 - c_1 p^{-c_2}$, which completes the proof.

4.5. *Proof of Corollary* 2. Let $\mathbb{P}_t$ denote the probability distribution of iterate γ_t in the MCMC algorithm. According to the definition of ε-mixing time, for any $t \ge \tau_{1/p}$, we are guaranteed that $|\mathbb{P}_t(\gamma^*) - \pi_n(\gamma^*)| \le \frac{1}{p}$. By Theorem 1, the posterior probability of γ^* satisfies $\pi_n(\gamma^*) \ge 1 - c_1 p^{-1}$ with probability at least $1 - c_2 p^{-c_3}$. By Theorem 2, the p^{-1}-mixing time $\tau_{1/p}$ satisfies

$$\tau_{1/p} \le 12 p s_0^2\big((\alpha n + \alpha s_0 + 2\kappa s_0)\log p + \log p + \log 2\big)$$

with probability at least $1 - c_3 p^{-c_4}$. Combining the three preceding displays, we find that $\mathbb{P}_t(\gamma^*) \ge 1 - (c_1 + 1)p^{-1}$, as claimed.

5. Discussion.

In this paper, we studied the computational complexity of MCMC methods for high-dimensional Bayesian linear regression under a sparsity constraint. We show that under a set of conditions that guarantees Bayesian variable-selection consistency, the corresponding MCMC algorithm achieves rapid mixing. Our result on the computational complexity of Bayesian variable-selection example provides insight into the dynamics of the Markov chain methods applied to statistical models with good asymptotic properties. It suggests that contraction properties of the posterior distribution are useful not only in guaranteeing desirable statistical properties such as parameter estimation or model selection consistency, but they also have algorithmic benefits in certifying the rapid mixing of the Markov chain methods designed to draw samples from the posterior.

As a future direction, it is interesting to investigate the mixing behavior of the MCMC algorithm when Bayesian variable selection fails. For example, both slow and fast mixing behavior are observed empirically in the intermediate SNR regime in our simulated examples and it would be interesting to understand this result theoretically. Another interesting direction is to consider the computational complexity of MCMC methods for models more complex than linear regression, for example, high-dimensional nonparametric additive regression. A third direction is to investigate whether the upper bound on mixing time provided in Theorem 2 is sharp up to constants.

SUPPLEMENTARY MATERIAL

Supplement to "On the computational complexity of high-dimensional Bayesian variable selection" (DOI: 10.1214/15-AOS1417SUPP; .pdf). Owing to space constraints, we have moved some materials and technical proofs to the Appendix, which is contained in the supplementary document.

REFERENCES

[1] AN, H., HUANG, D., YAO, Q. and ZHANG, C. (2008). Stepwise searching for feature variables in high-dimensional linear regression. Technical report, Dept. Statistics, London School of Economics.

[2] BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. MR2065192

[3] BELLONI, A. and CHERNOZHUKOV, V. (2009). On the computational complexity of MCMC-based estimators in large samples. *Ann. Statist.* **37** 2011–2055. MR2533478

[4] BHATTACHARYA, A., PATI, D., PILLAI, N. S. and DUNSON, D. B. (2015). Dirichlet–Laplace Priors for Optimal Shrinkage. *J. Amer. Statist. Assoc.* **110** 1479–1490. MR3449048

[5] BORGS, C., CHAYES, J. T., FRIEZE, A., KIM, J. H., TETALI, P., VIGODA, E. and VU, V. H. (1999). Torpid mixing of some Monte Carlo Markov chain algorithms in statistical physics. In 40*th Annual Symposium on Foundations of Computer Science* (*New York*, 1999) 218–229. IEEE Computer Soc., Los Alamitos, CA. MR1917562

[6] CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. MR3375874

[7] DIACONIS, P. and STROOCK, D. (1991). Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Probab.* **1** 36–61. MR1097463

[8] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[9] FERNÁNDEZ, C., LEY, E. and STEEL, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *J. Econometrics* **100** 381–427. MR1820410

[10] GELMAN, A. and RUBIN, D. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

[11] GEORGE, E. and MCCULLOCH, R. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

[12] GUAN, Y. and STEPHENS, M. (2011). Bayesian variable selection regression for Genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* **5** 1780–1815. MR2884922

[13] HANS, C., DOBRA, A. and WEST, M. (2007). Shotgun stochastic search for "large p" regression. *J. Amer. Statist. Assoc.* **102** 507–516. MR2370849

[14] HORN, R. A. and JOHNSON, C. R. (1985). *Matrix Analysis*. Cambridge Univ. Press, Cambridge. MR0832183

[15] ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158

[16] JONES, G. L. and HOBERT, J. P. (2004). Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. *Ann. Statist.* **32** 784–817. MR2060178

[17] KASS, R. E. and WASSERMAN, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.* **90** 928–934. MR1354008

[18] LAURENT, B. and MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** 1302–1338. MR1805785

[19] LEDOUX, M. (2001). *The Concentration of Measure Phenomenon. Mathematical Surveys and Monographs* **89**. Amer. Math. Soc., Providence, RI. MR1849347

[20] LEVIN, D. A., LUCZAK, M. J. and PERES, Y. (2010). Glauber dynamics for the mean-field Ising model: Cut-off, critical power law, and metastability. *Probab. Theory Related Fields* **146** 223–265. MR2550363

[21] LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. and BERGER, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103** 410–423. MR2420243

[22] MARTINELLI, F. and SINCLAIR, A. (2012). Mixing time for the solid-on-solid model. *Ann. Appl. Probab.* **22** 1136–1166. MR2977988

[23] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[24] MOSSEL, E. and VIGODA, E. (2006). Limitations of Markov chain Monte Carlo algorithms for Bayesian inference of phylogeny. *Ann. Appl. Probab.* **16** 2215–2234. MR2288719

[25] NARISETTY, N. N. and HE, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.* **42** 789–817. MR3210987

[26] SCHRECK, A., FORT, G., CORFF, S. L. and MOULINES, E. (2015). A shrinkage-thresholding Metropolis adjusted Langevin algorithm for Bayesian variable selection. Available at arXiv:1312.5658.

[27] SHANG, Z. and CLAYTON, M. K. (2011). Consistency of Bayesian linear model selection with a growing number of parameters. *J. Statist. Plann. Inference* **141** 3463–3474. MR2817355

[28] SHEN, X., PAN, W. and ZHU, Y. (2012). Likelihood-based selection and sharp parameter estimation. *J. Amer. Statist. Assoc.* **107** 223–232. MR2949354

[29] SINCLAIR, A. (1988). Algorithms for random generation and counting: A Markov chain approach. Ph.D. thesis, Univ. Edinburgh.

[30] SINCLAIR, A. (1992). Improved bounds for mixing rates of Markov chains and multicommodity flow. *Combin. Probab. Comput.* **1** 351–370. MR1211324

[31] SPARKS, D., KHARE, K. and GHOSH, M. (2015). Necessary and sufficient conditions for high-dimensional posterior consistency under g-priors. *Bayesian Anal.* **10** 627–664.

[32] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[33] WAINWRIGHT, M. J. (2009). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory* **55** 5728–5741. MR2597190

[34] WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873

[35] WOODARD, D. B. and ROSENTHAL, J. S. (2013). Convergence rate of Markov chain methods for genomic motif discovery. *Ann. Statist.* **41** 91–124. MR3059411

[36] YANG, Y., WAINWRIGHT, M. J. and JORDAN, M. I. (2016). Supplement to "On the computational complexity of high-dimensional Bayesian variable selection." DOI:10.1214/15-AOS1417SUPP.

[37] ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques* (P. K. Goel and A. Zellner, eds.). *Stud. Bayesian Econometrics Statist.* **6** 233–243. North-Holland, Amsterdam. MR0881437

[38] ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

[39] ZHANG, T. (2011). Adaptive forward–backward greedy algorithm for learning sparse representations. *IEEE Trans. Inform. Theory* **57** 4689–4708. MR2840485

[40] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

Y. YANG
DEPARTMENT OF EECS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: yy84@berkeley.edu

M. J. WAINWRIGHT
M. I. JORDAN
DEPARTMENT OF EECS AND STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: wainwrig@berkeley.edu
jordan@stat.berkeley.edu