SIAM J. MATH. DATA SCI.
Vol. 3, No. 3, pp. 854–872

Approximate Spectral Gaps for Markov Chain Mixing Times in High Dimensions*

Yves F. Atchadé†

Abstract. This paper introduces a concept of approximate spectral gap to analyze the mixing time of reversible Markov chain Monte Carlo (MCMC) algorithms for which the usual spectral gap is degenerate or almost degenerate. We use the idea to analyze an MCMC algorithm to sample from mixtures of densities. As an application we study the mixing time of a Gibbs sampler for variable selection in linear regression models. We show that, properly tuned, the algorithm has a mixing time that grows at most polynomially with the dimension. Our results also suggest that the mixing time improves when the posterior distribution contracts towards the true model and the initial distribution is well chosen.

Key words. Markov chain Monte Carlo algorithms, Markov chain mixing times, spectral gap, MCMC for mixtures of densities, high-dimensional linear regression models

AMS subject classifications. 60J05, 65C05, 65C60

DOI. 10.1137/19M1283082

1. Introduction.

Understanding the type of problems for which fast Markov chain Monte Carlo (MCMC) sampling is possible is a question of fundamental interest. The study of the size of the spectral gap is a widely used approach for that purpose. However, the technique can be too coarse when dealing with distributions with small isolated local modes. To be more precise, let π be some probability measure of interest on some measurable space $\mathcal{X}$, and let K be a Markov kernel with invariant distribution π. For the purpose of sampling from π using K, one can represent an isolated local mode as a subset A such that $K(x, \mathcal{X} \setminus A)$ is small compared to $\pi(\mathcal{X} \setminus A)$ for all $x \in A$. In this case, K will have a small conductance, and a small spectral gap. Note, however, that if $\pi(A)$ is also small (that is, we are dealing with a small isolated mode A), then since

$$\int_{\mathcal{X}\setminus A} \pi(\mathrm{d}x)K(x, A) = \int_A \pi(\mathrm{d}x)K(x, \mathcal{X} \setminus A),$$

we see that the set A will be typically hard to reach in the first place. Hence, any finite-length Markov chain $\{X_0, \ldots, X_n\}$, say, with transition kernel K and initialized in $\mathcal{X} \setminus A$ is unlikely to visit A. But even when A is never visited, and since $\pi(A)$ is small, X_n may still be a good approximate sample from π for large n. This implies that the poor mixing time predicted by the standard spectral gap may markedly differ from the actual behavior of these finite-length chains. Motivated by this problem, and building upon the s-conductance of Lovász

*Received by the editors August 23, 2019; accepted for publication (in revised form) June 3, 2021; published electronically August 24, 2021.
https://doi.org/10.1137/19M1283082
Funding: This work is partially supported by NSF grant DMS-1513040.
†Department of Mathematics and Statistics, Boston University, Boston, MA 02215 USA (atchade@bu.edu).

and Simonovits [14], we develop an idea of approximate spectral gap (that we call a ζ-spectral gap for some $\zeta \in [0,1)$) which allows us to measure the mixing time of a Markov chain while discounting the ill effects of overly small sets.

Mixtures are good examples of probability distributions with isolated local modes. We use the idea to analyze a class of MCMC algorithms to sample from mixtures of densities. Much is known on the computational complexity of various MCMC algorithms for log-concave densities (see, e.g., [14, 8, 13, 15], and [5] and the references therein). However, these results cannot be directly applied to mixtures, since a mixture of log-concave densities is not log-concave in general. By augmenting the variable of interest to include the mixing variable, a Gibbs sampler can be used to sample from a mixture. A very nice lower bound on the spectral gap of such Gibbs samplers is developed in [16]. We reexamine the argument in [16] using the ζ-spectral gap concept, leading to Theorem 3.1, which gives potentially better dependence on the dimension.

Our initial motivation into this work is in large-scale Bayesian variable selection problems. The Bayesian posterior distributions that arise from these problems are typically mixtures of log-concave densities with very large numbers of components, and the aforementioned Gibbs sampler is commonly used for sampling (see, e.g., [10, 21]). We analyze the algorithm in a regime where the posterior distribution is known to have good contraction properties. In that regime we show that the algorithm has a mixing time that grows exponentially fast with the sample size (Theorem 4.3). However, using the approximate spectral gap we also show that with a good initial distribution (warm-start) the mixing time of the algorithm grows only polynomially with the number of regressors. The power of the polynomial function depends mainly on the coherence and the eigenstructure of the regressors (Theorem 4.4).

The paper is organized as follows. We develop the concept of ζ-spectral gap in section 2. The main result there is Lemma 2.1. In section 3 we study the mixing time of mixtures of Markov kernels and derive (Theorem 3.1) a generalization of Theorem 1.2 of [16]. We put these two results together to analysis the linear regression model in section 4. The proofs of these results can be found in the accompanying supplementary material (supplmaterial.pdf [local/web 395KB]). Some numerical simulations are detailed in section 4.1.

2. Approximate spectral gaps for Markov chains. Let π be a probability measure on some measurable space $\mathcal{X}$ with sigma-algebra $\mathcal{B}$. For a function $f : \mathcal{X} \to \mathbb{R}$, we write $f \in \mathcal{B}$ to say that f is $\mathcal{B}$-measurable. We let $L^2(\pi)$ denote the Hilbert space of all real-valued square-integrable (with respect to π) functions on $\mathcal{X}$, equipped with the inner product $\langle f, g\rangle_\pi \stackrel{\text{def}}{=} \int_{\mathcal{X}} f(x)g(x)\pi(\mathrm{d}x)$ with associated norm $\|\cdot\|_2$. We will also make use of the essential supremum of f with respect to π defined as $\|f\|_\infty \stackrel{\text{def}}{=} \inf\{M \geq 0 : \pi(\{x \in \mathcal{X} : |f(x)| > M\}) = 0\}$. If P is a Markov kernel on $\mathcal{X}$, and $n \geq 1$ an integer, P^n denotes the nth iterate of P, defined recursively as $P^n(x, A) \stackrel{\text{def}}{=} \int_{\mathcal{X}} P^{n-1}(x, \mathrm{d}z)P(z, A)$, $x \in \mathcal{X}$, A measurable. For $f \in \mathcal{B}$, we define $Pf : \mathcal{X} \to \mathbb{R}$ as $Pf(x) \stackrel{\text{def}}{=} \int_{\mathcal{X}} P(x, \mathrm{d}z)f(z)$, $x \in \mathcal{X}$, whenever the integral is well defined. And if μ is a probability measure on $\mathcal{X}$, then μP is the probability on $\mathcal{X}$ defined as $\mu P(A) \stackrel{\text{def}}{=} \int_{\mathcal{X}} \mu(\mathrm{d}z)P(z, A)$, $A \in \mathcal{B}$. The total variation distance between two probability measures μ, ν is defined as

$$\|\mu - \nu\|_{\text{tv}} \stackrel{\text{def}}{=} 2 \sup_{A \in \mathcal{B}} (\mu(A) - \nu(A)) = \sup_{f \in \mathcal{B}:\, \|f\|_\infty \leq 1} \left(\int_{\mathcal{X}} f(x)\mu(\mathrm{d}x) - \int_{\mathcal{X}} f(x)\nu(\mathrm{d}x) \right).$$

Let K be a Markov kernel on $\mathcal{X}$ with invariant distribution π. Without changing notation we will view K as the linear operator on $L^2(\pi)$ that transforms f into Kf as defined above. We write $K^\star$ to denote the adjoint of K, that is, the linear operator on $L^2(\pi)$ such that $\langle Kf, g\rangle_\pi = \langle f, K^\star g\rangle_\pi$ for all $f, g \in L^2(\pi)$. We say that K is reversible with respect to π (π-reversible, for short) if $K = K^\star$, and we say that K is positive if it is π-reversible and $\langle f, Kf\rangle_\pi \geq 0$ for all $f \in L^2(\pi)$. Note that the operators $K^\star K$ and $KK^\star$ are always positive, since $\langle f, K^\star Kf\rangle_\pi = \langle Kf, Kf\rangle_\pi = \|Kf\|_2^2 \geq 0$, and similarly for $KK^\star$. The concept of spectral gap and the related Poincaré inequalities are commonly used to quantify Markov chain mixing times. For $f \in L^2(\pi)$, we set $\pi(f) \stackrel{\text{def}}{=} \int_{\mathcal{X}} f(x)\pi(\mathrm{d}x)$, $\mathsf{Var}_\pi(f) \stackrel{\text{def}}{=} \|f - \pi(f)\|_2^2$, and

$$\mathcal{E}_K(f,f) \stackrel{\text{def}}{=} \frac{1}{2}\int\int (f(y)-f(x))^2\pi(\mathrm{d}x)K(x,\mathrm{d}y) = \langle f, f\rangle_\pi - \langle f, Kf\rangle_\pi .$$

The spectral gap of K is then defined as

$$\begin{aligned}\lambda(K) &\stackrel{\text{def}}{=} \inf\left\{\frac{\mathcal{E}_K(f,f)}{\mathsf{Var}_\pi(f)},\ f \in L^2(\pi),\ \text{ s.t. }\ \mathsf{Var}_\pi(f) > 0\right\},\\ &= \inf\left\{1 - \langle f, Kf\rangle_\pi,\ f \in L^2(\pi),\ \pi(f) = 0,\ \pi(f^2) = 1\right\}.\end{aligned}$$

It is well known (see, for instance, [20, Corollary 2.14]) that if $\pi_0(\mathrm{d}x) = f_0(x)\pi(\mathrm{d}x)$ and $f_0 \in L^2(\pi)$, then

$$\|\pi_0 K^n - \pi\|_{\text{tv}} \leq \sqrt{\mathsf{Var}_\pi(f_0)}\,(1-\lambda(K^\star K))^{n/2}. \tag{2.1}$$

This result can also be derived from Lemma 2.1 below with $\zeta = 0$ (see 2.7). It follows from (2.1) that lower bounds on the spectral gap of $K^\star K$ can be used to derive upper bounds on the mixing time of K. Note from the definition that

$$\lambda(K^\star K) = 1 - \sup\left\{\|Kf\|_2^2,\ f \in L^2(\pi),\ \pi(f) = 0,\ \pi(f^2) = 1\right\} = 1 - \|K_0\|^2,$$

where K_0 denotes the restriction of K to $\{f \in L^2(\pi) :\ \pi(f) = 0\}$, and $\|K_0\|$ denotes its operator norm. Hence (2.1) can be also written as

$$\|\pi_0 K^n - \pi\|_{\text{tv}} \leq \sqrt{\mathsf{Var}_\pi(f_0)}\|K_0\|^n. \tag{2.2}$$

So far we have not made any assumption on K besides that it has invariant distribution π. If we assume that K is positive, then it is well known that $\|K_0\| = \sup\{\langle f, Kf\rangle_\pi : f \in L^2(\pi),\ \pi(f) = 0,\ \pi(f^2) = 1\}$. Hence in this case we have $\|K_0\| = 1-\lambda(K)$, and (2.2) becomes

$$\|\pi_0 K^n - \pi\|_{\text{tv}} \leq \sqrt{\mathsf{Var}_\pi(f_0)}\,(1-\lambda(K))^n. \tag{2.3}$$

In some problems the conductance of K is easier to control than the spectral gap. An interesting generalization of the conductance introduced by Lovász and Simonovits [14] (which we shall call here ζ-conductance) has proven very useful in problems where a warm-start to the Markov chain is available. For $\zeta \in [0, 1/2)$, we define the ζ-conductance of the Markov kernel K as

$$\Phi_\zeta(K) \stackrel{\text{def}}{=} \inf\left\{\frac{\int_A \pi(\mathrm{d}x)K(x,A^c)}{(\pi(A)-\zeta)(\pi(A^c)-\zeta)},\ \ \zeta < \pi(A) < \frac{1}{2}\right\},$$

where the infimum above is taken over measurable subsets of $\mathcal{X}$. Note that $\Phi_0(K)$ is the standard conductance. $\Phi_\zeta(K)$ captures the same concept of ergodic flow as $\Phi_0(K)$, except that in $\Phi_\zeta(K)$ we disregard sets that are either too small or too large under π. It turns out that $\Phi_\zeta(K)$ still controls the mixing time of K up to an additive constant that depends on ζ (see [14, Corollary 1.5]). There are many problems where a direct bound on the spectral gap instead of the conductance is easier, or yields better results. Motivated by the ζ-conductance, we introduce a concept similar to ζ-spectral gap that directly approximates the spectral gap. For $\zeta \in [0,1)$, we define the ζ-spectral gap of K as

$$\lambda_\zeta(K) \stackrel{\text{def}}{=} \inf\left\{ \frac{\mathcal{E}_K(f,f)}{\mathsf{Var}_\pi(f) - \frac{\zeta}{2}},\ \ f \in L^2(\pi),\ \ \text{s.t.}\ \ \mathsf{Var}_\pi(f) > \zeta,\ \text{and}\ \|f\|_\infty = 1 \right\}. \tag{2.4}$$

The definition can be adapted to norms other than the uniform norm. We focus on the uniform norm for convenience in the applications. We note that when K is positive, $\lambda_\zeta(K)$ is always in the interval $[0,2]$. In particular for any Markov kernel K, we always have $\lambda_\zeta(K^\star K) \in [0,2]$. To see this, given $f \in L^2(\pi)$, such that $\|f\|_\infty = 1$, and $\mathsf{Var}_\pi(f) > \zeta$, and writing $\bar{f} = f - \pi(f)$ so that $\mathsf{Var}_\pi(f) = \pi(\bar{f}^2)$, we have

$$\frac{\mathcal{E}_K(f,f)}{\mathsf{Var}_\pi(f) - \frac{\zeta}{2}} = \frac{\pi(\bar{f}^2) - \langle \bar{f}, K\bar{f}\rangle_\pi}{\pi(\bar{f}^2) - \frac{\zeta}{2}} \le \frac{\pi(\bar{f}^2)}{\pi(\bar{f}^2) - \frac{\zeta}{2}} < 2,$$

where the first inequality uses the positivity of K. The next proposition shows that the ζ-spectral gap can be used to bound convergence to stationarity.

Lemma 2.1. *Fix $\zeta \in [0,1)$. For every integer $N \ge 1$, and $f \in L^2(\pi)$, we have*

$$\mathsf{Var}_\pi(K^N f) \le \mathsf{Var}_\pi(f)\left(1 - \min(1, \lambda_\zeta(K^\star K))\right)^N + 2\zeta\|f\|_\infty^2. \tag{2.5}$$

If K is positive, then we have

$$\mathsf{Var}_\pi(K^N f) \le \mathsf{Var}_\pi(f)\left(1 - \min(1, \lambda_\zeta(K))\right)^N + 2\zeta\|f\|_\infty^2. \tag{2.6}$$

Proof. See section 5. ∎

Remark 2.2. The idea of approximate spectral gap developed here is somewhat similar to the concept of weak Poincaré inequality developed for continuous-time Markov semigroups with zero spectral gap ([12, 3]). The main difference is that weak Poincaré inequalities lead to subgeometric rates of convergence of the semigroup, whereas the idea of ζ-spectral gap as introduced here leads to a geometric convergence rate, plus an additive remainder that depends on ζ.

We conjecture[1] that for a positive kernel K, it holds that

$$1 - \min(1, \lambda_\zeta(K^\star K)) \le \left(1 - \min(1, \lambda_\zeta(K))\right)^2.$$

[1] I'm grateful to Daniel Rudolf for stimulating discussions on this problem.

This result would match the known behavior of the standard spectral gap ($\zeta = 0$) and would improve the power on the right-hand side of (2.6) from N to $2N$. We note that

$$1 - \min(1, \lambda_\zeta(K^\star K)) \le \sup\left\{\|Kf\|_2^2,\ f \in L^2(\pi), \pi(f) = 0, \pi(f^2) \in [1,2],\ \|f\|_\infty \le 2\sqrt{\frac{2}{\zeta}}\right\},$$

but relating the right-hand side of the last display back to $\lambda_\zeta(K)$ when K is positive has proven difficult.

Let ν be some arbitrary initial distribution on $\mathcal{X}$ that is absolutely continuous with respect to π with density, say, f_0. Then, noting that for any $f \in L^2(\pi)$,

$$\left|\nu K^N(f) - \pi(f)\right| = \left|\int_{\mathcal{X}} f_0(x)\left(K^N f(x) - \pi(f)\right)\pi(\mathrm{d}x)\right| \le \|f_0\|_2\sqrt{\mathsf{Var}_\pi(K^N f)},$$

we deduce from Lemma 2.1 that

$$\sup_{f\in\mathcal{B}:\ \|f\|_\infty\le 1}\left|\nu K^N(f) - \pi(f)\right| \le \|f_0\|_2\left(1 - \min(1, \lambda_\zeta(K^\star K))\right)^{N/2} + \|f_0\|_2\sqrt{2\zeta}, \tag{2.7}$$

which gives a bound on the convergence to stationarity of K in total variation. If K is positive, we can replace $\lambda_\zeta(K^\star K)$ by $\lambda_\zeta(K)$ in (2.7). Equation (2.7) captures the main idea of the paper: when a warm-start is available (that is, the term $\|f_0\|_2\sqrt{2\zeta}$ is small), the mixing time of K is well captured by $\lambda_\zeta(K^\star K)$, which can behave better than $\lambda(K^\star K)$—particularly in high-dimensional problems with small isolated modes.

2.1. Illustration with the small local mode example and further intuition. As in the introduction, suppose that $\mathcal{X} = \mathcal{X}_0 \cup (\mathcal{X}_0^c)$ for some measurable subset $\mathcal{X}_0$ of $\mathcal{X}$, and $\pi(\mathcal{X}_0^c)$ is small. The approximate spectral gap can be used to show that if the restriction of K to $\mathcal{X}_0$ is fast mixing and $\pi(\mathcal{X}_0^c)$ is small, then the Markov chain warm-started in $\mathcal{X}_0$ is also fast mixing. Let $K_{\mathcal{X}_0}$ be the restriction of K on $\mathcal{X}_0$ defined as

$$K_{\mathcal{X}_0}(x, \mathrm{d}y) = K(x, \mathrm{d}y) + \delta_x(\mathrm{d}y)K(x, \mathcal{X}_0^c),\quad x \in \mathcal{X}_0.$$

It is easy to show that the invariant distribution of $K_{\mathcal{X}_0}$ is $\pi_{\mathcal{X}_0}$, the restriction of π to $\mathcal{X}_0$, and the spectral gap of $K_{\mathcal{X}_0}$ is given by

$$\lambda_{\mathcal{X}_0}(K) \stackrel{\mathrm{def}}{=} \inf\left\{\frac{\frac{1}{2}\int_{\mathcal{X}_0}\int_{\mathcal{X}_0}\pi(\mathrm{d}x)K(x,\mathrm{d}y)(f(y)-f(x))^2}{\frac{1}{2}\int_{\mathcal{X}_0}\int_{\mathcal{X}_0}\pi(\mathrm{d}x)\pi(\mathrm{d}y)(f(y)-f(x))^2},\ f:\ \mathcal{X}\to\mathbb{R}\right\}, \tag{2.8}$$

where the infimum is taken over all functions for which the denominator is positive. The next result shows that the spectral gap of $K_{\mathcal{X}_0}$ is a lower bound for $\lambda_\zeta(K)$.

Lemma 2.3. *For $\zeta \in [0,1)$, if $\pi(\mathcal{X}_0) \ge 1 - \zeta/8$, then $\lambda_\zeta(K) \ge \lambda_{\mathcal{X}_0}(K)$.*

Proof. See section 6. ∎

Suppose that the initial distribution $\pi_0(\mathrm{d}x) = f_0(x)\pi(\mathrm{d}x)$ is such that $\|f_0\|_\infty \le B$ for some constant $B \ge 1$. In that case (2.7) gives, for all $n \ge 1$,

$$\|\pi_0 K^n - \pi\|_{\mathrm{tv}} \le B\left(\left(1 - \min(1, \lambda_\zeta(K^\star K))\right)^{N/2} + \sqrt{2\zeta}\right).$$

Fix $\zeta_0 \in (0,1)$, and take $\zeta = \zeta_0^2/(2B^2)$. Therefore, if $\pi(\mathcal{X}_0) \geq 1 - \zeta/8 = 1 - \zeta_0^2/(16B^2)$, by Lemma 2.3 we obtain the following bound on the mixing time of K using the spectral gap of $(K^\star K)_{\mathcal{X}_0}$:

$$\|\pi_0 K^N - \pi\|_{\mathrm{tv}} \leq 2\zeta_0 \quad \text{for all} \quad N \geq \frac{\log\left(\frac{B^2}{\zeta_0^2}\right)}{\lambda_{\mathcal{X}_0}(K^\star K)}. \tag{2.9}$$

The condition $\pi(\mathcal{X}_0) \geq 1 - \left(\frac{\zeta_0}{4B}\right)^2$ puts a constraint on the initial distribution π_0 and on the concentration properties of π on $\mathcal{X}_0$. The successful use of the technique typically hinges on controlling these two aspects.

Another possible approach to bounding the mixing time of K using information from the restricted kernel is to bound directly (for instance, using coupling) the total variation distance between $\pi_0 K^N$ and $\pi_0 (K_{\mathcal{X}_0})^N$. This has been explored in the literature [2, 6, 18], and more systematically by [24]. This approach typically works well when K has well-understood drift conditions. Another more classical approach to relating the mixing times of the restricted and unrestricted chains is via state decomposition theorems [17, 16, 11, 23, 7]. However, this involves the so-called projection chain that describes jumps between $\mathcal{X}_0$ and $\mathcal{X}_0^c$, which in the current setting will result in rather pessimistic bounds.

3. Application: Mixing times of mixtures of Markov kernels. To illustrate Lemma 2.1 we consider here the case where $\mathcal{X} = \mathbb{R}^p$ (equipped with its Borel σ-algebra $\mathcal{B}$ and its Lebesgue measure, which we write as $\mathrm{d}x$), and π is a discrete mixture of log-concave densities of the form

$$\pi(\mathrm{d}x) \propto \sum_{i \in \mathsf{I}} \pi(i,x)\mathrm{d}x, \tag{3.1}$$

where I is a nonempty countable set, and for each $i \in \mathsf{I}$, $\pi(i,\cdot) : \mathbb{R}^p \to [0,\infty)$ is a measurable and integrable function. Sampling from mixtures is more challenging than sampling from log-concave densities ([9]). A common strategy is to work with the joint distribution on $\mathsf{I} \times \mathcal{X}$ defined as

$$\bar{\pi}(D \times B) = \frac{\sum_{i \in D} \int_B \pi(i,x)\mathrm{d}x}{\sum_{i \in \mathsf{I}} \int_{\mathcal{X}} \pi(i,x)\mathrm{d}x}, \quad D \subseteq \mathsf{I},\ B \in \mathcal{B}. \tag{3.2}$$

Let $\pi(i|x) \propto \pi(i,x)$ (resp., $\pi(i) \propto \int_{\mathcal{X}} \pi(i,x)\mathrm{d}x$) denote the implied conditional (resp., marginal) distribution on I, and let $\pi_i(\mathrm{d}x) \propto \pi(i,x)\mathrm{d}x$ be the implied conditional distribution on $\mathcal{X}$. For each $i \in \mathsf{I}$, let K_i be a transition kernel on $\mathcal{X}$ with invariant distribution π_i. We then consider the Markov kernel K defined as

$$K(x,\mathrm{d}y) \stackrel{\mathrm{def}}{=} \sum_{i \in \mathsf{I}} \pi(i|x) K_i(x,\mathrm{d}y). \tag{3.3}$$

It is easy to check that for all $f, g \in L^2(\pi)$, it holds that

$$\langle f, Kg \rangle_\pi = \sum_{i \in \mathsf{I}} \pi(i) \langle f, K_i g \rangle_{\pi_i}.$$

This shows that if, for each $i \in \mathsf{I}$, K_i is π_i-reversible (resp., positive), then K is π-reversible (resp., positive). In particular if each K_i is an exact draw from π_i, then K is positive. In [16] the authors developed a very nice lower bound on the spectral gap of K knowing the spectral gaps of the K_i's. Their result goes as follows. Suppose that there exist $\kappa > 0$ and a connected graph on I such that whenever there is an edge between $i, j \in \mathsf{I}$, it holds that

$$\int_{\mathcal{X}} \min\left(\pi_i(x), \pi_j(x)\right) \mathrm{d}x \geq \kappa. \tag{3.4}$$

If $D(\mathsf{I})$ denotes the diameter of the graph thus defined, Theorem 1.2 of [16] says that

$$\lambda(K) \geq \frac{\kappa}{2D(\mathsf{I})} \min_{i \in \mathsf{I}} \left\{\pi(i)\lambda(K_i)\right\}. \tag{3.5}$$

The lower bound in (3.5) can be very small when some $\pi(i)$ are small, or when the overlap parameter κ is small (which corresponds to the existence of isolated local modes). We combine the approach in [16] with the canonical path argument of [22, 4] to develop a new bound on the ζ-spectral gap of K. We make the following assumption.

H1. *There exist $\mathsf{I}_0 \subseteq \mathsf{I}$, and $\{\mathsf{B}_i,\ i \in \mathsf{I}_0\}$ a family of nonempty measurable subsets of $\mathcal{X}$, with the following properties:*

1. *For each $i \in \mathsf{I}_0$, $\pi_i(\mathsf{B}_i) \geq 1/2$.*
2. *There exist $\kappa > 0$ and $\mathcal{G} \subset \{(i,j) \in \mathsf{I}_0 \times \mathsf{I}_0 :\ i \neq j\}$ such that*

$$\int_{\mathsf{B}_i \cap \mathsf{B}_j} \min\left(\frac{\pi_i(x)}{\pi_i(\mathsf{B}_i)}, \frac{\pi_j(x)}{\pi_j(\mathsf{B}_j)}\right) \mathrm{d}x \geq \kappa \tag{3.6}$$

   whenever $(i,j) \in \mathcal{G}$.

3. *For each distinct pair $i, j \in \mathsf{I}_0$, there exists a path $\gamma_{ij} = (i_0, \ldots, i_\ell)$ where each pair (i_{k-1}, i_k) belongs to $\mathcal{G}$, such that $i_0 = i$ and $i_\ell = j$. Furthermore we assume that an edge can appear at most once on a given path. Let $\Gamma \stackrel{\text{def}}{=} \{\gamma_{ij},\ (i,j) \in \mathsf{I}_0 \times \mathsf{I}_0, i \neq j\}$.*

One should view $\cup_{i \in \mathsf{I}_0}\{i\} \times \mathsf{B}_i$ as a subset of $\mathsf{I} \times \mathcal{X}$ that captures most of the probability mass of the joint distribution $\bar{\pi}$. We stress that we do not assume the sets $\{\mathsf{B}_i,\ i \in \mathsf{I}_0\}$ to be known, only that they exist. In the case where $\bar{\pi}$ is a posterior distribution from some Bayesian inference problems, such existence results, known as posterior contraction, can often be obtained under further statistical assumptions.

The graph $\mathcal{G}$ captures the proximity between the conditional distributions, and the total variation distance between adjacent conditional distributions (as captured by κ) is the key parameter that determines the mixing of the kernel K. Indeed, (3.6) implies that the total variation distance between the restriction of π_i to B_i and the restriction of π_j to B_j is at most $2(1-\kappa)$.

For $\gamma \in \Gamma$, let $|\gamma|$ be the number of edges on γ. We define

$$\mathsf{m} \stackrel{\text{def}}{=} \max_{\iota \in \mathsf{I}_0} \sum_{\gamma_{ij} \in \Gamma:\ \gamma_{ij} \ni \iota} |\gamma_{ij}| \frac{\pi(i)\pi(j)}{\pi(\iota)}, \tag{3.7}$$

where the summation is taken over all canonical paths γ_{ij} that go through node ι. For $i \in \mathsf{I}_0$ and a kernel P on $\mathcal{X}$ we also define

$$\lambda_i(P) \stackrel{\text{def}}{=} \inf \left\{ \frac{\frac{1}{2}\int_{\mathsf{B}_i}\int_{\mathsf{B}_i} \pi(\mathrm{d}x)P(x,\mathrm{d}y)(f(y)-f(x))^2}{\frac{1}{2}\int_{\mathsf{B}_i}\int_{\mathsf{B}_i} \pi(\mathrm{d}x)\pi(\mathrm{d}y)(f(y)-f(x))^2},\ f:\ \mathcal{X}\to\mathbb{R} \right\}, \tag{3.8}$$

where the infimum is taken over all bounded measurable functions f for which the denominator is positive.

Theorem 3.1. *Let π be as in (3.1), and K as in (3.3). Assume that H1 holds. Set $\bar{\mathsf{B}} \stackrel{\text{def}}{=} \cup_{i\in\mathsf{I}_0}\{i\}\times\mathsf{B}_i$. If for some $\zeta \in [0,1)$ it holds that*

$$\bar{\pi}(\bar{\mathsf{B}}) \geq 1 - \frac{\zeta}{10}, \tag{3.9}$$

then

$$\lambda_\zeta(K) \geq \left(\frac{\kappa}{1+8\mathsf{m}}\right) \min_{i\in\mathsf{I}_0} \lambda_i(K_i). \tag{3.10}$$

Proof. See section 7. ∎

Note that the constant m satisfies

$$\mathsf{m} \leq \frac{\mathsf{D}(\mathsf{I}_0)}{\min_{i\in\mathsf{I}_0}\pi(i)}. \tag{3.11}$$

Hence the bound in (3.10) improves upon (3.5), even when $\zeta = 0$. In problems where an exact draw from $\pi(\cdot|x)$ is not available, the kernel K in (3.3) is not usable. In these cases it is typical to replace those exact draws by MCMC. Extending Theorem 3.1 to such settings is an important problem that we leave for future research.

4. Example: Analysis of a Gibbs sampler. We consider the Bayesian treatment of a linear regression problem with response variable $z \in \mathbb{R}^n$, and covariate matrix $X \in \mathbb{R}^{n\times p}$, with a spike-and-slab prior distribution on the regression parameter $\theta \in \mathbb{R}^p$ as in [10, 21]. More precisely, for some variable selection parameter $\delta \in \Delta \stackrel{\text{def}}{=} \{0,1\}^p$ and positive parameters ρ_0, ρ_1, we assume that the components of θ are conditionally independent, and $\theta_j|\{\delta = 1\}$ has density $\mathbf{N}(0,\rho_1^{-1})$, and $\theta_j|\{\delta = 0\}$ has density $\mathbf{N}(0,\rho_0^{-1})$, where $\mathbf{N}(\mu, v^2)$ denotes the univariate Gaussian distribution with mean μ and variance v^2. We further assume that given $\mathsf{q} \in (0,1)$, the prior distribution of δ is a product of Bernoulli with success probability q, and restricted to be in $\Delta_s \stackrel{\text{def}}{=} \{\delta \in \Delta :\ \|\delta\|_0 \leq s\}$ for some sparsity level s specified by the user. The resulting posterior distribution on $\Delta \times \mathbb{R}^p$ is

$$\Pi(\delta, \mathrm{d}\theta|z) \propto \left(\frac{\mathsf{q}}{1-\mathsf{q}}\right)^{\|\delta\|_0} \mathbf{1}_{\Delta_s}(\delta) \frac{e^{-\frac{1}{2}\theta' D_{(\delta)}^{-1}\theta}}{\sqrt{\det\left(2\pi D_{(\delta)}\right)}} e^{-\frac{1}{2\sigma^2}\|z-X\theta\|_2^2} \mathrm{d}\theta, \tag{4.1}$$

where $D_{(\delta)} \in \mathbb{R}^{p\times p}$ is a diagonal matrix with jth diagonal element equal to ρ_1^{-1} if $\delta_j = 1$, and ρ_0^{-1} if $\delta_j = 0$. The regression error σ is assumed to be known. The posterior conditional

distribution $\Pi(\delta|\theta,z)$ is a product of independent Bernoulli distributions constrained to be s-sparse:

$$\Pi(\delta|\theta,z) \;\propto\; \mathbf{1}_{\Delta_s}(\delta)\prod_{j=1}^{p}\left[\mathsf{q}_j\right]^{\delta_j}\left[1-\mathsf{q}_j\right]^{1-\delta_j}, \qquad \mathsf{q}_j \stackrel{\text{def}}{=} \frac{1}{1+\left(\frac{1}{\mathsf{q}}-1\right)\sqrt{\frac{\rho_0}{\rho_1}}e^{\frac{1}{2}(\rho_1-\rho_0)\theta_j^2}}. \tag{4.2}$$

We will assume that sampling from (4.2) is easy. This is the case when $s = p$ (by direct independent sampling), or when s is large (by a simple rejection scheme). A Metropolis–Hastings scheme could also be used, but we will focus our analysis on cases where an exact draw is made from (4.2). The conditional distribution of θ given δ is $\mathbf{N}_p(m_\delta, \sigma^2\Sigma_\delta)$, with m_δ and Σ_δ given by

$$m_\delta \stackrel{\text{def}}{=} \Sigma_\delta X'z \quad \text{and} \quad \Sigma_\delta \stackrel{\text{def}}{=} \left(X'X+\sigma^2 D_{(\delta)}^{-1}\right)^{-1}. \tag{4.3}$$

Put together, these two conditional distributions yield a simple Gibbs sampling algorithm for (4.1) with transition kernel given by

$$K(u,\mathrm{d}\theta) \stackrel{\text{def}}{=} \sum_{\omega\in\Delta_s}\Pi(\omega|u,z)\Pi(\mathrm{d}\theta|\omega,z), \tag{4.4}$$

with invariant distribution

$$\Pi(\mathrm{d}\theta|z) \propto \sum_{\delta\in\Delta_s}\left(\frac{\mathsf{q}}{1-\mathsf{q}}\right)^{\|\delta\|_0}\frac{e^{-\frac{1}{2}\theta' D_{(\delta)}^{-1}\theta}}{\sqrt{\det\left(2\pi D_{(\delta)}\right)}}e^{-\frac{1}{2\sigma^2}\|z-X\theta\|_2^2}\mathrm{d}\theta. \tag{4.5}$$

As pointed out above, K is a positive Markov kernel on the $L^2(\Pi)$. The posterior distribution (4.5) was analyzed in [21] from a statistical viewpoint. They show that $\Pi(\cdot|z)$ contracts and recovers well the true underlying signal as $n, p \to \infty$, provided that ρ_0 grows faster than n, and one sets ρ_1 to be of order n/p^2 and under some additional statistical assumptions. Therefore we shall analyze the mixing of the Markov kernel K in (4.4) in that regime.

To proceed we introduce some notation. For $\delta \in \Delta$ and $\theta \in \mathbb{R}^p$, we write θ_δ as short for the componentwise product of θ and δ, and we define $\delta^c \stackrel{\text{def}}{=} 1-\delta$, that is, $\delta_j^c = 1-\delta_j$, $1 \le j \le p$. For a matrix $A \in \mathbb{R}^{q\times p}$, A_δ (resp., A_{δ^c}) denotes the matrix of $\mathbb{R}^{q\times\|\delta\|_0}$ (resp., $\mathbb{R}^{q\times(p-\|\delta\|_0)}$) obtained by keeping only the columns of A for which $\delta_j = 1$ (resp., $\delta_j = 0$). When $\delta = e_j$ (the jth canonical unit vector of $\mathbb{R}^p$) we write A_δ (resp., A_{δ^c}) as A_j (resp., A_{-j}). For two elements δ, δ' of Δ, we write $\delta \supseteq \delta'$ to mean that for all j, $\delta_j = 1$ whenever $\delta'_j = 1$. The support of a vector $u \in \mathbb{R}^p$ is the vector $\mathsf{supp}(u) \in \Delta$ such that $\mathsf{supp}(u)_j = 1$ if and only if $|u_j| > 0$. An important role is played in the analysis by the matrices

$$L_\delta \stackrel{\text{def}}{=} I_n + \frac{1}{\sigma^2}XD_{(\delta)}X'$$

and the coherence of X defined as

$$\mathcal{C}(s) \stackrel{\text{def}}{=} \max_{\delta\in\Delta_s}\max_{j\neq\ell}\frac{\left|X_j'L_\delta^{-1}X_\ell\right|}{\sqrt{n\log(p)}}.$$

We will make the assumption that $\mathcal{C}(s)$ does not grow with p. It can be easily checked that if the columns of X are orthogonal, then $\mathcal{C}(s) = 0$. Furthermore, it can be shown that if X is a realization of random matrix with i.i.d. standard Gaussian entries, and provided that $n \geq As^2 \log(p)$, it holds that $\mathcal{C}(s) \leq c$ for some absolute constants c, A. We refer the reader to the supplementary material for details. For any integer $a \in \{1, \ldots, p\}$ we define

$$\varpi_a \stackrel{\text{def}}{=} \min_{\delta:\, \|\delta\|_0 \leq a} \inf \left\{ \frac{v' \left(X'_{\delta^c} L_\delta^{-1} X_{\delta^c} \right) v}{n\|v\|_2^2},\ v \in \mathbb{R}^{p-\|\delta\|_0},\ 0 < \|v\|_0 \leq a \right\}. \tag{4.6}$$

Remark 4.1. ϖ_a is a form restricted eigenvalue of the matrix X. It can be shown that if X is a random matrix with i.i.d. standard Gaussian entries, then $\varpi_a > 0$ for a of order $n/\log(p)$. We refer the reader to the supplementary material for details. □

We make the following assumptions.

H2.

1. *There exists a parameter value $\theta_\star \in \mathbb{R}^p$ with sparsity support $\delta_\star \in \Delta_s$, with $\|\delta_\star\|_0 = s_\star$, such that $p^{s_\star}\Pi(\delta_\star|z) \geq 1$.*
2. *For some constant $u > 0$, the prior parameter q satisfies*
$$\frac{\mathsf{q}}{1-\mathsf{q}} = \frac{1}{p^u}. \tag{4.7}$$
3. *The matrix X is nonrandom and such that*
$$\|X_j\|_2^2 = n, \quad j = 1, \ldots, p. \tag{4.8}$$
Furthermore there exists an integer $s_0 \geq s$, such that $\varpi_{s_0} > 0$, where ϖ_{s_0} is as defined as in (4.6).
4. *The prior parameters ρ_0, ρ_1 are positive and satisfy*
$$\rho_0 \geq c_1 \varpi_1 n, \quad \sigma^2 \rho_1 \leq \left(1 - \frac{\rho_1}{\rho_0}\right) n, \quad \text{and} \quad \sqrt{1 + \frac{ns}{\sigma^2 \rho_1}} \leq p^a \tag{4.9}$$
for some absolute constants $c_1 \geq 1, a > 0$, where ϖ_1 is as defined in (4.6).

Remark 4.2. It is well known that some form of restricted strong convexity is needed for signal recovery in high dimensions. Here this assumption takes the form $\varpi_{s_0} > 0$. In (4.9) we focus on the regime where ρ_0 grows as least linearly with n. This is a regime in which the posterior is known to have good contraction properties. The last two parts of (4.9) are easily satisfied and are imposed mostly to obtain simple mathematical formulas.

For some constant $c_0 > 0$, we introduce the event

$$\mathcal{E}_0 \stackrel{\text{def}}{=} \left\{ z \in \mathbb{R}^n : \max_{\delta \in \Delta_s} \sup_{1 \leq j \leq p} \frac{1}{\sigma} \left| \langle L_\delta^{-1} X_j, z - X\theta_\star \rangle \right| \leq \sqrt{c_0 n \log(p)} \right\}.$$

We note if $z \sim \mathbf{N}(X\theta_\star, \sigma^2 I_n)$ and $\|X_j\|_2 \leq \sqrt{n}$, then the event $z \in \mathcal{E}_0$ holds with high probability, with $c_0 = 2(s+1)$.

Theorem 4.3. *Suppose that H2 holds, and $z \in \mathcal{E}_0$. Fix $\varepsilon \in (0,1)$. Let $\nu_0 = \Pi(\cdot|\delta^{(i)}, z)$ for some arbitrary $\delta^{(i)} \in \Delta_s$. Suppose that we choose u in H2 large enough such that*

$$u > 2\max\left(2, \frac{\varrho}{\varpi_1}\right), \quad \text{where} \quad \varrho \stackrel{\text{def}}{=} \left(\sigma\sqrt{c_0} + \|\theta_\star\|_1 \mathcal{C}(s)\right)^2, \tag{4.10}$$

and the sample size n satisfies

$$n \geq \frac{A_0 u \sigma^2 \varrho \log(p)}{\varpi_1^2 \, \underline{\theta}_\star^2}, \quad \text{where} \quad \underline{\theta}_\star \stackrel{\text{def}}{=} \min_{j:\, \delta_{\star j}=1} |\theta_{\star j}|, \tag{4.11}$$

for some absolute constant A_0. Then there exists a constant C_0 that does not depend on n, p, or ε such that for all

$$N \geq C_0 \, s \left(\log\left(\frac{1}{\varepsilon}\right) + s_\star \|\theta_\star\|_\infty^2 n\right) e^{2\rho_0 \|\theta_\star\|_\infty^2}, \tag{4.12}$$

we have

$$\|\nu_0 K^N - \Pi\|_{\text{tv}} \leq \varepsilon.$$

Proof. See section (SM2) in the supplementary material. ∎

Theorem 4.3 follows from the standard spectral gap bound (2.3) together with Theorem 3.1 that we apply with $\zeta = 0$. In other words, we did not actually exploit the approximate spectral gap of K in Theorem 4.3. The result shows that when the prior parameter ρ_0 is taken as $\rho_0 = c_1 n$ (as required by [21] for good statistical behavior of the posterior), the mixing of the kernel K scales at most as $e^{\|\theta_\star\|_\infty^2 n}$. This result is consistent with the behaviors observed in the simulations. That said, it is important to add that (4.12) is an upper bound on the mixing time which may not be tight, and as such does not prove slow mixing.

In the regime considered here, posterior contraction holds and the posterior distribution assigns increasingly small probability to $\{\delta : \ \delta \not\supseteq \delta_\star\}$. The slow mixing obtained above is actually the result of the difficulty in the chain moving between $\{\delta : \ \delta \not\supseteq \delta_\star\}$ and $\{\delta : \ \delta \supseteq \delta_\star\}$. However, if the Markov chain is initialized in $\{\delta : \ \delta \supseteq \delta_\star\}$, it almost never transitions to $\{\delta : \ \delta \not\supseteq \delta_\star\}$, but nevertheless still recovers approximately well the posterior Π, since most of the probability mass is in $\{\delta : \ \delta \supseteq \delta_\star\}$. Using the approximate spectral gap, we now show that the latter Markov chain has a better mixing time than the conclusion of Theorem 4.3. Note that since $\delta_\star$ is not known, it is not possible to simply truncate the state space and apply classical spectral gap tools to the restriction of K to $\{\delta : \ \delta \supseteq \delta_\star\}$.

To derive this result, we shall focus on the unconstrained case where $s = p$ in (4.1). But any other value of $s \in \{s_0, \ldots, p\}$ will also work. We formalize the posterior contraction as follows. Given $k \geq 0$, we define

$$\mathcal{D}_k \stackrel{\text{def}}{=} \{\delta \in \Delta : \ \delta \supseteq \delta_\star, \ \|\delta\|_0 \leq \|\delta_\star\|_0 + k\},$$

which collects models that contain the true model $\delta_\star$ and have at most k false positives, and we introduce the event

$$\mathcal{E} \stackrel{\text{def}}{=} \mathcal{E}_0 \cap \left\{z \in \mathbb{R}^n : \quad \Pi(\mathcal{D}_k|z) \geq 1 - \frac{1}{p^{\frac{u}{2}(k+1)}} \ \text{ for all } k \geq 0\right\}.$$

We will say that posterior contraction holds when $z \in \mathcal{E}$. We will not directly establish this property. However, several existing works suggest that this description of the posterior contraction of $\Pi(\cdot|z)$ holds. For instance, under assumptions similar to those above, [21] shows that $\Pi(\mathcal{D}_0|Z) \geq 1 - \frac{a_1}{p^{a_2}}$ with high probability for positive constants a_1, a_2, and [1] shows that $z \in \mathcal{E}$ with high probability for a slightly modified version of the posterior distribution (4.1).

Theorem 4.4. *Assume H2 and $s = p$ in (4.1). Fix $\varepsilon \in (0,1)$. Suppose that $\nu_0 = \Pi(\cdot|\delta^{(i)}, z)$ for some $\delta^{(\mathrm{i})} \supseteq \delta_\star$ and such that*

$$\|\delta^{(\mathrm{i})}\|_0 \leq s_\star + \frac{u}{4(u+a)}(s_0 - s_\star) - \frac{u \log\left(\frac{40}{\varepsilon}\right)}{4(u+a)\log(p)}, \tag{4.13}$$

where a is as in H2. Suppose also that (4.10) and (4.11) hold. Then there exists a constant C_0 that does not depend on n, p, or ε such that for all $z \in \mathcal{E}$, and all

$$N \geq C_0 \|\delta^{(\mathrm{i})}\|_0 \left[\log\left(\varepsilon^{-1}\right) + \|\delta^{(\mathrm{i})}\|_0 u \log(p)\right] p^{\frac{\rho_0}{n}\frac{2\varrho}{\varpi_1}}, \tag{4.14}$$

we have

$$\|\nu_0 K^N - \Pi\|_{\mathrm{tv}} \leq \varepsilon.$$

Proof. See section (SM3) in the supplementary material. ∎

Condition (4.13) restricts the number of false positives of $\delta^{(\mathrm{i})}$ in the initial distribution. This condition can be relaxed if the contraction of Π on $\mathcal{D}_k$ is faster than the polynomial form assumed in the event $\mathcal{E}$.

Theorem 4.4 shows that when posterior contraction holds ($z \in \mathcal{E}$), and a good initial distribution is used, the mixing time of K is polynomial in the dimension p and less sensitive to large values of ρ_0. Instead, the mixing time depends mainly on the coherence parameter $\mathcal{C}(s)$ and the restricted eigenvalue ϖ_1. One clear roadblock to the practical use of this result is finding the initial $\delta^{(\mathrm{i})}$ such that $\delta^{(\mathrm{i})} \supseteq \delta_\star$, since $\delta_\star$ is typically unknown. In practice, various frequentist estimators such as the lasso can be used. At least in a high signal-to-noise-ratio setting the lasso estimator is known to contain the true model under mild assumptions. We refer the reader to, for instance, [19].

One of the first papers to analyze the mixing times of the MCMC algorithm in high-dimensional linear regression models and highlight fast/slow mixing behaviors is [25]. Their posterior distribution is slightly different from what we looked at in this work. Specifically [25] applied a Metropolis–Gibbs sampler to the marginal distribution of δ, whereas we consider here a data-augmentation sampler applied to the marginal distribution of θ. These authors show that in general their sampler has a mixing time that is exponential in p unless the state space is restricted to models δ for which $\|\delta\|_0 \leq s$ for some threshold s, in which case the worst-case mixing time is $O(s^2 np \log(p))$.

4.1. Numerical illustrations.

We illustrate some of the conclusions with the following simulation study. We consider a linear regression model with Gaussian noise $\mathbf{N}(0, \sigma^2)$, where σ^2 is set to 1. We experiment with sample size $n = p$ and dimensions $p \in \{500, 1000, 2000, 3000, 4000\}$. We take $X \in \mathbb{R}^{n \times p}$ as a random matrix with i.i.d. rows drawn from $\mathbf{N}_p(0, \Sigma)$ under two scenarios: a low coherence setting where $\Sigma = I_p$, and a high coherence where $\Sigma_{ij} = 0.9^{|j-i|}$.

Table 1

Average empirical mixing time of the samplers in a low-coherence setting. Based on 50 simulation replications. The numbers in parenthesis are standard errors. The notation > a means that some (or all) of the replicated mixing times have been truncated.

		$p = 500$	$p = 1{,}000$	$p = 2{,}000$	$p = 3{,}000$	$p = 4{,}000$
FN	$\rho_0 = n$	866.3 (3,204)	423.6 (2,735)	147.1 (575)	> 437.3	> 871.0
	$\rho_0 = n^{1.5}$	> 11,125.8	> 13,662.6	> 13,2371.6	> 15,948.0	> 16237.3
	Yang et al.	5,244.2 (1,379)	12,208.5 (2,463)	27,617.6 (5,803)	43,821.9 (6,453)	54,697.9 (5,611)
no FN	$\rho_0 = n$	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
	$\rho_0 = n^{1.5}$	30.9 (81)	43.7 (55)	123.2 (251)	241.2 (535)	215.3 (250)
	Yang et al.	5,191.0 (1,503)	11,975.9 (2,769)	26,877.8 (4,786)	42,285.7 (8,721)	56,264.3 (10,362)

After sampling, we normalized the columns of X to each have norm $\sqrt{n}$. We fix the number of nonzero coefficients to $s_\star = 10$, and $\delta_\star$ is given by

$$\delta_\star = (\underbrace{1, \ldots, 1}_{10}, \underbrace{0, \ldots, 0}_{p-10}).$$

The nonzero coefficients of $\theta_\star$ are uniformly drawn from $(-a-1, -a) \cup (a, a+1)$, where

$$a = 4\sqrt{\frac{\log(p)}{n}}.$$

We use the following prior parameters values:

$$u = 2, \quad \rho_1 = \frac{n}{p^{2.1}}, \quad \rho_0 \in \left\{ \frac{n}{\sigma^2}, \frac{n^{1.5}}{\sigma^2} \right\}.$$

These scalings of ρ_0 and ρ_1 roughly match the recommendations of [21] to get posterior contraction of $\Pi(\cdot|z)$. We use an initial distribution $\nu_0 = \Pi(\cdot|\delta^{(\mathsf{i})}, z)$, where $\delta^{(\mathsf{i})}$ is such that $\|\delta^{(\mathsf{i})} - \delta_\star\|_0 = 2p/10$, with two scenarios: a scenario FN (false negative) where 5 out of 10 of the true positives of $\delta_\star$ are set to 0, and a scenario no FN where $\delta^{(\mathsf{i})}$ has only false positives. To monitor the mixing, we compute the sensitivity and the precision at iteration k as

$$\mathsf{SEN}_k = \frac{1}{s_\star} \sum_{j=1}^{p} \mathbf{1}_{\{|\delta_{k,j}|>0\}} \mathbf{1}_{\{|\delta_{\star,j}|>0\}}, \qquad \mathsf{PREC}_k = \frac{\sum_{j=1}^{p} \mathbf{1}_{\{|\delta_{k,j}|>0\}} \mathbf{1}_{\{|\delta_{\star,j}|>0\}}}{\sum_{j=1}^{p} \mathbf{1}_{\{|\delta_{k,j}|>0\}}}.$$

We empirically measure the mixing time of the algorithm as the first time k where both SEN_k and PREC_k reach 1, truncated to 2×10^4—that is, we stop any run that has not mixed by 20,000 iterations. For the sampler of [25], we stop any run that has not mixed by 10^5 iterations. The average empirical mixing times thus obtained (based on 50 independent MCMC replications) are presented in Tables 1 and 2.

We can make the following observations.

1. There is sharp difference in behavior between the low- and high-coherence settings.
2. As predicted by our theory, the Markov kernel K mixes better when there is no false negative in the initialization. The algorithm of [25] seems impervious to the initialization. It should be noted in comparing the two algorithms that an iteration of the algorithm of [25] costs roughly p times less than an iteration of the Markov kernel K.

Table 2

Average empirical mixing time of the samplers in a high-coherence setting. Based on 50 simulation replications. The numbers in parenthesis are standard errors. The notation > a means that some (or all) of the replicated mixing times have been truncated.

		$p = 500$	$p = 1{,}000$	$p = 2{,}000$	$p = 3{,}000$	$p = 4{,}000$
FN	$\rho_0 = n$	> 20,000	> 19,200	> 18,400	> 17,870	> 19129.1
	$\rho_0 = n^{1.5}$	> 20,000	> 20,000	> 20,000	> 20,000	> 20,000
	Yang et al.	> 100,000	> 91,177	> 75,373	> 83,246	> 84,972
no FN	$\rho_0 = n$	> 880.1	> 1,200.1	> 400.9	> 800.96	> 900.1
	$\rho_0 = n^{1.5}$	> 416.8	> 1,246.2	> 874.2	> 425.2	> 313.6
	Yang et al.	> 98,067	> 87,424	> 73,253	> 77,902	> 82,205

3. The third observation that can be drawn from the results is that when there are false negatives, the Markov kernel K mixes better with $\rho_0 = n/\sigma^2$, compared to $\rho_0 > n/\sigma^2$, as predicted by our result. The difference is less noticeable in the high-coherence setting. This observation is also explained by our bound, since in a high-coherence setting the parameter ϱ is expected to be large. Another observation here is that when there are false negatives in the initialization, the mixing time becomes highly variable (several runs have hit the wallclock).
4. Finally, we notice that the theory of [25] does not fully describe the behavior of their algorithm, as we see a significant loss of performance in their algorithm with high-coherence design matrices, which cannot be clearly explained by their result.

5. Proof of Lemma 2.1. Fix $\zeta \in [0, 1)$, and take $f \in L^2(\pi)$. Since $\pi(f) = \pi(Kf)$ and $\mathsf{Var}_\pi(f) = \langle f, f\rangle_\pi - \pi(f)^2$, we have

$$\begin{aligned}(5.1)\quad \mathsf{Var}_\pi(Kf) - \mathsf{Var}_\pi(f) &= \langle Kf, Kf\rangle_\pi - \langle f, f\rangle_\pi = \langle f, K_\star K f\rangle_\pi - \langle f, f\rangle_\pi \\ &= -\frac{1}{2}\int\int (f(y) - f(x))^2\, \pi(\mathrm{d}x)(K^\star K)(x, \mathrm{d}y).\end{aligned}$$

Using the last display together with the definition of $\mathcal{E}_{K^\star K}(f, f)$, we conclude that for all $f \in L^2(\pi)$,

$$(5.2)\qquad \mathsf{Var}_\pi(Kf) \le \mathsf{Var}_\pi(f) - \mathcal{E}_{K^\star K}(f, f).$$

Suppose that $+\infty > \|f\|_\infty > 0$. If $\mathsf{Var}_\pi(f) > \zeta\|f\|_\infty^2$, then by (5.2)

$$\begin{aligned}\mathsf{Var}_\pi(Kf) &\le \mathsf{Var}_\pi(f) - \|f\|_\infty^2 \mathcal{E}_{K^\star K}\left(\frac{f}{\|f\|_\infty}, \frac{f}{\|f\|_\infty}\right) \\ &\le \mathsf{Var}_\pi(f) - \|f\|_\infty^2 \lambda_\zeta(K^\star K)\left(\mathsf{Var}_\pi\left(\frac{f}{\|f\|_\infty}\right) - \frac{\zeta}{2}\right) \\ &\le \mathsf{Var}_\pi(f)\,(1 - \min(1, \lambda_\zeta(K^\star K))) + \zeta\|f\|_\infty^2 \lambda_\zeta(K^\star K).\end{aligned}$$

If $\mathsf{Var}_\pi(f) \le \zeta\|f\|_\infty^2$, then by (5.2)

$$\mathsf{Var}_\pi(Kf) \le \mathsf{Var}_\pi(f)\,(1 - \min(1, \lambda_\zeta(K^\star K))) + \zeta\|f\|_\infty^2 \lambda_\zeta(K^\star K).$$

But if $\|f\|_\infty = 0$, then $\mathsf{Var}_\pi(f) = 0$, and hence $\mathsf{Var}_\pi(Kf) = 0$ by (5.2), so that the last display continues to hold. Similarly, if $\|f\|_\infty = +\infty$, the last display continues to hold. We conclude that for all $f \in L^2(\pi)$,

$$\mathsf{Var}_\pi(Kf) \le \mathsf{Var}_\pi(f)\,(1 - \min(1, \lambda_\zeta(K^\star K))) + \zeta \|f\|_\infty^2 \lambda_\zeta(K^\star K).$$

We iterate the above inequality to deduce that, for all $f \in L^2(\pi)$ and for all $n \ge 1$,

$$\begin{aligned}\mathsf{Var}_\pi(K^n f) &\le \mathsf{Var}_\pi(f)\,(1 - \min(1, \lambda_\zeta(K^\star K)))^n \\ &\quad + \zeta \lambda_\zeta(K^\star K) \sum_{j \ge 0} (1 - \min(1, \lambda_\zeta(K^\star K)))^j \, \|K^{n-j-1} f\|_\infty^2 \\ &\le \mathsf{Var}_\pi(f)\,(1 - \lambda_\zeta(K^\star K))^n + \frac{\zeta \lambda_\zeta(K^\star K)}{\min(1, \lambda_\zeta(K^\star K))} \|f\|_\infty^2 \\ &\le \mathsf{Var}_\pi(f)\,(1 - \lambda_\zeta(K^\star K))^n + 2\zeta \|f\|_\infty^2,\end{aligned}$$

where the last inequality uses the $\lambda_\zeta(K^\star K) \in [0, 2]$. If K is positive, then $K^\star K = K^2$, and K admits a square root: there exists a bounded π-reversible operator S such that $S^2 = K$, and S commutes with K. Furthermore, with $\mathbb{I}$ denoting the identity operator, $\mathbb{I} - K$ is also a positive operator: $\mathbb{I} - K$ is clearly π-reversible, and $\langle f, (\mathbb{I} - K) f \rangle_\pi = \|f\|_2^2 - \langle f, Kf \rangle_\pi \ge 0$, using the fact that the operator norm of K is less than or equal to one. Hence

$$\langle f, Kf \rangle_\pi - \langle f, K^\star K f \rangle_\pi = \left\langle f, (K - K^2) f \right\rangle_\pi = \langle f, S(I - K) S f \rangle_\pi = \langle Sf, (I - K) S f \rangle_\pi \ge 0.$$

Therefore when K is positive we can replace (5.1) by

$$\mathsf{Var}_\pi(Kf) \le \mathsf{Var}_\pi(f) - \frac{1}{2} \int \int (f(y) - f(x))^2 \, \pi(\mathrm{d}x) K(x, \mathrm{d}y)$$

and proceed as above to obtain the stated bound. This ends the proof. ∎

6. Proof Lemma 2.3. Take a measurable function $f : \mathcal{X} \to \mathbb{R}$ such that $\|f\|_\infty = 1$ and $\mathsf{Var}_\pi(f) > \zeta$. We have

$$\begin{aligned}2\mathsf{Var}_\pi(f) &= \int_{\mathcal{X}_0} \int_{\mathcal{X}_0} (f(y) - f(x))^2 \pi(\mathrm{d}x) \pi(\mathrm{d}y) \\ &\quad + 2 \int_{\mathcal{X}_0} \int_{\mathcal{X} \setminus \mathcal{X}_0} (f(y) - f(x))^2 \pi(\mathrm{d}x) \pi(\mathrm{d}y) + \int_{\mathcal{X} \setminus \mathcal{X}_0} \int_{\mathcal{X} \setminus \mathcal{X}_0} (f(y) - f(x))^2 \pi(\mathrm{d}x) \pi(\mathrm{d}y) \\ &\le \int_{\mathcal{X}_0} \int_{\mathcal{X}_0} (f(y) - f(x))^2 \pi(\mathrm{d}x) \pi(\mathrm{d}y) + 8 \|f\|_\infty^2 \pi(\mathcal{X} \setminus \mathcal{X}_0) \pi(\mathcal{X}_0) + 4 \|f\|_\infty^2 \pi(\mathcal{X} \setminus \mathcal{X}_0)^2 \\ &\le \int_{\mathcal{X}_0} \int_{\mathcal{X}_0} (f(y) - f(x))^2 \pi(\mathrm{d}x) \pi(\mathrm{d}y) + 8 \pi(\mathcal{X} \setminus \mathcal{X}_0).\end{aligned}$$

Using $\pi(\mathcal{X}_0) \ge 1 - (\zeta/8)$, we get

$$2 \left(\mathsf{Var}_\pi(f) - \frac{\zeta}{2} \right) \ge \int_{\mathcal{X}_0} \int_{\mathcal{X}_0} \pi(\mathrm{d}x) \pi(\mathrm{d}y) (f(y) - f(x))^2.$$

Hence

$$\frac{\mathcal{E}(f,f)}{\mathsf{Var}_\pi(f)-\frac{\zeta}{2}} \geq \frac{\int_{\mathcal{X}0}\int_{\mathcal{X}0}\pi(\mathrm{d}x)K(x,\mathrm{d}y)(f(y)-f(x))^2}{\int_{\mathcal{X}_0}\int_{\mathcal{X}0}\pi(\mathrm{d}x)\pi(\mathrm{d}y)(f(y)-f(x))^2} \geq \lambda_{\mathcal{X}_0}.$$

The statement bound easily follows. ∎

7. Proof of Theorem 3.1. We rely on the following lemma due to [16] (inequality (47)).

Lemma 7.1. *Let $\nu(\mathrm{d}x)=f_\nu(x)\mathrm{d}x$, $\mu(\mathrm{d}x)=f_\mu(x)\mathrm{d}x$ be two probability measures on some measurable space with reference measure $\mathrm{d}x$, such that $\int\min(f_\mu(x),f_\nu(x))\mathrm{d}x>\epsilon$ for some $\epsilon>0$. Then for any measurable function h such that $\int h^2(x)\nu(\mathrm{d}x)<\infty$ and $\int h^2(x)\mu(\mathrm{d}x)<\infty$, we have*

$$\int(h(y)-h(x))^2\mu(\mathrm{d}y)\nu(\mathrm{d}x) \leq \frac{2-\epsilon}{2\epsilon}\left[\int(h(y)-h(x))^2\mu(\mathrm{d}y)\mu(\mathrm{d}x)+\int(h(y)-h(x))^2\nu(\mathrm{d}y)\nu(\mathrm{d}x)\right].$$

Choose $f\in L^2(\pi)$ such that $\|f\|_\infty=1$. We define

$$\mathcal{E}_i(f,f)\stackrel{\text{def}}{=}\frac{1}{2}\int_{\mathsf{B}_i}\int_{\mathsf{B}_i}(f(y)-f(x))^2\,\pi_i(\mathrm{d}x)K_i(x,\mathrm{d}y).$$

From the definition

$$2\mathcal{E}(f,f)=\int_{\mathcal{X}}\int_{\mathcal{X}}(f(y)-f(x))^2\,\pi(\mathrm{d}x)\left[\sum_{i\in\mathsf{I}}\pi(i|x)K_i(x,\mathrm{d}y)\right] =\sum_{i\in\mathsf{I}}\pi(i)\int_{\mathcal{X}}\int_{\mathcal{X}}(f(y)-f(x))^2\,\pi_i(\mathrm{d}x)K_i(x,\mathrm{d}y) \geq 2\sum_{i\in\mathsf{I}}\pi(i)\mathcal{E}_i(f,f)\geq 2\sum_{i\in\mathsf{I}_0}\pi(i)\mathcal{E}_i(f,f). \tag{7.1}$$

Using $\mathsf{I}\times\mathcal{X}=\bar{\mathsf{B}}\cup\bar{\mathsf{B}}^c$, where $\bar{\mathsf{B}}^c$ denotes the complement of $\bar{\mathsf{B}}$, and $\|f\|_\infty=1$, we have

$$\begin{aligned}2\mathsf{Var}_\pi(f)=&\int_{\bar{\mathsf{B}}}\int_{\bar{\mathsf{B}}}(f(y)-f(x))^2\,\bar{\pi}(\mathrm{d}i,\mathrm{d}x)\bar{\pi}(\mathrm{d}j,\mathrm{d}y)\\&+2\int_{\bar{\mathsf{B}}}\int_{\bar{\mathsf{B}}^c}(f(y)-f(x))^2\,\bar{\pi}(\mathrm{d}i,\mathrm{d}x)\bar{\pi}(\mathrm{d}j,\mathrm{d}y)\\&+\int_{\bar{\mathsf{B}}^c}\int_{\bar{\mathsf{B}}^c}(f(y)-f(x))^2\,\bar{\pi}(\mathrm{d}i,\mathrm{d}x)\bar{\pi}(\mathrm{d}j,\mathrm{d}y)\\&\leq\int_{\bar{\mathsf{B}}}\int_{\bar{\mathsf{B}}}(f(y)-f(x))^2\,\bar{\pi}(\mathrm{d}i,\mathrm{d}x)\bar{\pi}(\mathrm{d}j,\mathrm{d}y)+10\bar{\pi}(\mathsf{B}^c).\end{aligned} \tag{7.2}$$

Expanding the first term on the right-hand side of (7.2) and using (3.9), it follows that

$$2\left(\mathsf{Var}_\pi(f)-\frac{\zeta}{2}\right)\le \sum_{i\in \mathsf{I}_0}\pi(i)^2\int_{\mathsf{B}_i}\int_{\mathsf{B}_i}(f(y)-f(x))^2\,\pi_i(\mathrm{d}x)\pi_i(\mathrm{d}y) + \sum_{i\neq j,\ i,j\in\mathsf{I}_0}\pi(i)\pi(j)\pi_i(\mathsf{B}_i)\pi_j(\mathsf{B}_j)\int_{\mathsf{B}_i}\int_{\mathsf{B}_j}(f(y)-f(x))^2\,\frac{\pi_i(\mathrm{d}x)}{\pi_i(\mathsf{B}_i)}\frac{\pi_j(\mathrm{d}y)}{\pi_j(\mathsf{B}_j)}. \tag{7.3}$$

Given an edge e in $\mathcal{G}$, let us write $e=(e_-,,e_+)$ to denote the two incident nodes of the edge. For $i\neq j\in\mathsf{I}_0$, let γ_{ij} denote the chosen canonical path between i and j, and let $i_0,i_1,\dots,i_\ell$ be the nodes on that canonical path (with $i_0=i$ and $i_\ell=j$). By introducing generic variables $z_{i_k}\in\mathsf{B}_{i_k}$, one can write $f(z_{i_\ell})-f(z_{i_0})=\sum_{k=1}^{\ell}f(z_{i_k})-f(z_{i_{k-1}})$. Using this and the Cauchy–Schwarz inequality, we have

$$\int_{\mathsf{B}_i}\int_{\mathsf{B}_j}(f(y)-f(x))^2\,\frac{\pi_i(\mathrm{d}x)}{\pi_i(\mathsf{B}_i)}\frac{\pi_j(\mathrm{d}y)}{\pi_j(\mathsf{B}_j)} \le |\gamma_{ij}|\sum_{e\in\gamma_{ij}}\int_{\mathsf{B}_{e_-}}\int_{\mathsf{B}_{e_+}}(f(y)-f(x))^2\,\frac{\pi_{e_-}(\mathrm{d}x)}{\pi_{e_-}(\mathsf{B}_{e_-})}\frac{\pi_{e_+}(\mathrm{d}y)}{\pi_{e_+}(\mathsf{B}_{e_+})}, \tag{7.4}$$

where $|\gamma_{ij}|$ denotes the number of edges on the canonical path γ_{ij}. By (3.6) and Lemma 7.1, and by using also the assumption that $\pi_i(\mathsf{B}_i)\ge 1/2$, the summation on the right-hand side of (7.4) is upper bounded by

$$\begin{aligned}&\frac{4}{\kappa}\sum_{e\in\gamma_{ij}}\int_{\mathsf{B}_{e_-}}\int_{\mathsf{B}_{e_-}}(f(y)-f(x))^2\,\pi_{e_-}(\mathrm{d}x)\pi_{e_-}(\mathrm{d}y)\\ &\qquad+\frac{4}{\kappa}\sum_{e\in\gamma_{ij}}\int_{\mathsf{B}_{e_+}}\int_{\mathsf{B}_{e_+}}(f(y)-f(x))^2\,\pi_{e_+}(\mathrm{d}x)\pi_{e_+}(\mathrm{d}y)\\ &\qquad\le\frac{8}{\kappa}\sum_{\iota\in\gamma_{ij}}\int_{\mathsf{B}_\iota}\int_{\mathsf{B}_\iota}(f(y)-f(x))^2\,\pi_\iota(\mathrm{d}x)\pi_\iota(\mathrm{d}y),\end{aligned}$$

where the summation $e\in\gamma_{ij}$ is taken over all edges along the path γ_{ij}, whereas the summation $\iota\in\gamma_{ij}$ is taken over all nodes ι along the path γ_{ij} including i and j. Hence

$$\begin{aligned}&\sum_{i\neq j,\ i,j\in\mathsf{I}_0}\pi(i)\pi(j)\pi_i(\mathsf{B}_i)\pi_j(\mathsf{B}_j)\int_{\mathsf{B}_i}\int_{\mathsf{B}_j}(f(y)-f(x))^2\,\frac{\pi_i(\mathrm{d}x)}{\pi_i(\mathsf{B}_i)}\frac{\pi_j(\mathrm{d}y)}{\pi_j(\mathsf{B}_j)}\\ &\qquad\le\frac{8}{\kappa}\sum_{\iota\in\mathsf{I}_0}\pi(\iota)\int_{\mathsf{B}_\iota}\int_{\mathsf{B}_\iota}(f(y)-f(x))^2\,\pi_\iota(\mathrm{d}x)\pi_\iota(\mathrm{d}y)\sum_{\gamma_{ij}\in\Gamma:\,\gamma_{ij}\ni\iota}|\gamma_{ij}|\frac{\pi(i)\pi(j)}{\pi(\iota)},\end{aligned} \tag{7.5}$$

which together with (7.3) yields

$$2\left(\mathsf{Var}_\pi(f)-\frac{\zeta}{2}\right)\le\left(1+\frac{8\mathsf{m}}{\kappa}\right)\sum_{i\in\mathsf{I}_0}\pi(i)\int_{\mathsf{B}_i}\int_{\mathsf{B}_i}(f(y)-f(x))^2\,\pi_i(\mathrm{d}x)\pi_i(\mathrm{d}y). \tag{7.6}$$

From the definition of $\lambda_i(K_i)$, we have

$$\int_{\mathsf{B}_i}\int_{\mathsf{B}_i}(f(y)-f(x))^2\,\pi_i(\mathrm{d}x)\pi_i(\mathrm{d}y) \le \frac{2\mathcal{E}_i(f,f)}{\lambda_i(K_i)}, \tag{7.7}$$

which we use in (7.6), to arrive at

$$\left(\mathsf{Var}_\pi(f)-\frac{\zeta}{2}\right) \le \frac{\left(1+\frac{8\mathsf{m}}{\kappa}\right)}{\min_{i\in\mathsf{I}_0}\lambda_i(K_i)}\sum_{i\in\mathsf{I}_0}\pi(i)\mathcal{E}_i(f,f). \tag{7.8}$$

Inequalities (7.8) and (7.1) together yield

$$\frac{\mathcal{E}(f,f)}{\left(\mathsf{Var}_\pi(f)-\frac{\zeta}{2}\right)} \ge \frac{\min_{i\in\mathsf{I}_0}\lambda_i(K_i)}{1+\frac{8\mathsf{m}}{\kappa}} \ge \frac{\kappa}{1+8\mathsf{m}}\min_{i\in\mathsf{I}_0}\lambda_i(K_i),$$

which, together with the definition (2.4) and $\kappa \le 1$, implies the stated bound. ∎

Acknowledgments. I'm grateful to Joonha Park, Daniel Rudolf, and the anonymous referees for comments, discussions, and suggestions that have significantly improved this work from the first manuscript.

REFERENCES

[1] Y. Atchade and A. Bhattacharyya, *An Approach to Large-Scale Quasi-Bayesian Inference with Spike-and-Slab Priors*, preprint, https://arxiv.org/abs/1803.10282, 2018.
[2] N. Bou-Rabee and M. Hairer, *Nonasymptotic mixing of the MALA algorithm*, IMA J. Numer. Anal., 3 (2013), pp. 380–110.
[3] P. Cattiaux and A. Guillin, *Trends to equilibrium in total variation distance*, Ann. Inst. H. Poincaré Probab. Statist., 45 (2009), pp. 117–145.
[4] P. Diaconis and D. Stroock, *Geometric bounds for eigenvalues of Markov chains*, Ann. Appl. Probab., 1 (1991), pp. 36–61.
[5] R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu, *Log-concave Sampling: Metropolis-Hastings Algorithms Are Fast!*, preprint, https://arxiv.org/abs/1801.02309, 2018.
[6] A. Eberle, *Error bounds for Metropolis–Hastings algorithms applied to perturbations of Gaussian measures in high dimensions*, Ann. Appl. Probab., 24 (2014), pp. 337–377.
[7] C. Efthymiou, T. P. Hayes, D. Stefankovic, E. Vigoda, and Y. Yin, *Convergence of MCMC and loopy BP in the tree uniqueness region for the hard-core model*, in Proceedings of the 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS), IEEE Computer Society, 2016, pp. 704–713.
[8] A. Frieze, R. Kannan, and N. Polson, *Sampling from log-concave distributions*, Ann. Appl. Probab., 4 (1994), pp. 812–837.
[9] R. Ge, H. Lee, and A. Risteski, *Simulated Tempering Langevin Monte Carlo* II: *An Improved Proof using Soft Markov Chain Decomposition*, preprint, https://arxiv.org/abs/1812.00793, 2018.
[10] E. I. George and R. E. McCulloch, *Approaches to Bayesian variable selection*, Statist. Sinica, 7 (1997), pp. 339–373.
[11] M. Jerrum, J.-B. Son, P.Tetali, and E. Vigoda, *Elementary bounds on Poincaré and log-Sobolev constants for decomposable Markov chains*, Ann. Appl. Probab., 14 (2004), pp. 1741–1765.
[12] T. M. Liggett, *l_2 rates of convergence for attractive reversible nearest particle systems: The critical case*, Ann. Probab., 19 (1991), pp. 935–959.
[13] L. Lovász, *Hit-and-run mixes fast*, Math. Program., 86 (1999), pp. 443–461.

[14] L. Lovász and M. Simonovits, *Random walks in a convex body and an improved volume algorithm*, Random Structures Algorithms, 4 (1993), pp. 359–412.
[15] L. Lovász and S. Vempala, *The geometry of logconcave functions and sampling algorithms*, Random Structures Algorithms, 30 (2007), pp. 307–358.
[16] N. Madras and D. Randall, *Markov chain decomposition for convergence rate analysis*, Ann. Appl. Probab., 12 (2002), pp. 581–606.
[17] R. A. Martin and D. Randall, *Sampling adsorbing staircase walks using a new Markov chain decomposition method*, in Proceedings of the 41st Annual Symposium on Foundations of Computer Science, FOCS '00, IEEE Computer Society, 2000, pp. 492–502.
[18] F. Medina-Aguayo, D. Rudolf, and N. Schweizer, *Perturbation bounds for Monte Carlo within Metropolis via restricted approximations*, Stochastic Process. Appl., 130 (2020), pp. 2200–2227.
[19] N. Meinshausen and B. Yu, *Lasso-type recovery of sparse representations for high-dimensional data*, Ann. Statist., 37 (2009), pp. 246–270.
[20] R. Montenegro and P. Tetali, *Mathematical aspects of mixing times in Markov chains*, Found. Trends Theoret. Comput. Sci., 1 (2006), pp. 237–354.
[21] N. Narisetty and X. He, *Bayesian variable selection with shrinking and diffusing priors*, Ann. Statist., 42 (2014), pp. 789–817.
[22] A. Sinclair, *Improved bounds for mixing rates of Markov chains and multicommodity flow*, Combin. Probab. Comput., 1 (1992), pp. 351–370.
[23] D. B. Woodard, S. C. Schmidler, and M. Huber, *Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions*, Ann. Appl. Probab., 19 (2009), pp. 617–640.
[24] J. Yang and J. S. Rosenthal, *Complexity Results for MCMC Derived from Quantitative Bounds*, preprint, https://arxiv.org/abs/1708.00829, 2019.
[25] Y. Yang, M. J. Wainwright, and M. I. Jordan, *On the computational complexity of high-dimensional Bayesian variable selection*, Ann. Statist., 44 (2016), pp. 2497–2532.