Approximate spectral gaps for Markov chains 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 a 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 chains mixing times, Spectral gaps, Canonical paths, MCMC for mixtures of densities, High-dimensional linear regression models

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

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 to gain insight into the behavior of MCMC algorithms. However this technique may be inapropriate when dealing with distributions with small isolated local modes. To be more precise, let π be some probability measure of interest on some measure 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 (to which K is sensitive) 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. Nevertheless, 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 on the s-conductance of L. Lovasz and M. Simonovits ([Lovász and Simonovits(1993)]), we develop an idea of approximate spectral gap (that we call 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-effect of overly small (and potentially problematic) sets.

---

*

Funding: This work is partially supported by the NSF grant DMS1513040.

†Boston University, Department of Mathematics and Statistics, 111 Cummington Mall, Boston, 02215, MA, United States (atchade@bu.edu, http://math.bu.edu/people/atchade/).

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. [Lovász and Simonovits(1993), Frieze et al.(1994), Lovász(1999), Lovász and Vempala(2007)], and [Dwivedi et al.(2018)] 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 [Madras and Randall(2002)]. We re-examine [Madras and Randall(2002)]'s argument using the concept of ζ-spectral gap, leading to Theorem 3.1 that 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. [George and McCulloch(1997), Narisetty and He(2014)]). We show that when properly tuned, the algorithm has a mixing time that grows at most polynomially with p, the number of regressors in the model (Theorem 4.2). Our result derived from the approximate spectral gap also suggests that the mixing time improves when a good initial distribution is used, provided that posterior contraction towards the true model holds (Theorem 4.3).

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 [Madras and Randall(2002)]. We put these two results together to analysis the linear regression model in Section 4. Some numerical simulations are detailed in Section 4.1.

2. Approximate spectral gaps for Markov chains.

Let π be a probability measure on some Polish space $(\mathcal{X}, \mathcal{B})$ (where $\mathcal{B}$ is its Borel sigma-algebra), equipped with a reference sigma-finite measure denoted $\mathrm{d}x$. In the applications that we have in mind, $\mathcal{X}$ is the Euclidean space $\mathbb{R}^p$ equipped with its Lebesgue measure. We assume that π is absolutely continuous with respect to $\mathrm{d}x$, and we will abuse notation and use π to denote both π and its density: $\pi(\mathrm{d}x) = \pi(x)\mathrm{d}x$. We let $L^2(\pi)$ denote the Hilbert space of all real-valued square-integrable (wrt 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 $\|\cdot\|_{2,\pi}$. More generally, for $s \geq 1$, we set $\|f\|_{s,\pi} \stackrel{\text{def}}{=} \left(\int_{\mathcal{X}} |f(x)|^s \pi(\mathrm{d}x)\right)^{1/s}$. For $s = +\infty$, $\|f\|_{s,\pi}$ is defined as the essential supremum of $|f|$ with respect to π. If P is a Markov kernel on $\mathcal{X}$, and $n \geq 1$ an integer, P^n denotes the n-th 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. If $f : \mathcal{X} \to \mathbb{R}$ is a measurable function, then $Pf : \mathcal{X} \to \mathbb{R}$ is the function defined as $Pf(x) \stackrel{\text{def}}{=} \int_{\mathcal{X}} P(x, \mathrm{d}z)f(z)$, $x \in \mathcal{X}$, assuming that 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

$$\|\mu - \nu\|_{\mathrm{tv}} \stackrel{\text{def}}{=} 2 \sup_{A \in \mathcal{B}} (\mu(A) - \nu(A)).$$

Let K be a Markov kernel on $\mathcal{X}$ that is reversible with respect to π. That is for all $A, B \in \mathcal{B}$,

$$\int_A \pi(\mathrm{d}x) \int_B K(x, \mathrm{d}y) = \int_B \pi(\mathrm{d}x) \int_A K(x, \mathrm{d}y).$$

We will also assume throughout that K is lazy in the sense that $K(x, \{x\}) \geq \frac{1}{2}$. The concept of spectral gap and the related Poincare's inequalities are commonly used to quantify Markov chains' mixing times. For $f \in L^2(\pi)$, we set $\pi(f) \stackrel{\text{def}}{=} \int_{\mathcal{X}} f(x)\pi(\mathrm{d}x)$, $\mathsf{Var}_\pi(f) \stackrel{\text{def}}{=} \|f - \pi(f)\|_{2,\pi}^2$, and $\mathcal{E}(f, f) \stackrel{\text{def}}{=} \frac{1}{2} \int\int (f(y) - f(x))^2 \pi(\mathrm{d}x)K(x, \mathrm{d}y)$. The spectral gap of K is then defined as

$$\mathsf{SpecGap}(K) \stackrel{\text{def}}{=} \inf \left\{ \frac{\mathcal{E}(f, f)}{\mathsf{Var}_\pi(f)},\ f \in L^2(\pi), \text{ s.t. } \mathsf{Var}_\pi(f) > 0 \right\}.$$

It is well-known (see for instance [Montenegro and Tetali(2006)] Corollary 2.15) that if $\pi_0(\mathrm{d}x) = f_0(x)\pi(\mathrm{d}x)$, and $f_0 \in L^2(\pi)$, then

$$\|\pi_0 K^n - \pi\|_{\mathrm{tv}}^2 \leq \mathsf{Var}_\pi(f_0) \left(1 - \mathsf{SpecGap}(K)\right)^n. \tag{2.1}$$

Therefore, lower-bounds on the spectral gap can be used to derive upper-bounds on the mixing time of K. In many examples, the conductance of K is easier to control than the spectral gap. In these examples the concept of s-conductance introduced by L. Lovacz and M. Simonivits ([Lovász and Simonovits(1993)]) as a generalization of the conductance has proven very useful, particularly 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. Plainly put, $\Phi_\zeta(K)$ 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 [Lovász and Simonovits(1993)] Corollary 1.5). One important drawback of the ζ-conductance is that the arguments that relate $\Phi_\zeta(K)$ to the mixing time of K (Theorem 1.4 of [Lovász and Simonovits(1993)]) is rather involved, and this has limited the scope and the usefulness of the concept. Furthermore there are some problems where direct bound on the spectral gap instead of the conductance is easier, or yields better results.

Motivated by the ζ-conductance, we introduce a similar concept of ζ-spectral gap that directly approximates the spectral gap. Let $\|\cdot\|_\star : L^2(\pi) \to [0, \infty]$ denote a norm-like function on $L^2(\pi)$ with the following properties: (i) $\|\alpha f\|_\star = |\alpha|\|f\|_\star$, (ii) if $\|f\|_\star = 0$ then $\mathsf{Var}_\pi(f) = 0$, and (iii)

$$\|Kf\|_\star \leq \|f\|_\star,\quad f \in L_\star^2(\pi), \tag{2.2}$$

where $L_\star^2(\pi) \stackrel{\text{def}}{=} \{f \in L^2(\pi) : \|f\|_\star < \infty\}$. For $\zeta \in (0, 1)$, we define the ζ-spectral gap of K as

$$\mathsf{SpecGap}_\zeta(K) \stackrel{\text{def}}{=} \inf \left\{ \frac{\mathcal{E}(f, f)}{\mathsf{Var}_\pi(f) - \frac{\zeta}{2}},\ f \in L_\star^2(\pi),\ \mathsf{Var}_\pi(f) > \zeta, \text{ and } \|f\|_\star = 1 \right\}. \tag{2.3}$$

We note that $\mathsf{SpecGap}_\zeta(K)$ depends on the choice of $\|\cdot\|_\star$. We note also that if $\zeta = 0$ and $\|f\|_\star = \|f\|_{2,\pi}$, then we recover $\mathsf{SpecGap}_0(K) = \mathsf{SpecGap}(K)$. Furthermore, given $f \in L^2(\pi)$, and writing $\bar f = f - \pi(f)$, we have

$$\frac{\mathcal{E}(f,f)}{\mathsf{Var}_\pi(f) - \frac{\zeta}{2}} = \frac{\pi(\bar f^2) - \left\langle \bar f, P\bar f\right\rangle_\pi}{\pi(\bar f^2) - \frac{\zeta}{2}}.$$

By the lazyness of the chain, $\left\langle \bar f, P\bar f\right\rangle_\pi \geq \pi(\bar f^2)/2$, and we deduce that $\mathsf{SpecGap}_\zeta(K)$ is a quantity that always belongs to the interval $[0,1]$. The idea is somewhat similar to the concept of weak Poincare inequality developed for continuous-time Markov semigroups with zero spectral gap ([Liggett(1991), Cattiaux and Guillin(2009)]). One key difference is that weak Poincare inequalities lead to sub-geometric 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. More precisely, we have the following analog of (2.1). The proof is similar to the proof of (2.1).

Lemma 2.1. *Suppose that K is π-reversible, lazy, and satisfies (2.2). Fix $\zeta \in [0,1)$. Suppose that $\pi_0(\mathrm{d}x) = f_0(x)\pi(\mathrm{d}x)$ for a function $f_0 \in L^2_\star(\pi)$. Then for all integer $n \geq 1$, we have*

$$\|\pi_0 K^n - \pi\|_{\mathrm{tv}}^2 \leq \mathsf{Var}_\pi(K^n f_0) \leq \mathsf{Var}_\pi(f_0)\left(1 - \mathsf{SpecGap}_\zeta(K)\right)^n + \zeta\|f_0\|_\star^2.$$

Proof. See Section 5.1. ■

It is also possible to control similarly the convergence to stationarity in the 1-Wasserstein metric. Indeed, for any $h \in L^2(\pi)$ we have

$$|\pi_0 K^n(h) - \pi(h)| = \left|\int_{\mathcal{X}} h(x)\left(K^n f_0(x) - 1\right)\pi(\mathrm{d}x)\right| \leq \|h\|_{2,\pi}\sqrt{\mathsf{Var}_\pi(K^n f_0)}. \tag{2.4}$$

Hence, if $\mathcal{X}$ is a metric space and π is such that any Lipschitz function h on $\mathcal{X}$ belongs to $L^2(\pi)$ (basically π has finite second moments), then under the assumptions of Lemma 2.1 we have,

$$\begin{aligned}\mathsf{W}_1(\pi_0 K^n, \pi) &\stackrel{\mathrm{def}}{=} \sup_{h:\ \|h\|_{\mathsf{Lip}}=1} |\pi_0 K^n(h) - \pi(h)| \\ &\leq \sup_{h:\ \|h\|_{\mathsf{Lip}}=1} \|h\|_{2,\pi}\sqrt{\mathsf{Var}_\pi(f_0)\left(1 - \mathsf{SpecGap}_\zeta(K)\right)^n + \zeta\|f_0\|_\star^2},\end{aligned} \tag{2.5}$$

where $\|h\|_{\mathsf{Lip}} \stackrel{\mathrm{def}}{=} \sup_{x\neq y}|h(y) - h(x)|/\mathsf{d}(y,x)$ is the Lipschitz norm of h, and where d is the metric on $\mathcal{X}$.

2.1. Illustration with the small local mode example. We now illustrate how the approximate spectral gap can be used with the conceptual example described in the introduction. For that purpose, in this section we assume that $\mathcal{X} = \mathcal{X}_0 \cup (\mathcal{X}_0^c)$ for some measurable subset $\mathcal{X}_0$ of $\mathcal{X}$. We aim to capture the intuition that when $\mathcal{X}_0^c$ is small under π, a Markov chain with transition kernel K started in $\mathcal{X}_0$ typically does not suffer from the local modes in $\mathcal{X}_0^c$.

Let $\mathcal{B}_{\mathcal{X}_0}$ be the trace sigma-algebra of $\mathcal{B}$ on $\mathcal{X}_0$. Let $K_{\mathcal{X}_0}$ be the restriction of K on $\mathcal{X}_0$. That is $K_{\mathcal{X}_0}$ is the transition kernel on $(\mathcal{X}_0, \mathcal{B}_{\mathcal{X}_0})$ defined as

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

Using the reversibility of K, 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

$$\mathsf{SpecGap}_{\mathcal{X}_0}(K) \stackrel{\text{def}}{=} \inf\left\{\frac{1}{2}\frac{\int_{\mathcal{X}_0}\int_{\mathcal{X}_0}\pi(\mathrm{d}x)K(x,\mathrm{d}y)(f(y)-f(x))^2}{\frac{1}{2}\int_{\mathcal{X}_0}\int_{\mathcal{X}_0}\pi(\mathrm{d}x)\pi(\mathrm{d}y)(f(y)-f(x))^2},\ f:\ \mathcal{X}\to\mathbb{R}\right\}, \tag{2.6}$$

where the infimum is taken over all functions $f \in L^2_\star(\pi)$ such that $\int_{\mathcal{X}_0}\int_{\mathcal{X}_0}\pi(\mathrm{d}x)\pi(\mathrm{d}y)(f(y)-f(x))^2 > 0$. The next result shows that the spectral gap of $K_{\mathcal{X}_0}$ is a lower bound for $\mathsf{SpecGap}_\zeta(K)$.

Lemma 2.2. *For $\zeta \in (0,1)$, and $\|\cdot\|_\star = \|\cdot\|_{m,\pi}$, for some $m \in (2,+\infty]$, if $\pi(\mathcal{X}_0) \geq 1 - \left(\frac{\zeta}{10}\right)^{1+\frac{2}{m-2}}$ then we have*

$$\mathsf{SpecGap}_\zeta(K) \geq \mathsf{SpecGap}_{\mathcal{X}_0}(K).$$

Proof. See Section 5.2. ■

Fix $\zeta_0 \in (0,1)$. Suppose that we choose the initial distribution π_0 such that $\|f_0\|_{m,\pi} \leq B$, for some constant $B \geq 1$. In that case Lemma 2.1 with $\|\cdot\|_\star = \|\cdot\|_{m,\pi}$, and $\zeta = \zeta_0^2/(B^2)$ gives for all $n \geq 1$,

$$\|\pi_0 K^n - \pi\|_{\mathrm{tv}}^2 \leq B^2\left(1 - \mathsf{SpecGap}_\zeta(K)\right)^n + \zeta_0^2. \tag{2.7}$$

Therefore, if $\pi(\mathcal{X}_0) \geq 1 - \left(\frac{\zeta}{10}\right)^{1+\frac{2}{m-2}}$, by Lemma 2.2 we obtain the following bound on the mixing time:

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

In other words the mixing time of K can indeed be controlled by the spectral gap of $K_{\mathcal{X}_0}$. The condition $\pi(\mathcal{X}_0) \geq 1 - \left(\frac{\zeta_0^2}{10B^2}\right)^{1+\frac{2}{m-2}}$ puts a stringent 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. Further illustrations are given below.

2.2. Extension to reversible Markov semigroups. The idea can also be applied to continuous-time Markov processes. We refer the reader to ([Bakry et al.(2013)]) for an introduction to Markov semigroups. We consider a reversible Markov semigroup $K = \{K_t,\ t \geq 0\}$, where for each t, K_t is a Markov kernel on $(\mathcal{X}, \mathcal{B})$ that is reversible with respect to π. Let G denote the generator of the semi-group that we assumed well-defined on a dense subspace $\mathcal{A}$ of $L^2(\pi)$ that is stable under G and K_t such that for all $t \geq 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}K_t f = K_t G f = G K_t f, \quad f \in \mathcal{A}. \tag{2.8}$$

We make also the assumption that the domain $\mathcal{A}$ contains constant functions and is equipped with a norm $\|\cdot\|_\star$ such that $\|f\|_\star = 0$ implies that $\mathsf{Var}_\pi(f) = 0$, and for all $t \geq 0$

$$\|K_t f\|_\star \leq \|f\|_\star, \quad f \in \mathcal{A}. \tag{2.9}$$

The Dirichlet form of K is defined as

$$\mathcal{E}(f,f) \stackrel{\text{def}}{=} -\int_{\mathcal{X}} f(x)Gf(x)\pi(\mathrm{d}x).$$

For $\zeta \in [0,1)$, we can define the ζ-spectral gap of the semi-group K as

$$\lambda_\zeta(K) \stackrel{\text{def}}{=} \inf\left\{\frac{-\int_{\mathcal{X}} f(x)Gf(x)\pi(\mathrm{d}x)}{\mathsf{Var}_\pi(f) - \zeta},\ \ f \in \mathcal{A},\ \mathsf{Var}_\pi(f) > \zeta,\ \text{and } \|f\|_\star = 1\right\}. \tag{2.10}$$

We have the analog of Lemma 2.1.

Lemma 2.3. *Suppose that the semigroup K satisfies (2.9). Let $\nu(\mathrm{d}x) = f(x)\pi(\mathrm{d}x)$ be a probability measure on $\mathcal{X}$, where $f \in \mathcal{A}$. Let $\zeta \in [0,1)$ be such that $\lambda_\zeta(K) > 0$. Then for all $t \geq 0$ we have*

$$\|\nu K_t - \pi\|_{\mathrm{tv}}^2 \leq \mathsf{Var}_\pi(K_t f) \leq \mathsf{Var}_\pi(f) e^{-2\lambda_\zeta(K)t} + \zeta\|f\|_\star^2.$$

Proof. See Section 5.3. ∎

For $\zeta = 0$, $\lambda_\zeta(K)$ corresponds to the classical spectral gap of the semigroup and Lemma 2.3 is the classical exponential convergence of the semigroup. This result can be applied to Langevin diffusion processes. Suppose that $\mathcal{X} = \mathbb{R}^p$ equipped with the Lebesgue measure, and $\pi(\mathrm{d}x) = e^{-U(x)}/Z$, for a function $U : \mathbb{R}^p \to \mathbb{R}$ that is differentiable with Lipschitz gradient. The Langevin diffusion process for π defines a reversible Markov semigroup with invariant distribution π. The convergence rate of the semigroup toward π is a key ingredient in the analysis of several recent MCMC algorithms, including the unadjusted Langevin algorithm and stochastic gradient Langevin dynamics ([Welling and Teh(2011), Raginsky et al.(2017)]). When U is convex, the semigroup is known to possess a spectral gap ([Bobkov(1999)]). Various extensions beyond the convex case are also known and are well discussed in ([Bakry et al.(2008)]). Lemma 2.3 offers another route, one that might be more effective when a good initial distribution is available, and π has well-understood concentration properties. We leave the details as possible future research.

3. Application: mixing times of mixtures of Markov kernels.

To illustrate Lemma 2.1 we consider here the case where $\mathcal{X} = \mathbb{R}^p$, 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 finite set, and for $i \in \mathsf{I}$, $\pi(i,\cdot) : \mathbb{R}^p \to [0,\infty)$ is a measurable function. As mentioned in the introduction, much is known on the computational complexity of various MCMC algorithms for log-concave densities. However these results cannot be directly applied to mixtures, since for instance a mixture of log-concave densities is not log-concave in general. Sampling from mixtures is more challenging than sampling from log-concave densities. For instance it is shown in [Ge et al.(2018)] that no polynomial-time MCMC algorithm exists to sample from mixtures of densities with inequal covariance matrix, if the algorithm uses only the marginal density of the mixture and its derivative. However this result does not cover the most commonly used strategy to deal with mixtures, namely the Gibbs sampler.

Gibbs sampling type algorithms 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 assume that K_i is reversible with respect to π_i, and ergodic (phi-irreducible and aperiodic). We then consider the Markov kernel K defined as

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

that is reversible with respect to π as in (3.1). In [Madras and Randall(2002)] 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 graph on I such that whenever there is an edge between $i, j \in \mathsf{I}$, it holds

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

If $D(\mathsf{I})$ denotes the diameter of the graph thus defined[1], Theorem 1.2 of [Madras and Randall(2002)] says that

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

The lower bound in (3.5) can be very small when I is large, particularly if some $\pi(i)$ are exponentially small. We combine the approach in ([Madras and Randall(2002)]) with the canonical path argument of ([Sinclair(1992), Diaconis and Stroock(1991)]) 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 property.*

1. *For each $i \in \mathsf{I}_0$, $\pi_i(\mathsf{B}_i) \geq 1/2$.*
2. *There exist $\kappa > 0$ and a connected graph $\mathcal{G}$ on I_0 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 there is an edge in $\mathcal{G}$ between i and j.

[1]The diameter of a graph is the length (the number of edges) of the longest among all the shortest paths between all pairs of vertices.

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 $\bar\pi$. The graph $\mathcal{G}$ captures the proximity between the conditional distributions. 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 $2(1-\kappa)$.

Since $\mathcal{G}$ is assumed connected, for any distinct pair $i,j\in\mathsf{I}_0$ we can find and pick a path γ_{ij} that connects i and j. We call γ_{ij} the canonical path from i to j. The number of edges on γ_{ij} is denoted $|\gamma_{ij}|$. We then define

$$\mathsf{m}_1 \stackrel{\text{def}}{=} \max_{\iota\in\mathsf{I}_0} \sum_{i,j\in\mathsf{I}_0:\ \gamma_{ij}\ni\iota} |\gamma_{ij}|\frac{\pi(i)\pi(j)}{\pi(\iota)}, \tag{3.7}$$

where the summation is taken over all distinct pair (i,j) whose canonical path γ_{ij} goes through node ι. We define the local spectral gap of K_i as $\mathsf{SpecGap}_i(K_i) = \mathsf{SpecGap}_{\mathsf{B}_i}(K_i)$, where $\mathsf{SpecGap}_{\mathsf{B}_i}(K_i)$ is defined as in (2.6).

Theorem 3.1. *Let π as in (3.1), and K as in (3.3). Assume that H1 holds and K satisfies (2.2) with some chosen pseudo-norm $\|\cdot\|_\star$. Set $\bar{\mathsf{B}} \stackrel{\text{def}}{=} \cup_{i\in I_0}\{i\}\times\mathsf{B}_i$ and assume that there exists $\zeta\in[0,1)$ such that for any function $f\in L^2_\star(\pi)$ satisfying $\|f\|_\star = 1$, it holds*

$$\begin{aligned} 2\int_{\bar{\mathsf{B}}}\int_{\bar{\mathsf{B}}^c} \left(f(y)-f(x)\right)^2 \bar\pi(\mathrm{d}i,\mathrm{d}x)\bar\pi(\mathrm{d}j,\mathrm{d}y) \\ + \int_{\bar{\mathsf{B}}^c}\int_{\bar{\mathsf{B}}^c} \left(f(y)-f(x)\right)^2 \bar\pi(\mathrm{d}i,\mathrm{d}x)\bar\pi(\mathrm{d}j,\mathrm{d}y) \le \zeta, \end{aligned} \tag{3.8}$$

where $\bar{\mathsf{B}}^c \stackrel{\text{def}}{=} (I\times\mathcal{X})\setminus\bar{\mathsf{B}}$. Then

$$\mathit{SpecGap}_\zeta(K) \ge \left(\frac{\kappa}{1+8\mathsf{m}_1}\right)\min_{i\in I_0}\mathit{SpecGap}_i(K_i). \tag{3.9}$$

Proof. See Section 5.4. ∎

Remark 3.2. The condition (3.8) can be easily handled. For instance if $\|\cdot\|_\star = \|\cdot\|_{\pi,m}$ for some $m\in(2,\infty]$, then by Holder's inequality the left hand side of (3.8) is easily bounded from above by $10\bar\pi(\bar{\mathsf{B}}^c)^{1-2/m}$. In that case (3.8) holds if $\bar{\mathsf{B}}$ satisfies $\bar\pi(\bar{\mathsf{B}})\ge 1-(\zeta/10)^{1+2/(m-2)}$.

Note that the constant m_1 satisfies

$$\mathsf{m}_1 \le \frac{\mathsf{D}(\mathsf{I}_0)}{\min_{i\in\mathsf{I}_0}\pi(i)}. \tag{3.10}$$

Hence the bound in (3.9) improves on (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. Theorem 3.1 can be extended to such settings. However we will not pursue this here for lack of space.

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 ([George and McCulloch(1997), Narisetty and He(2014)]). 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 $\Delta_s \stackrel{\text{def}}{=} \{\delta \in \Delta : \|\delta\|_0 \le s\}$, 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 j-th diagonal element equal to ρ_1^{-1} if $\delta_j = 1$, and ρ_0^{-1} if $\delta_j = 0$. Note that we can always set $s = p$. The regression error σ is assumed known. This model is very popular in the applications. Indeed, the posterior conditional distribution $\Pi(\delta|\theta, z)$ is a product of independent Bernoulli distributions constrained to be s-sparse:

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

where $A \stackrel{\text{def}}{=} (1-\mathsf{q})\mathsf{q}^{-1}\sqrt{\rho_0/\rho_1}$. 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). Given δ, 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 yields a simple Gibbs sampling algorithm for (4.1). We consider the following version that is modified so that the resulting Markov chain is lazy as required by our theory.

[Algorithm 4] For some initial distribution ν_0 on $\mathbb{R}^p$, draw $u_0 \sim \nu_0$. Given $u_0, \ldots, u_k$ for some $k \ge 0$, draw independently $I_{k+1} \sim \mathsf{Ber}(0.5)$.

1. If $I_{k+1} = 0$, set $u_{k+1} = u_k$.
2. If $I_{k+1} = 1$,
   (a) Draw $\delta \sim \Pi(\cdot|u_k, z)$ as given in (4.2), and
   (b) draw $u_{k+1} \sim \mathbf{N}_p(m_\delta, \sigma^2\Sigma_\delta)$ as given in (4.3).

We analyze the mixing time of the marginal chain $\{u_k,\ k \geq 0\}$ from Algorithm 4. As easily seen, $\{u_k,\ k \geq 0\}$ is a Markov chain with invariant distribution

$$\Pi(\mathrm{d}\theta|z) \propto \sum_{\delta\in\Delta_s} \left(\frac{\mathsf{q}}{1-\mathsf{q}}\right)^{\|\delta\|_0} \frac{e^{-\frac{1}{2}\theta' D_{(\delta)}^{-1}\theta}}{\sqrt{\det\left(2\pi D_{(\delta)}\right)}} e^{-\frac{1}{2\sigma^2}\|z-X\theta\|_2^2}\mathrm{d}\theta, \tag{4.4}$$

which is of the form (3.1), and with transition kernel

$$K(u,\mathrm{d}\theta) \stackrel{\mathrm{def}}{=} \sum_{\omega\in\Delta} \Pi(\omega|u,z)\left[\frac{1}{2}\delta_u(\mathrm{d}\theta) + \frac{1}{2}\Pi(\mathrm{d}\theta|\omega,z)\right], \tag{4.5}$$

which is of the form (3.3).

To proceed we introduce some notations. For $\delta \in \Delta$, and $\theta \in \mathbb{R}^p$, we write θ_δ as a short for the component-wise product of θ and δ, and we define $\delta^c \stackrel{\mathrm{def}}{=} 1-\delta$, that is $\delta_j^c = 1-\delta_j$, $1 \leq j \leq p$. 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 j-th 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 $\delta_j = 1$ 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

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

and the coherence of X defined as

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

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, then and provided that $n \geq As^2\log(p)$, it holds $\mathcal{C}(s) \leq c$ for some absolute constants c, A. We refer the reader to the Appendix for details. We make the following regularity assumption on the matrix X.

H2. 1. *The matrix X is non-random and normalized such that*

$$\|X_j\|_2^2 = n, \quad j = 1,\ldots,p. \tag{4.6}$$

Furthermore, there exists an integer $s_0 \in \{1,\ldots,p-1\}$, such that

$$\lambda \stackrel{\mathrm{def}}{=} \min_{\delta:\ \|\delta\|_0\leq s_0}\ \inf\left\{\frac{v'\left(X_{\delta^c}' L_\delta^{-1} X_{\delta^c}\right)v}{n\|v\|_2^2},\ v\in\mathbb{R}^{p-\|\delta\|_0},\ 0<\|v\|_0\leq s_0\right\} > 0.$$

Remark 4.1. The matrix L_δ^{-1} can be loosely interpreted as the projector on the orthogonal of the space spanned by the columns of X_δ. Therefore, H2 rules out settings where a small number of columns of X have the same column span as the column span of X. Indeed signal recovery becomes nearly impossible in such settings. In can be shown that if X is a random matrix with i.i.d. standard Gaussian entries then $\lambda > 0$ for s_0 of order $n/\log(p)$. We refer the reader to the Appendix for details. □

We also make some very mild assumptions pertaining to the prior parameters and to the existence of a true model.

H3. 1. *There exists a true value of the parameter $\theta_\star \in \mathbb{R}^p$ 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 prior parameters ρ_0, ρ_1 satisfy*

$$0 < \rho_1 < \rho_0, \quad \sigma^2\rho_1 \leq \left(1 - \frac{\rho_1}{\rho_0}\right) n, \quad \textit{and} \quad \sqrt{1 + \frac{ns}{\sigma^2\rho_1}} \leq p^a, \tag{4.8}$$

for some absolute constant $a > 0$.

The last two parts of Condition (4.8) are easily satisfied and are imposed mostly to obtain simple mathematical formulas. For some constant $c_0 > 0$, we introduce the event

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

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 4.2. *Suppose that H2-H3 hold. Fix $\zeta_0 \in (0,1)$. Suppose that s, the sparsity level of the posterior distribution (4.1) is chosen such that $0 < s \leq s_0$ with s_0 as in H2, and Algorithm 4 is initialized from $\nu_0 = \Pi(\cdot|\delta^{(i)}, z)$, for some arbitrary $\delta^{(i)} \in \Delta_s$. Take $z \in \mathcal{E}_0$, suppose that we choose u large enough such that*

$$u > 2\max\left(2, \frac{\varrho}{\lambda}\right), \quad \textit{where} \quad \varrho \stackrel{\text{def}}{=} \left(\sigma\sqrt{c_0} + \|\theta_\star\|_1 \mathcal{C}(s)\right)^2, \tag{4.9}$$

and the sample size n satisfies

$$n \geq \frac{A_0 u \sigma^2 s_\star \log(p)}{\lambda^2 \underline{\theta}_\star^2}, \quad \textit{where} \quad \underline{\theta}_\star \stackrel{\text{def}}{=} \min_{j:\, \delta_{\star j}=1} |\theta_{\star j}|, \tag{4.10}$$

for some absolute constant A_0. Set

$$\lambda_1 \stackrel{\text{def}}{=} \min_{1\leq j\leq p} \min_{\delta\in\Delta_s} \frac{X_j' L_\delta^{-1} X_j}{n}.$$

Then there exists a constant A_1 that does not depend on n, p nor ζ_0 such that for all

$$(4.11)\quad N \geq A_1\, s\left[\log\left(\frac{1}{\zeta_0}\right) + \frac{su(1+\|\theta_\star\|_\infty^2)n}{\sigma^2\lambda}\right] \times \max\left(1, \sqrt{\frac{n}{\sigma^2\rho_0}}\right) \times \max\left(1, e^{\frac{1}{4\sigma^2}(8\sigma^2\rho_0 - n\lambda_1)}\right) \times p^{\frac{\rho_0}{n}\frac{2\varrho}{\lambda_1^2}},$$

we have

$$\|\nu_0 K^N - \Pi(\cdot|z)\|_{\mathrm{tv}} \leq \zeta_0.$$

Proof. See Section 5.5. ■

We note that our condition (4.9) is analogous to Condition C of [Yang et al.(2016)]. The main term in the bound (4.11) is

$$\max\left(1, e^{\frac{1}{4\sigma^2}(8\sigma^2\rho_0 - n\lambda_1)}\right) p^{\frac{\rho_0}{n}\frac{2\varrho}{\lambda_1^2}},$$

which highlights the important impact of the prior parameter ρ_0 on the mixing of the algorithm. If ρ_0 is chosen as $\rho_0 \leq n\lambda_1/(8\sigma^2)$, then by (4.11), the mixing time scales as $O(p^{\varrho/\lambda_1})$. Note that the ratio ϱ/λ_1 depends mainly on the correlation between the columns of X. Our simulation results indeed confirm that dependence of the mixing time on X, however the polynomial scaling $O(p^{\varrho/\lambda_1})$ predicted by the theorem may be conservative.

In contrast, if $\rho_0 > n\lambda_1/(8\sigma^2)$ the bound predicts a mixing time that scales as $O(e^{2\rho_0} p^{\frac{\rho_0}{n}\frac{2\varrho}{\lambda_1^2}})$, which is worst than $O(e^n p^{\varrho/\lambda_1})$. This said, it is important to add that (4.11) is an upper bound on the mixing time which may not be tight, and as such does not prove slow mixing.

We contrast these findings with the posterior contraction properties of the posterior distribution. According to [Narisetty and He(2014)], as $n, p \to \infty$, we need to let ρ_0 grow faster than n, and let ρ_1 be of order n/p^2 in order to guarantee posterior contraction of Π. And in their simulation section these authors suggest using $\rho_0 = 10n/\sigma^2$ (although it is unclear whether posterior contraction holds in that regime). In these regimes our results suggest that the mixing time of Algorithm 4 grows faster than $O(e^n p^{\varrho/\lambda_1})$. This description matches well with our numerical experiments. But again (4.11) is only an upper bound on the mixing time, and as such does not establish slow mixing.

Note that when posterior contraction holds the posterior distribution assigns increasingly small probability to $\{\delta :\ \delta \not\supseteq \delta_\star\}$. Hence a chain that starts in $\{\delta :\ \delta \supseteq \delta_\star\}$ may have markedly different mixing time than what is predicted by Theorem 4.2. To formalize this, we shall focus on the unconstrained case where $s = p$ in (4.1). 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

$$\mathcal{E} \stackrel{\text{def}}{=} \Big\{ z \in \mathbb{R}^n : \quad \Pi(\mathcal{D}_k|z) \geq 1 - \frac{1}{p^{\frac{u}{2}(k+1)}}, \quad \text{for all } k \geq 0,$$
$$\text{and} \quad \max_{\delta \supseteq \tilde{\delta}_\star :\, \|\delta\|_0 \leq s_0} \ \sup_{1 \leq j \leq p} \frac{1}{\sigma} \left| \left\langle L_\delta^{-1} X_j, z - X\theta_\star \right\rangle \right| \leq \sqrt{c_0 n \log(p)} \Big\},$$

for some constant c_0. 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 similar assumptions as above, [Narisetty and He(2014)] show 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 [Atchade and Bhattacharyya(2018)] shows that $z \in \mathcal{E}$ with high probabiity for a slightly modified version of the posterior distribution (4.1).

Theorem 4.3. *Assume H2-H3 and $s = p$ in (4.1). Fix $\zeta_0 \in (0,1)$. Suppose that Algorithm 4 is initialized from $\nu_0 = \Pi(\cdot|\delta^{(i)}, z)$, for some $\delta^{(\mathrm{i})} \in \mathcal{D}_{(s_0 - s_\star)}$ such that $\mathsf{FP} \stackrel{\text{def}}{=} \|\delta^{(\mathrm{i})}\|_0 - s_\star$ satisfies*

$$\mathsf{FP} \leq \frac{u}{4(u+a)}(k+1) + \frac{\log\left(\frac{80}{\zeta_0^2}\right)}{2(u+a)\log(p)}, \tag{4.12}$$

for some integer $k \leq s_0 - s_\star$. Suppose also that (4.9) and (4.10) hold. Then there exists a constant A that does not depend on n, p nor ζ_0 such that for all $z \in \mathcal{E}$, and all

$$N \geq A\, \mathsf{FP} \left[\log\left(\zeta_0^{-1}\right) + \mathsf{FP} u \log(p)\right] p^{\frac{2\rho_0}{n} \frac{\varrho}{\lambda_1^2}}, \tag{4.13}$$

we have

$$\|\nu_0 K^N - \Pi(\cdot|z)\|_{\mathrm{tv}} \leq \zeta_0.$$

Proof. See Section 5.6. ■

Condition (4.12) restricts the number of false-positives of the initial model $\delta^{(\mathrm{i})}$ compared to s_0. 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 4.3 suggests that when posterior contraction holds ($z \in \mathcal{E}$), the mixing time of Algorithm 4 with a good initialization is less sensitive to large values of ρ_0 (the term $e^{\frac{1}{4\sigma^2}(8\sigma^2\rho_0 - n\lambda_1)}$ no longer appear in (4.13)). For instance with $\rho_0 = n\lambda_1/2$ the mixing time is at most $O(\mathsf{FP}^2 p^{\varrho/\lambda_1})$, which is better $O(e^n p^{\varrho/\lambda_1})$.

One clear roadblock toward the practical use of this result is finding the initial $\delta^{(\mathrm{i})}$ such that $\delta^{(\mathrm{i})} \supseteq \delta_\star$. 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 (similar to H2). We refer the reader for instance to ([Meinshausen and Yu(2009)]).

One of the first paper that analyzes the mixing times of MCMC algorithm in high-dimensional linear regression models and highlights fast/slow mixing behaviors is [Yang et al.(2016)].

Their posterior distribution is slightly different from what we looked at in this work. Specifically [Yang et al.(2016)] applied a Metropolized-Gibbs sampler to the marginal distribution of δ, whereas we consider here a Gibbs sampler applied to the joint distribution of (δ, 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 $O(s^2 np \log(p))$. To the extent that our bound in Theorem 4.2 is tight, the better rate obtained by these authors can perhaps be interpretated as the positive effect of marginalization and collapsing in Gibbs sampling ([Liu(1994)]).

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 dimension $p \in \{500, 1000, 2000, 3000, 4000\}$. 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 $\Sigma_{ij} = 0.9^{|j-i|}$. After sampling, we normalized the columns of X to each have norm $\sqrt{n}$. We fix the number of non-zero 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 non-zero 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 matches the recommendations of [Narisetty and He(2014)] to get posterior contraction of $\Pi(\cdot|z)$. We use an initial distribution $\nu_0 = \Pi(\cdot|\delta^{(\mathrm{i})}, z)$, where $\delta^{(\mathrm{i})}$ is such that $\|\delta^{(\mathrm{i})} - \delta_\star\|_0 = 2p/10$, with two scenarios. A scenario FN (false negative), where 5 out of 10 of the true positive of $\delta_\star$ are set to 0, and a scenario $\mathsf{no\ FN}$, where $\delta^{(\mathrm{i})}$ 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 20000 iterations. For the sampler of [Yang et al.(2016)], we stop any run that has not mixed by 10^5 iterations. The average empirical mixing time thus obtained (based on 50 independent MCMC replications) are presented in Table 1 and Table 2.

We can make the following observations.

		$p = 500$	$p = 1000$	$p = 2000$	$p = 3000$	$p = 4000$
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.	$5,191.0(1,503)$	$11,975.9(2,769)$	$26,877.8(4,786)$	$42,285.7(8,721)$	$56,264.3(10,362)$

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.

		$p = 500$	$p = 1000$	$p = 2000$	$p = 3000$	$p = 4000$
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.	$> 98,067$	$> 87,424$	$> 73,253$	$> 77,902$	$> 82,205$

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.

1. There is sharp difference in behavior between the low and high coherence settings.
2. As predicted by our theory, Algorithm 4 mixes better when there is no false-negative in the initialization. The algorithm of [Yang et al.(2016)] seems impervious to the initialization. It should be noted in comparing the two algorithms, that an iteration of the algorithm of [Yang et al.(2016)] costs roughly p times less than an iteration of Algorithm 4.
3. The third observation that can be drawn from the results is that when there are false-negatives, Algorithm 4 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 [Yang et al.(2016)] does not fully describe the behavior of their algorithm, as we see a significant degradation of performance in their algorithm with high coherence design matrices, which cannot be clearly explained by their result.

Overall, based on our theoretical analysis and the simulation study, our recommendation when using Algorithm 4 is to set $\rho_0 = n/\sigma^2$, and to the extent possible to use the lasso sparsity structure as initialization (or some other similar high-dimensional frequentist estimator).

5. Proofs. The proof of Theorem 3.1 relies on the following lemma due to [Madras and Randall(2002)]. For a proof see their inequality (47). A direct argument by coupling can also be easily constructed.

Lemma 5.1. *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*

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

5.1. Proof of Lemma 2.1. We first note that if a probability measure ν is absolutely continuous with respect to π with Radon-Nikodym derivative f_ν, then for any $A \in \mathcal{B}$,

$$\begin{aligned} \nu K(A) &= \int \nu(\mathrm{d}x)K(x, A) = \int\int f_\nu(x)\mathbf{1}_A(y)\pi(\mathrm{d}x)K(x,\mathrm{d}y) \\ &= \int\int \mathbf{1}_A(x) f_\nu(y)\pi(\mathrm{d}x)K(x,\mathrm{d}y) = \int_A \pi(\mathrm{d}x)\int K(x,\mathrm{d}y)f_\nu(y), \end{aligned}$$

where the third equality uses the reversibility of K. This calculation says that νK is also absolutely continuous with respect to π with Radon-Nikodym derivative $x \mapsto Kf_\nu(x) \stackrel{\mathrm{def}}{=} \int K(x,\mathrm{d}y)f_\nu(y)$. More generally $\frac{\mathrm{d}(\nu K^n)}{\mathrm{d}\pi}(\cdot) = K^n f_\nu(\cdot)$, and

$$\begin{aligned} \|\nu K^n - \pi\|_{\mathrm{tv}}^2 &= \left(\int \left|\frac{\mathrm{d}(\nu K^n)}{\mathrm{d}\pi}(x) - 1\right| \pi(\mathrm{d}x)\right)^2 \\ &= \left(\int |K^n f_\nu(x) - 1|\, \pi(\mathrm{d}x)\right)^2 \\ &\leq \|K^n f_\nu - 1\|_{2,\pi}^2 \\ &= \mathsf{Var}_\pi(K^n f_\nu). \end{aligned} \tag{5.1}$$

Take $f \in L^2(\pi)$. Since $\pi(f) = \pi(Kf)$, we have

$$\mathsf{Var}_\pi(Kf) - \mathsf{Var}_\pi(f) = \langle Kf, Kf\rangle_\pi - \langle f, f\rangle_\pi = -\frac{1}{2}\int\int (f(y) - f(x))^2\, \pi(\mathrm{d}x)K^2(x,\mathrm{d}y), \tag{5.2}$$

where the last equality exploits the reversibility of K. By the lazyness of K we have

$$\int\int (f(y) - f(x))^2\, \pi(\mathrm{d}x)K^2(x,\mathrm{d}y) \geq \int\int (f(y) - f(x))^2\, \pi(\mathrm{d}x)K(x,\mathrm{d}y).$$

A proof of this statement is given for instance in [Montenegro and Tetali(2006)] (Equation 2.12). Using the last display together with (5.2), and the definition of $\mathcal{E}(f,f)$, we conclude that for all $f \in L^2(\pi)$,

$$\mathsf{Var}_\pi(Kf) \le \mathsf{Var}_\pi(f) - \mathcal{E}(f,f). \tag{5.3}$$

Fix $\zeta \in (0,1)$, and take $f \in L^2_\star(\pi)$. Suppose that $\|f\|_\star > 0$. If $\mathsf{Var}_\pi(f) \le \zeta\|f\|_\star^2$, then, by (5.3), $\mathsf{Var}_\pi(Kf) \le \min(\mathsf{Var}_\pi(f), \zeta\|f\|_\star^2)$. But if $\mathsf{Var}_\pi(f) > \zeta\|f\|_\star^2$, then by (5.3),

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

Note also that if $\|f\|_\star = 0$, then $\mathsf{Var}_\pi(f) = 0$ by the listed properties of $\|\cdot\|_\star$, and $\mathsf{Var}_\pi(Kf) = 0$ by (5.3), so that the last display continue to hold. We conclude that for all $f \in L^2_\star(\pi)$,

$$\mathsf{Var}_\pi(Kf) \le \mathsf{Var}_\pi(f)\left(1 - \mathsf{SpecGap}_\zeta(K)\right) + \zeta\|f\|_\star^2 \mathsf{SpecGap}_\zeta(K).$$

Given that $Kf \in L^2_\star(\pi)$ for all $f \in L^2_\star(\pi)$, we can iterate the above inequality to deduce that for all $f \in L^2_\star(\pi)$, and for all $n \ge 1$,

$$\begin{aligned}
\mathsf{Var}_\pi(K^n f) \le \mathsf{Var}_\pi(f)\left(1 - \mathsf{SpecGap}_\zeta(K)\right)^n \\
+ \zeta \mathsf{SpecGap}_\zeta(K) \sum_{j \ge 0} \left(1 - \mathsf{SpecGap}_\zeta(K)\right)^j \|K^{n-j-1} f\|_\star^2 \\
\le \mathsf{Var}_\pi(f)\left(1 - \mathsf{SpecGap}_\zeta(K)\right)^n + \zeta\|f\|_\star^2.
\end{aligned}$$

Now, if $\pi_0 = f_0\pi$, the last display combined with (5.1) implies that

$$\|\pi_0 K^n - \pi\|_{\mathrm{tv}}^2 \le \mathsf{Var}_\pi(K^n f_0) \le \mathsf{Var}(f_0)\left(1 - \mathsf{SpecGap}_\zeta(K)\right)^n + \zeta\|f_0\|_\star^2,$$

as claimed. □

5.2. Proof Lemma 2.2. Take $f : \mathcal{X} \to \mathbb{R}$ such that $\mathsf{Var}_\pi(f) > \zeta$, and $\|f\|_\star = \|f\|_{m,\pi} = 1$. We have

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

Using the convexity inequality $(a+b)^2 \leq 2a^2 + 2b^2$, and Holder's inequality,

$$\int_{\mathcal{X}_0}\int_{\mathcal{X}\setminus\mathcal{X}_0} (f(y)-f(x))^2\pi(\mathrm{d}x)\pi(\mathrm{d}y)$$
$$\leq 2\pi(\mathcal{X}_0)\int_{\mathcal{X}\setminus\mathcal{X}_0} f(x)^2\pi(\mathrm{d}x) + 2\pi(\mathcal{X}\setminus\mathcal{X}_0)\int_{\mathcal{X}_0} f(x)^2\pi(\mathrm{d}x)$$
$$\leq 2\pi(\mathcal{X}_0)\pi(\mathcal{X}\setminus\mathcal{X}_0)^{1-\frac{2}{m}}\|f\|^2_{m,\pi} + 2\pi(\mathcal{X}\setminus\mathcal{X}_0)\|f\|^2_{m,\pi} \leq 4\pi(\mathcal{X}\setminus\mathcal{X}_0)^{1-\frac{2}{m}}.$$

With similar calculation,

$$\int_{\mathcal{X}\setminus\mathcal{X}0}\int_{\mathcal{X}\setminus\mathcal{X}_0} (f(y)-f(x))^2\pi(\mathrm{d}x)\pi(\mathrm{d}y) \leq 4\pi(\mathcal{X}\setminus\mathcal{X}_0)\pi(\mathcal{X}\setminus\mathcal{X}_0)^{1-\frac{2}{m}} \leq 2\pi(\mathcal{X}\setminus\mathcal{X}_0)^{1-\frac{2}{m}}.$$

Using $\pi(\mathcal{X}_0) \geq (\zeta/10)^{1+2/(m-2)}$, we get

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

Hence

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

The statement bound easily follows. □

5.3. Proof of Lemma 2.3. Take $f \in \mathcal{A}$. Without any loss of generality we assume that $\pi(f) = 0$. Then

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{Var}_\pi(K_t f) = \frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{X}} (K_t f)^2(x)\pi(\mathrm{d}x) = 2\int_{\mathcal{X}} K_t f(x) G K_t f(x)\pi(\mathrm{d}x). \tag{5.4}$$

Suppose that $\|K_t f\|_\star > 0$. If $\mathsf{Var}_\pi(K_t f/\|K_t f\|_\star) > \zeta$, then from (5.4) and the definition of $\lambda_\zeta(K)$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{Var}_\pi(K_t f) &\leq -2\|K_t f\|^2_\star \lambda_\zeta(K)\left(\mathsf{Var}_\pi\left(\frac{K_t f}{\|K_t f\|_\star}\right) - \zeta\right)\\ &\leq -2\lambda_\zeta(K)\mathsf{Var}_\pi(K_t f) + 2\zeta\lambda_\zeta(K)\|K_t f\|^2_\star.\end{aligned} \tag{5.5}$$

However, if $\mathsf{Var}_\pi(K_t f/\|K_t f\|_\star) \leq \zeta$, we see that the right-hand side of (5.5) is nonnegative, whereas from (5.4) and the properties of the generator we see that the left-hand side of (5.5) is nonpositive. Note also that (5.5) continue to hold when $\|K_t f\|_\star = 0$. Hence for all $f \in \mathcal{A}$, and for all $t \geq 0$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{Var}_\pi(K_t f) \leq -2\lambda_\zeta(K)\mathsf{Var}_\pi(K_t f) + 2\zeta\lambda_\zeta(K)\|f\|^2_\star. \tag{5.6}$$

The lemma then follows from Gronwall's lemma. More precisely, set $\alpha = \zeta\|f\|_\star^2$, $\beta = 2\lambda_\zeta(K)$, and $u(t) = \mathsf{Var}_\pi(K_t f)$. Hence (5.6) reads $u'(t) \leq -\beta u(t) + \alpha\beta$. Setting $v(t) = e^{-\beta t}$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{u(t)}{v(t)}\right) = \frac{u'(t)v(t) - v'(t)u(t)}{v(t)^2} = \frac{u'(t) + \beta u(t)}{v(t)} \leq \alpha\beta e^{\beta t}.$$

Integrating both sides yields the stated bound. □

5.4. Proof of Theorem 3.1.

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

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

Using $\bar{\mathsf{B}} = \cup_{i\in\mathsf{I}_0}\{i\}\times\mathsf{B}_i$, and $\bar{\mathsf{B}}^c \stackrel{\text{def}}{=} (\mathsf{I}\times\mathcal{X})\setminus\bar{\mathsf{B}}$, we have,

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

For $\bar{\mathsf{B}}$ as in (3.8), and expanding the first term on the right hand side of (5.8) it follows that

$$\begin{aligned}
(5.9)\quad 2\left(\mathsf{Var}_\pi(f) - \frac{\zeta}{2}\right) &\leq \sum_{i\in\mathsf{I}_0} \pi(i)^2\int_{\mathsf{B}_i}\int_{\mathsf{B}_i} (f(y) - f(x))^2\, \pi_i(\mathrm{d}x)\pi_i(\mathrm{d}y) \\
&\quad + \sum_{i\neq j,\, i,j\in\mathsf{I}_0} \pi(i)\pi(j)\pi_i(\mathsf{B}_i)\pi_j(\mathsf{B}_j)\int_{\mathsf{B}_i}\int_{\mathsf{B}_j} (f(y) - f(x))^2\, \frac{\pi_i(\mathrm{d}x)}{\pi_i(\mathsf{B}_i)}\frac{\pi_j(\mathrm{d}y)}{\pi_j(\mathsf{B}_j)}.
\end{aligned}$$

Given an edge e in $\mathcal{G}$, let us write e_- and e_+ to denote the two incident nodes of the edge. For $i \neq j \in \mathsf{I}_0$, let γ_{ij} denotes the chosen canonical path between i and j, and let $i_0, i_1, \ldots, i_\ell$ 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

$$\begin{aligned}
(5.10)\quad &\int_{\mathsf{B}_i}\int_{\mathsf{B}_j} (f(y) - f(x))^2\, \frac{\pi_i(\mathrm{d}x)}{\pi_i(\mathsf{B}_i)}\frac{\pi_j(\mathrm{d}y)}{\pi_j(\mathsf{B}_j)} \\
&\quad \leq |\gamma_{ij}|\sum_{e\in\gamma_{ij}}\int_{\mathsf{B}_{e_-}}\int_{\mathsf{B}_{e_+}} (f(y) - f(x))^2\, \frac{\pi_{e_-}(\mathrm{d}x)}{\pi_{e_-}(\mathsf{B}_{e_-})}\frac{\pi_{e_+}(\mathrm{d}y)}{\pi_{e_+}(\mathsf{B}_{e_+})},
\end{aligned}$$

where $|\gamma_{ij}|$ denotes the number of edges on the canonical path γ_{ij}. By Lemma 5.1 and using also the assumption that $\pi_i(\mathsf{B}_i) \geq 1/2$, the summation on the right-hand side of (5.10) is upper bounded by

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

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

$$\sum_{i\neq j,\, i,j\in\mathsf{I}_0}\pi(i)\pi(j)\pi_i(\mathsf{B}_i)\pi_j(\mathsf{B}_j)\int_{\mathsf{B}_i}\int_{\mathsf{B}_j}(f(y)-f(x))^2\,\frac{\pi_i(\mathrm{d}x)}{\pi_i(\mathsf{B}_i)}\frac{\pi_j(\mathrm{d}y)}{\pi_j(\mathsf{B}_j)}$$
$$\leq \frac{8}{\kappa}\sum_{\iota\in\mathsf{I}_0}\pi(\iota)\int_{\mathsf{B}_\iota}\int_{\mathsf{B}_\iota}(f(y)-f(x))^2\,\pi_\iota(\mathrm{d}x)\pi_\iota(\mathrm{d}y)\sum_{i,j\in\mathsf{I}_0:\,\gamma_{ij}\ni\iota}|\gamma_{ij}|\frac{\pi(i)\pi(j)}{\pi(\iota)}, \tag{5.11}$$

which together with (5.9) yields

$$2\left(\mathsf{Var}_\pi(f)-\frac{\zeta}{2}\right) \leq \left(1+\frac{8\mathsf{m}_1}{\kappa}\right)\sum_{i\in\mathsf{I}_0}\pi(i)\int_{\mathsf{B}_i}\int_{\mathsf{B}_i}(f(y)-f(x))^2\,\pi_i(\mathrm{d}x)\pi_i(\mathrm{d}y). \tag{5.12}$$

From the definition of $\mathsf{SpecGap}_i(K_i)$, we have

$$\int_{\mathsf{B}_i}\int_{\mathsf{B}_i}(f(y)-f(x))^2\,\pi_i(\mathrm{d}x)\pi_i(\mathrm{d}y) \leq \frac{2\mathcal{E}_i(f,f)}{\mathsf{SpecGap}_i(K_i)}, \tag{5.13}$$

which we use in (5.12), to arrive at

$$\left(\mathsf{Var}_\pi(f)-\frac{\zeta}{2}\right) \leq \frac{\left(1+\frac{8\mathsf{m}_1}{\kappa}\right)}{\min_{i\in\mathsf{I}_0}\mathsf{SpecGap}_i(K_i)}\sum_{i\in\mathsf{I}_0}\pi(i)\mathcal{E}_i(f,f). \tag{5.14}$$

(5.14) and (5.7) together yield,

$$\frac{\mathcal{E}(f,f)}{\left(\mathsf{Var}_\pi(f)-\frac{\zeta}{2}\right)} \geq \frac{\min_{i\in\mathsf{I}_0}\mathsf{SpecGap}_i(K_i)}{1+\frac{8\mathsf{m}_1}{\kappa}} \geq \frac{\kappa}{1+8\mathsf{m}_1}\min_{i\in\mathsf{I}_0}\mathsf{SpecGap}_i(K_i),$$

which together with the definition (2.3) implies the stated bound. □

5.5. Proof of Theorem 4.2. We start with some basic calculations on the model.

Lemma 5.2. *For $\delta, \vartheta \in \Delta$ such that $\vartheta \supseteq \delta$, setting $\tau \stackrel{\text{def}}{=} \frac{1}{\sigma^2}\left(\frac{1}{\rho_1} - \frac{1}{\rho_0}\right)$, we have*

$$\frac{\Pi(\vartheta|z)}{\Pi(\delta|z)} = \left(\frac{1}{p^u}\right)^{\|\vartheta\|_0 - \|\delta\|_0} \frac{e^{\frac{\tau}{2\sigma^2} z' L_\delta^{-1} X_{(\vartheta-\delta)} \left(I_{\|\vartheta-\delta\|_0} + \tau X'_{(\vartheta-\delta)} L_\delta^{-1} X_{(\vartheta-\delta)}\right)^{-1} X'_{(\vartheta-\delta)} L_\delta^{-1} z}}{\sqrt{\det\left(I_{\|\vartheta-\delta\|_0} + \tau X'_{(\vartheta-\delta)} L_\delta^{-1} X_{(\vartheta-\delta)}\right)}}. \tag{5.15}$$

Proof. We start with some basic calculations on the model. For any $\vartheta, \delta \in \Delta$, we have

$$\begin{aligned}
\frac{\Pi(\vartheta|z)}{\Pi(\delta|z)} &= \frac{\omega_\vartheta}{\omega_\delta}\left(\frac{\rho_1}{\rho_0}\right)^{\frac{\|\vartheta\|_0 - \|\delta\|_0}{2}} \frac{\int_{\mathbb{R}^p} e^{-\frac{1}{2\sigma^2}\|z - Xu\|_2^2 - \frac{1}{2} u' D_{(\vartheta)}^{-1} u} \mathrm{d}u}{\int_{\mathbb{R}^p} e^{-\frac{1}{2\sigma^2}\|z - Xu\|_2^2 - \frac{1}{2} u' D_{(\delta)}^{-1} u} \mathrm{d}u}. \\
&= \frac{\omega_\vartheta}{\omega_\delta}\left(\frac{\rho_1}{\rho_0}\right)^{\frac{\|\vartheta\|_0 - \|\delta\|_0}{2}} \frac{\sqrt{\det\left(\sigma^2 D_{(\delta)}^{-1} + X'X\right)}}{\sqrt{\det\left(\sigma^2 D_{(\vartheta)}^{-1} + X'X\right)}} \frac{e^{\frac{1}{2\sigma^2} z' X \left(\sigma^2 D_{(\vartheta)}^{-1} + X'X\right)^{-1} X' z}}{e^{\frac{1}{2\sigma^2} z' X \left(\sigma^2 D_{(\delta)}^{-1} + X'X\right)^{-1} X' z}}.
\end{aligned}$$

By the determinant lemma ($\det(A + UV') = \det(A)\det(I_m + V'A^{-1}U)$ valid for any invertible matrix $A \in \mathbb{R}^{n\times n}$, and $U, V \in \mathbb{R}^{n\times m}$) we have

$$\left(\frac{\rho_1}{\rho_0}\right)^{\frac{\|\vartheta\|_0 - \|\delta\|_0}{2}} \frac{\sqrt{\det\left(\sigma^2 D_{(\delta)}^{-1} + X'X\right)}}{\sqrt{\det\left(\sigma^2 D_{(\vartheta)}^{-1} + X'X\right)}} = \sqrt{\frac{\det\left(I_n + \frac{1}{\sigma^2} X D_{(\delta)} X'\right)}{\det\left(I_n + \frac{1}{\sigma^2} X D_{(\vartheta)} X'\right)}}.$$

By the Woodbury identity which states that for any set of matrices U, V, A, C with matching dimensions, $(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$, we have

$$\begin{aligned}
X\left(\sigma^2 D_{(\delta)}^{-1} + X'X\right)^{-1} X' &= \frac{1}{\sigma^2} X D_{(\delta)} X' - \frac{1}{\sigma^4} X D_{(\delta)} X' \left(I_n + \frac{1}{\sigma^2} X D_{(\delta)} X'\right)^{-1} X D_{(\delta)} X' \\
&= I_n - \left(I_n + \frac{1}{\sigma^2} X D_{(\delta)} X'\right)^{-1}.
\end{aligned}$$

so that,

$$\frac{e^{\frac{1}{2\sigma^2} z' X \left(\sigma^2 D_{(\vartheta)}^{-1} + X'X\right)^{-1} X' z}}{e^{\frac{1}{2\sigma^2} z' X \left(\sigma^2 D_{(\delta)}^{-1} + X'X\right)^{-1} X' z}} = \frac{e^{\frac{1}{2\sigma^2} z' \left(I_n + \frac{1}{\sigma^2} X D_{(\delta)} X'\right)^{-1} z}}{e^{\frac{1}{2\sigma^2} z' \left(I_n + \frac{1}{\sigma^2} X D_{(\vartheta)} X'\right)^{-1} z}}.$$

We combine these developments together to conclude that

$$\frac{\Pi(\vartheta|z)}{\Pi(\delta|z)} = \frac{\omega_\vartheta}{\omega_\delta} \sqrt{\frac{\det\left(L_\delta\right)}{\det\left(L_\vartheta\right)}} \frac{e^{\frac{1}{2\sigma^2} z' L_\delta^{-1} z}}{e^{\frac{1}{2\sigma^2} z' L_\vartheta^{-1} z}}, \tag{5.16}$$

where, for $\delta \in \Delta$, we recall the definition $L_\delta \stackrel{\text{def}}{=} I_n + \frac{1}{\sigma^2} X D_{(\delta)} X'$. If $\vartheta \supseteq \delta$, setting $\tau \stackrel{\text{def}}{=} \frac{1}{\sigma^2}\left(\frac{1}{\rho_1} - \frac{1}{\rho_0}\right) < 1/(\sigma^2\rho_1)$, it is easily seen that

$$L_\vartheta = L_\delta + \tau \sum_{j:\ \delta_j=0, \vartheta_j=1} X_j X_j'.$$

The determinant lemma then gives

$$\frac{\det(L_\vartheta)}{\det(L_\delta)} = \det\left(I_{\|\vartheta-\delta\|_0} + \tau X'_{(\vartheta-\delta)} L_\delta^{-1} X_{(\vartheta-\delta)}\right).$$

And the Woodbury identity gives

$$L_\vartheta^{-1} = L_\delta^{-1} - \tau L_\delta^{-1} X_{(\vartheta-\delta)} \left(I_{\|\vartheta-\delta\|_0} + \tau X'_{(\vartheta-\delta)} L_\delta^{-1} X_{(\vartheta-\delta)}\right)^{-1} X'_{(\vartheta-\delta)} L_\delta^{-1}.$$

Combining the last two display in (5.16) yields the stated results. ■

Lemma 5.3. *Assume H2. Let ϱ and $\underline{\theta}_\star$ be as in Theorem 4.2. For $z \in \mathcal{E}_0$, we have*

$$\max_{\delta\in\Delta_s} \max_{j:\ \delta_{\star j}=0} |X_j' L_\delta^{-1} z| \leq \sqrt{\varrho n \log(p)}, \tag{5.17}$$

$$\text{and} \quad \max_{\delta\in\Delta_s} \max_{j:\ \delta_{\star j}=1} |X_j' L_\delta^{-1} z| \leq \|\theta_\star\|_\infty n + \sqrt{\varrho n \log(p)}.$$

Furthermore, if $n \geq 4\varrho \log(p)/(\underline{\theta}_\star^2 \lambda_1^2)$, then

$$\min_{\delta\in\Delta_s} \min_{j:\ \delta_{\star j}=1} |X_j' L_\delta^{-1} z| \geq \frac{\lambda_1}{2} \underline{\theta}_\star n.$$

Proof. Set $V \stackrel{\text{def}}{=} (z - X\theta_\star)/\sigma$, so that

$$z = \sigma V + \sum_{k:\ \delta_{\star k}=1} \theta_{\star k} X_k,$$

and

$$X_j' L_\delta^{-1} z = \sigma X_j' L_\delta^{-1} V + \sum_{k:\ \delta_{\star k}=1} \theta_{\star,k} X_j' L_\delta^{-1} X_k.$$

For $z \in \mathcal{E}_0$, $|X_j' L_\delta^{-1} V| \leq \sqrt{c_0 n \log(p)}$. If $\delta_{\star j} = 0$ and $\delta_{\star k} = 1$, then $|X_j' L_\delta^{-1} X_k| \leq \mathcal{C}(s)\sqrt{n\log(p)}$. Hence

$$\max_{\delta\in\Delta_s} \max_{j:\ \delta_{\star j}=0} |X_j' L_\delta^{-1} z| \leq \left(\sigma\sqrt{c_0} + \mathcal{C}(s) \sum_{k:\ \delta_{\star k}=1} |\theta_{\star k}|\right)\sqrt{n\log(p)} \leq \sqrt{\varrho n \log(p)}.$$

If $\delta_{\star j} = 1$, then

$$X_j' L_\delta^{-1} z = \sigma X_j' L_\delta^{-1} V + \theta_{\star,j} X_j' L_\delta^{-1} X_j + \sum_{k\neq j:\ \delta_{\star k}=1} \theta_{\star,k} X_j' L_\delta^{-1} X_k.$$

Since $X_j' L_\delta^{-1} X_j \leq \|X_j\|_2^2 = n$, this implies, as we have done above that $|X_j' L_\delta^{-1} z| \leq \|\theta_\star\|_\infty n + \sqrt{\varrho n \log(p)}$. Similarly, if $\delta_{\star j} = 1$, then $|X_j' L_\delta^{-1} z| \geq \frac{1}{2}|\theta_{\star,j}| X_j' L_\delta^{-1} X_j$, provided that we have $\sqrt{\varrho n \log(p)} \leq \frac{1}{2}|\theta_{\star,j}| X_j' L_\delta^{-1} X_j$. Then using the definition of λ_1, we get

$$\min_{\delta \in \Delta_s} \min_{j:\, \delta_{\star j}=1} |X_j' L_\delta^{-1} z| \geq \frac{\lambda_1}{2} \underline{\theta}_\star n. \qquad \blacksquare$$

Proof of Theorem 4.2. Fix $\zeta_0 \in (0,1)$. We will apply Lemma 2.1 with $\|\cdot\|_\star = \|\cdot\|_{\pi,\infty}$. Since Kf is bounded when f is bounded, the kernel K satisfies (2.2) with this choice of $\|\cdot\|_\star$. We recall that the initial distribution is taken as $\nu_0 = \Pi(\cdot|\delta^{(\mathrm{i})}, z)$, for some initial choice $\delta^{(\mathrm{i})} \in \Delta_s$. Let f_0 be the density of ν_0 with respect to $\Pi(\cdot|z)$. We Lemma 2.1 with $\zeta = 0$ to conclude that

$$\|\nu_0 K^N - \Pi(\cdot|z)\|_{\mathrm{tv}}^2 \leq \zeta_0^2, \quad \text{for} \quad N \geq \frac{1}{\mathsf{SpecGap}_0(K)} \log\left(\frac{\mathsf{Var}_\pi(f_0)}{\zeta_0^2}\right). \tag{5.18}$$

To bound the spectral gap we apply Theorem 3.1 with the choices $\zeta = 0$, $\mathsf{I} = \Delta$, $\mathsf{I}_0 = \Delta_s$, and $\mathsf{B}_\delta = \mathbb{R}^p$, and with a graph on Δ_s constructed as follows: we put an edge between $\delta^{(1)}$ and $\delta^{(2)}$ if $\delta^{(1)} \supseteq \delta^{(2)}$, or $\delta^{(2)} \supseteq \delta^{(1)}$, and $\|\delta^{(2)} - \delta^{(1)}\|_0 = 1$ (in other words the models $\delta^{(1)}$ and $\delta^{(2)}$ differ only in one variable). Clearly (3.8) holds, since $\Pi(\Delta_s|z) = 1$. We then conclude from Theorem 3.1 that

$$\mathsf{SpecGap}_0(K) \geq \frac{\kappa}{1 + 8\mathsf{m}_1}. \tag{5.19}$$

To bound the constants κ and m_1 we develop a similar argument as in [Yang et al.(2016)]. Given $\delta \in \Delta_s$, we call $\min(\delta, \delta_\star)$ the skeleton of δ, and we let $\mathcal{S} \stackrel{\text{def}}{=} \{\min(\delta, \delta_\star),\ \delta \in \Delta_s\}$ be the set of all possible skeletons. Basically $\mathcal{S}$ is the set of submodels of the true model $\delta_\star$. Given $\delta \in \Delta_s$, we build our canonical path from δ to $\delta_\star$ as follows. First we build a path from δ to its skeleton (that is $\min(\delta, \delta_\star)$) by successively removing from the model δ the variables X_j for which $\delta_j = 1$ and $\delta_{\star j} = 0$, in reverse index ordering. Then we build a path from the skeleton to $\delta_\star$ by adding to the skeleton the variables X_j for which $\delta_j = 0$ and $\delta_{\star j} = 1$ in their index ordering. For example, if $p = 6$, $\delta_\star = (1,1,1,0,0)$ and $\delta = (0,0,1,0,1,1)$, then our canonical path from δ to $\delta_\star$ is

$$(0,0,1,0,1,1) \to (0,0,1,0,1,0) \to (0,0,1,0,0,0) \to (1,0,1,0,0,0) \to (1,1,1,0,0,0).$$

Given $\delta^{(1)}, \delta^{(2)} \in \Delta_s$, let $\delta^{(1,2)}$ be the node where the canonical path from $\delta^{(1)}$ to $\delta_\star$ and the canonical path from $\delta^{(2)}$ to $\delta_\star$ meet for the first time. Our canonical path $\gamma_{\delta^{(1)},\delta^{(2)}}$ between $\delta^{(1)}$ and $\delta^{(2)}$ is then defined as follows. Follow the canonical path from $\delta^{(1)}$ towards $\delta_\star$ until $\delta^{(1,2)}$, then reverse direction and follow the path from $\delta^{(1,2)}$ until $\delta^{(2)}$. For instance if $p = 6$, $\delta_\star = (1,1,1,0,0,0)$ and $\delta^{(1)} = (0,1,0,0,1,1)$, and $\delta^{(2)} = (1,1,0,1,1,0)$, then $\delta^{(1,2)} = (1,1,0,0,0)$, and our chosen canonical path from $\delta^{(1)}$ to $\delta^{(2)}$ is

$$(0,1,0,0,1,1) \to (0,1,0,0,1,0) \to (0,1,0,0,0,0) \to (1,1,0,0,0,0) \to (1,1,0,1,0,0) \to (1,1,0,1,1,0).$$

We claim that for the canonical paths constructed above we have

$$\mathsf{m}_1 \stackrel{\text{def}}{=} \max_{\delta\in\Delta_s} \sum_{\delta^{(1)},\delta^{(2)}\in\Delta_s:\ \gamma_{\delta^{(1)},\delta^{(2)}}\ni\delta} |\gamma_{\delta^{(1)},\delta^{(2)}}| \frac{\pi(\delta^{(1)}|z)\pi(\delta^{(2)}|z)}{\pi(\delta|z)} \leq 8s, \tag{5.20}$$

and

$$\begin{aligned}\kappa \stackrel{\text{def}}{=} \min_{\delta^{(1)}\sim\delta^{(2)}} \int_{\mathbb{R}^p} \min\left(\Pi(\theta|\delta^{(1)},z), \Pi(\theta|\delta^{(2)},z)\right) \mathrm{d}\theta \\ \geq \frac{1}{2}\min\left(1, \sqrt{\frac{\sigma^2\rho_0}{2n}}\right)\min\left(1, e^{\frac{1}{4\sigma^2}(n\lambda_1-8\sigma^2\rho_0)}\right) p^{-\frac{2\rho_0}{n}\frac{\varrho}{\lambda^2}}.\end{aligned} \tag{5.21}$$

where the minimum is taken over all connected pairs of nodes $\delta^{(1)},\delta^{(2)}$. Furthermore, we claim that we can bound the variance of the initial density and get

$$\log\left(\frac{\mathsf{Var}_\pi(f_0)}{\zeta_0^2}\right) \leq A\left(\log\left(\frac{1}{\zeta_0}\right) + \frac{su(1+\|\theta_\star\|_\infty^2)n}{\sigma^2\lambda}\right), \tag{5.22}$$

for some absolute constant A. (5.20) and (5.21) shows that

$$\mathsf{SpecGap}_0(K) \geq \frac{A}{s}\min\left(1, \sqrt{\frac{\sigma^2\rho_0}{2n}}\right)\min\left(1, e^{\frac{1}{4\sigma^2}(n\lambda_1-8\sigma^2\rho_0)}\right) p^{-\frac{2\rho_0}{n}\frac{\varrho}{\lambda^2}} \tag{5.23}$$

for some absolute constant A. We put (5.23) together with (5.22) and (5.18) to reach the stated conclusion. The remaining of the proof consists in establishing the claims (5.20), (5.21) and (5.22).

Proof of Equation (5.20). For $\delta^{(1)},\delta^{(2)}\in\Delta_s$, we will use the obvious bound

$$|\gamma_{\delta^{(1)},\delta^{(2)}}| \leq 2s.$$

Given $\delta\in\Delta_s$, we denote $\Lambda(\delta)$ the set of all $\delta^{(1)}\in\Delta_s$ such that the canonical path from $\delta^{(1)}$ to $\delta_\star$ goes through δ. Using this we can bound m_1 as

$$\mathsf{m}_1 \leq 2s \max_{\delta\in\Delta_s} \sum_{\delta^{(1)}\in\Lambda(\delta)} \sum_{\delta^{(2)}\in\Delta_s} \frac{\pi(\delta^{(1)}|z)\pi(\delta^{(2)}|z)}{\pi(\delta|z)} \leq 2s \max_{\delta\in\Delta_s} \sum_{\delta^{(1)}\in\Lambda(\delta)} \frac{\pi(\delta^{(1)}|z)}{\pi(\delta|z)}. \tag{5.24}$$

Let $\mathcal{S} \stackrel{\text{def}}{=} \{\min(\delta,\delta_\star),\ \delta\in\Delta_s\}$ be the set of all possible skeletons. Take $\delta^{(1)}\in\Lambda(\delta)$. We will distinguish whether $\delta\in\mathcal{S}$ or not. Suppose $\delta\notin\mathcal{S}$. Therefore, traveling the canonical path from $\delta^{(1)}$ toward $\delta_\star$ we arrive at δ by removing only non-significant variables. Therefore, assuming that $\|\delta^{(1)}\|_0 = \|\delta\|_0 + \ell$, and using (5.15), and H2, we have

$$\frac{\pi(\delta^{(1)}|z)}{\pi(\delta|z)} \leq \frac{1}{p^{u\ell}}\exp\left(\frac{\tau}{2\sigma^2(1+n\tau\lambda)}\sum_{j:\ \delta_j^{(1)}=1,\delta_j=0}(X_j'L_\delta^{-1}z)^2\right) \leq \frac{e^{\frac{\ell\bar{Q}_0}{n\lambda}}}{p^{u\ell}}, \tag{5.25}$$

where $\bar{Q}_0 = \max_{j:\, \delta^{(1)}_j=1, \delta_j=0} (X_j' L_\delta^{-1} z)^2$. From Lemma 5.3, we get $\bar{Q}_0 \leq \varrho n \log(p)$. Using this and the trivial inequality $\binom{p}{\ell} \leq p^\ell$, it follows that

$$\sum_{\delta^{(1)} \in \Lambda(\delta)} \frac{\pi(\delta^{(1)}|z)}{\pi(\delta|z)} \leq \sum_{\ell=0}^{s-\|\delta\|_0} \sum_{\delta^{(1)} \in \Lambda(\delta):\, \|\delta^{(1)}\|_0 = \|\delta\|_0 + \ell} \frac{\pi(\delta^{(1)}|z)}{\pi(\delta|z)} \leq \sum_{\ell=0}^{s} \left(\frac{p^{\frac{\varrho}{2\sigma^2\lambda}}}{p^{u-1}} \right)^\ell \leq 2,$$

under the assumption that $\sigma^2 u \lambda \geq \varrho$, and $u > 4$. Suppose now that $\delta \in \mathcal{S}$. Then $\Lambda(\delta)$ is comprised of the elements of Δ_s whose skeletons are subsets of δ. Hence

$$\sum_{\delta^{(1)} \in \Lambda(\delta)} \frac{\pi(\delta^{(1)}|z)}{\pi(\delta|z)} = \sum_{\delta_0 \in \mathcal{S}: \delta \supseteq \delta_0} \frac{\pi(\delta_0|z)}{\pi(\delta|z)} \sum_{\delta^{(1)} \in \Lambda(\delta):\, \min(\delta^{(1)}, \delta_\star) = \delta_0} \frac{\pi(\delta^{(1)}|z)}{\pi(\delta_0|z)}.$$

The inner summation can be upper bounded by 2 as above. If $\delta \supseteq \delta_0$ and $\|\delta\|_0 = \|\delta_0\|_0 + r$, we apply (5.15) again and get,

$$\frac{\pi(\delta_0|z)}{\pi(\delta|z)} \leq \left(p^u \sqrt{1 + \frac{n s_\star}{\sigma^2 \rho_1}}\, e^{-\frac{\tau \bar{Q}_3}{2\sigma^2(1+\tau s_\star n)}} \right)^r \leq \left(p^{u+a} e^{-\frac{\bar{Q}_3}{4\sigma^2 s_\star n}} \right)^r,$$

where we use H3-(3) to obtain $\tau/(1+\tau s_\star n) \geq 1/(2 s_\star n)$, and $\sqrt{1 + \frac{n s_\star}{\sigma^2 \rho_1}} \leq p^a$, and where $\bar{Q}_3 \stackrel{\text{def}}{=} \min_{j:\, \delta_{0j}=0, \delta_{\star j}=1} \left(X_j' L_{\delta_0}^{-1} z \right)^2$. From Lemma 5.3 we get $\bar{Q}_3 \geq \frac{\underline{\theta}_\star^2}{4} \lambda_1^2 n^2$, under the sample condition $n \geq 4\varrho \log(p)/(\underline{\theta}_\star^2 \lambda_1^2)$ which is implied by (4.10). We conclude that

$$\max_{\delta \in \mathcal{S}} \sum_{\delta^{(1)} \in \Lambda(\delta)} \frac{\pi(\delta^{(1)}|z)}{\pi(\delta|z)} \leq 2 \sum_{\delta_0 \in \mathcal{S}: \delta \supseteq \delta_0} \frac{\pi(\delta_0|z)}{\pi(\delta|z)} \leq 2 \sum_{r=0}^{s_\star} s_\star^r \left(p^{u+a} e^{-\frac{\underline{\theta}_\star^2 \lambda_1^2 n}{4\sigma^2 s_\star}} \right)^r \leq 4,$$

using the sample size condition in (4.10). This proves the claim (5.20).

Proof of Equation (5.21). Fix $\delta^{(1)}, \delta^{(2)} \in \Delta_s$, such that $\delta^{(1)} \supseteq \delta^{(2)}$, or $\delta^{(2)} \subseteq \delta^{(1)}$, and $\|\delta^{(2)} - \delta^{(1)}\|_0 = 1$. Without any loss of generality, suppose that $\delta^{(2)} \supseteq \delta^{(1)}$, and their difference occurs on component j: $\delta^{(2)}_j = 1$, while $\delta^{(1)}_j = 0$. Then for all $\theta \in \mathbb{R}^p$, we have

$$\frac{\Pi(\theta|\delta^{(1)}, z)}{\Pi(\theta|\delta^{(2)}, z)} = \left(\frac{\int_{\mathbb{R}^p} e^{-\frac{1}{2\sigma^2}\|z - X\theta\|_2^2 - \frac{1}{2}\theta' D^{-1}_{(\delta^{(2)})}\theta} \mathrm{d}\theta}{\int_{\mathbb{R}^p} e^{-\frac{1}{2\sigma^2}\|z - X\theta\|_2^2 - \frac{1}{2}\theta' D^{-1}_{(\delta^{(1)})}\theta} \mathrm{d}\theta} \right) e^{-(\rho_0 - \rho_1)\frac{\theta_j^2}{2}}.$$

Let A denote the ratio of integrals in the last display. We can then write

$$\int_{\mathbb{R}^p} \min\left(\Pi(\theta|\delta^{(1)}, z), \Pi(\theta|\delta^{(2)}, z) \right) \mathrm{d}\theta = \int_{\mathbb{R}} \min\left(1, A e^{-(\rho_0-\rho_1)\frac{\theta_j^2}{2}} \right) \Pi(\theta_j|\delta^{(2)}, z) \mathrm{d}\theta_j.$$

Recall from (4.3) that the j marginal under $\Pi(\theta_j|\delta^{(2)}, z)$ is the Gaussian distribution $\mathbf{N}(\mu_j, \sigma_j^2)$, where

$$\sigma_j = \sigma \sqrt{e_j' \Sigma_{\delta^{(2)}} e_j}, \quad \text{and} \quad \mu_j = e_j' \Sigma_{\delta^{(2)}} X' z, \quad 1 \leq j \leq p,$$

and where e_j denotes the j-th unit vector. Hence, for $Z \sim \mathbf{N}(0,1)$,

$$(5.26)\quad \int_{\mathbb{R}^p} \min\left(\Pi(\theta|\delta^{(1)}, z), \Pi(\theta|\delta^{(2)}, z)\right) \mathrm{d}\theta = \mathbb{E}\left[\min\left(1, Ae^{-\frac{(\rho_0-\rho_1)}{2}(\mu_j+\sigma_j Z)^2}\right)\right]$$
$$\geq \frac{1}{2}\min\left(1, Ae^{-\frac{(\rho_0-\rho_1)}{2}(|\mu_j|+\sigma_j)^2}\right) \geq \frac{1}{2}\min\left(1, Ae^{-\rho_0(\mu_j^2+\sigma_j^2)}\right),$$

using the fact that for any nonnegative function f, $\mathbb{E}(f(Z)) \geq \mathbb{P}(|Z| \leq 1) \min_{z:|z|\leq 1} f(z)$. By matrix block inversion, we work out σ_j^2 to

$$(5.27)\quad \sigma_j^2 = \frac{\sigma^2}{\sigma^2\rho_1 + X_j'\left(I_n + \frac{1}{\sigma^2}X_{-j}D_{(\delta^{(2)},j)}X_{-j}'\right)^{-1}X_j} = \frac{\sigma^2}{\sigma^2\rho_1 + X_j' L^{-1}_{\delta^{(1)}_{-j}} X_j} \leq \frac{\sigma^2}{\lambda_1 n},$$

where $D_{(\delta^{(2)},j)} = D_{(\delta^{(1)},j)}$ is the $(p-1)$-dimensional matrix obtained by removing the j-th row and the j-th column of $D_{(\delta^{(2)})}$, and $L_{\delta^{(1)}_{-j}} = I_n + \frac{1}{\sigma^2}X_{-j}D_{(\delta^{(1)},j)}X_{-j}'$. By block inversion the mean μ_j can be written as

$$(5.28)\quad \mu_j = e_1 \begin{pmatrix} X_j'X_j + \sigma^2\rho_1 & X_j'X_{-j} \\ X_{-j}'X_j & X_{-j}'X_{-j} + \sigma^2 D^{-1}_{(\delta^{(2)},j)} \end{pmatrix}^{-1} \begin{pmatrix} X_j' z \\ X_{-j}' z \end{pmatrix} = \frac{X_j' L^{-1}_{\delta^{(1)}_{-j}} z}{\sigma^2\rho_1 + X_j' L^{-1}_{\delta^{(1)}_{-j}} X_j}.$$

Consider first the case where j is such that $\delta_{\star,j} = 0$. Note that $X_j' L^{-1}_{\delta^{(1)}_{-j}} X_j \geq X_j' L^{-1}_{\delta^{(1)}} X_j \geq n\lambda_1$. Therefore, and using Lemma 5.3, and $z \in \mathcal{E}_0$, we obtain

$$|\mu_j| \leq \frac{1}{n\lambda_1}\sqrt{\varrho n \log(p)} = \frac{1}{\lambda_1}\sqrt{\frac{\varrho \log(p)}{n}}.$$

Consider now the case where $\delta_{\star,j} = 1$. Then we have

$$\mu_j^2 \leq \frac{1}{(X_j' L^{-1}_{\delta^{(1)}_{-j}} X_j)^2}\left(\theta_{\star j} X_j' L^{-1}_{\delta^{(1)}_{-j}} X_j + \sigma\sqrt{c_0 n \log(p)} + \sum_{k:\ \delta_{\star k}=1} \theta_{\star k} X_j' L^{-1}_{\delta^{(1)}_{-j}} X_k\right)^2$$
$$\leq \frac{2}{(X_j' L^{-1}_{\delta^{(1)}_{-j}} X_j)^2}\left(\theta_{\star j}^2 (X_j' L^{-1}_{\delta^{(1)}_{-j}} X_j)^2 + \varrho n \log(p)\right)$$
$$\leq 2\theta_{\star j}^2 + \frac{\varrho \log(p)}{\lambda_1^2 n}.$$

On the other hand, using (5.15), the ratio of integrals A gives

$$A = \sqrt{\frac{\rho_0}{\rho_1}} \frac{1}{\sqrt{1 + \tau X_j' L^{-1}_{\delta^{(1)}} X_j}} \exp\left(\frac{1}{2\sigma^2}\frac{\tau(X_j' L^{-1}_{\delta^{(1)}} z)^2}{1 + \tau X_j' L^{-1}_{\delta^{(1)}} X_j}\right),$$

where we recall that $\tau = (\rho_1^{-1} - \rho_0^{-1})/\sigma^2$. Note that if $\delta_{\star j} = 0$, the term inside the exponential in this last expression of A grows like $\varrho \log(p)/\lambda_1$ which is not fast enough to face off with the term $-\rho_0(\mu_j^2 + \sigma_j^2)$. Hence we use instead the trivial lower bound $A \geq 1$ together with the upper bounds on μ_j and σ_j^2 obtained above and (5.26) to conclude that

$$\text{(5.29)} \qquad \int_{\mathbb{R}^p} \min\left(\Pi(\theta|\delta^{(1)}, z), \Pi(\theta|\delta^{(2)}, z)\right) \mathrm{d}\theta \geq \frac{1}{2} e^{-\rho_0(\mu_j^2+\sigma_j^2)} \geq \frac{1}{2} \exp\left(-\frac{2\rho_0}{n} \frac{\varrho \log(p)}{\lambda_1^2}\right).$$

However if $\delta_{\star j} = 1$, By Lemma 5.3, and under the sample size condition (4.10) we have

$$(X_j' L_{\delta^{(1)}}^{-1} z)^2 \geq \frac{\theta_{\star j}^2}{2} (X_j' L_{\delta^{(1)}}^{-1} X_j)^2.$$

Noting that $1 \leq \tau X_j' L_{\delta^{(1)}}^{-1} X_j$, we deduce that

$$A \geq \sqrt{\frac{\rho_0}{\rho_1}} \frac{1}{\sqrt{1 + \frac{n}{\sigma^2 \rho_1}}} e^{\frac{\theta_{\star j}^2 X_j' L_{\delta^{(1)}}^{-1} X_j}{4\sigma^2}}.$$

It follows in this case that

$$\text{(5.30)} \qquad \int_{\mathbb{R}^p} \min\left(\Pi(\theta|\delta^{(1)}, z), \Pi(\theta|\delta^{(2)}, z)\right) \mathrm{d}\theta \geq \frac{1}{2} \min\left(1, A e^{-\rho_0\left(2\theta_{\star j}^2 + \frac{\varrho \log(p)}{\lambda^2 n} + \frac{\sigma^2}{\lambda n}\right)}\right)$$
$$\geq \frac{1}{2} \min\left(1, \sqrt{\frac{\sigma^2 \rho_0}{\sigma^2 \rho_1 + n}}\right) \min\left(1, e^{\frac{\theta_{\star j}^2 \left(X_j' L_{\delta^{(1)}}^{-1} X_j - 8\sigma^2 \rho_0\right)}{4\sigma^2}} e^{-2\rho_0 \frac{\varrho \log(p)}{\lambda^2 n}}\right)$$
$$\geq \frac{1}{2} \min\left(1, \sqrt{\frac{\sigma^2 \rho_0}{2n}}\right) \min\left(1, e^{\frac{1}{4\sigma^2}(n\lambda_1 - 8\sigma^2 \rho_0)}\right) p^{-\frac{2\rho_0}{n} \frac{\varrho}{\lambda^2}},$$

where we have used the fact that $\min(1, ab) \geq \min(1, a) \min(1, b)$ valid for all nonnegative numbers a, b, c. We combine (5.29) and (5.30) to obtain (5.21).

Proof of Equation (5.22). Since $\Pi(\theta|z) = \sum_\vartheta \Pi(\vartheta|z)\Pi(\theta|\vartheta, z) \geq \Pi(\delta^{(\mathsf{i})}|z)\Pi(\theta|\delta^{(\mathsf{i})}, z)$, we have

$$f_0(\theta) = \frac{\Pi(\theta|\delta^{(\mathsf{i})}, z)}{\Pi(\theta|z)} \leq \frac{1}{\Pi(\delta^{(\mathsf{i})}|z)} = \frac{1}{\Pi(\delta_\star|z)} \frac{\Pi(\delta_\star|z)}{\Pi(\delta_\star^{(\mathsf{i})}|z)} \frac{\Pi(\delta_\star^{(\mathsf{i})}|z)}{\Pi(\delta^{(\mathsf{i})}|z)},$$

where $\delta_\star^{(\mathsf{i})} \stackrel{\text{def}}{=} \min(\delta^{(\mathsf{i})}, \delta_\star)$. We apply (5.15) twice (to each ratio), and use H2, to get

$$\frac{\Pi(\delta_\star|z)}{\Pi(\delta_\star^{(\mathsf{i})}|z)} \frac{\Pi(\delta_\star^{(\mathsf{i})}|z)}{\Pi(\delta|z)} \leq p^{u(\|\delta\|_0 - \|\delta_\star\|_0)} \sqrt{\det\left(I_{\|\delta\|_0 - \|\delta_\star^{(\mathsf{i})}\|_0} + \tau X'_{(\delta - \delta_\star^{(\mathsf{i})})} L_{\delta_\star^{(\mathsf{i})}}^{-1} X_{(\delta - \delta_\star^{(\mathsf{i})})}\right)}$$
$$\times e^{\frac{\tau}{2\sigma^2} z' L_{\delta_\star^{(\mathsf{i})}}^{-1} X_{(\delta_\star - \delta_\star^{(\mathsf{i})})} \left(I_{\|\delta_\star - \delta_\star^{(\mathsf{i})}\|_0} + \tau X'_{(\delta_\star - \delta_\star^{(\mathsf{i})})} L_{\delta_\star^{(\mathsf{i})}}^{-1} X_{(\delta_\star - \delta_\star^{(\mathsf{i})})}\right)^{-1} X'_{(\delta_\star - \delta_\star^{(\mathsf{i})})} L_{\delta_\star^{(\mathsf{i})}}^{-1} z}$$
$$\leq p^{u(\|\delta\|_0 - s_\star)} \left(1 + \frac{n\|\delta\|_0}{\sigma^2 \rho_1}\right)^{\frac{\|\delta\|_0}{2}} e^{\frac{1}{2\sigma^2 n\lambda} \|X'_{(\delta_\star - \delta_\star^{(\mathsf{i})})} L_{\delta_\star^{(\mathsf{i})}}^{-1} z\|_2^2}.$$

Under the assumption $p^{us_\star}\Pi(\delta_\star|z) \geq 1$ (H3-(1)), and since $\|\delta\|_0 \leq s$, we conclude that

$$\|f_0\|_{\pi,\infty} \leq p^{us}\left(1+\frac{ns}{\sigma^2\rho_1}\right)^{\frac{s}{2}} e^{\frac{s_\star \bar{Q}_1}{2\sigma^2 n\lambda}} \leq p^{(u+a)s} e^{\frac{s_\star \bar{Q}_1}{2\sigma^2 n\lambda}}, \tag{5.31}$$

where the second inequality uses (4.8), and where $\bar{Q}_1 = \max_{j:\delta_{\star,j}=1}(X_j' L^{-1}_{\delta_\star^{(\mathsf{i})}} z)^2$. From Lemma 5.3, we get $\bar{Q}_1 \leq 4n^2\|\theta_\star\|_\infty^2$, using the sample size condition (4.10). (5.31) then becomes

$$\sqrt{\mathsf{Var}_\pi(f_0)} \leq \|f_0\|_{\pi,\infty} \leq p^{(u+a)s} e^{\frac{2s_\star\|\theta_\star\|_\infty^2 n}{\sigma^2\lambda}} \leq e^{\frac{As(1+\|\theta_\star\|_\infty^2)n}{\sigma^2\lambda}},$$

for some absolute constant A. The claim follows by taking the log. ■

5.6. Proof of Theorem 4.3. The proof is very similar to the proof of Theorem 4.2. Fix $\zeta_0 \in (0,1)$, and $z \in \mathcal{E}$. First we bound the uniform norm of the density of the initial distribution ν_0 as in (5.31). Noting here that the skeleton of $\delta^{(\mathsf{i})}$ is $\delta_\star$, we get the simpler bound

$$\|f_0\|_{\pi,\infty} \leq 2\left(p^u\sqrt{1+\frac{n\mathsf{FP}}{\sigma^2\rho_1}}\right)^{\mathsf{FP}} \leq 2p^{(u+a)\mathsf{FP}}.$$

In view of this bound, we set

$$\zeta = \frac{\zeta_0^2}{8}p^{-2(u+a)\mathsf{FP}}, \tag{5.32}$$

which gives $\zeta\|f_0\|_{\pi,\infty}^2 \leq \zeta_0^2/2$. Therefore, we can readily apply Lemma 2.1 with this particular value of ζ to get

$$\|\nu_0 K^N - \Pi(\cdot|z)\|_{\mathrm{tv}}^2 \leq \zeta_0^2, \quad \text{for} \quad N \geq \frac{1}{\mathsf{SpecGap}_\zeta(K)}\log\left(\frac{1}{\zeta}\right). \tag{5.33}$$

We lower bound the approximate spectral gap via Theorem 3.1, and using the same approach as in Theorem 4.2. We apply Theorem 3.1 with the choices $\mathsf{I} = \Delta$, $\mathsf{I}_0 = \mathcal{D}_k$ endowed with the same graph as in proof of Theorem 4.2, and $\mathsf{B}_\delta = \mathbb{R}^p$. First we need to check (3.8). For $z \in \mathcal{E}$, ζ as in (5.32), we have

$$\frac{10}{\zeta}(1-\Pi(\mathcal{D}_k|z)) \leq \frac{80}{\zeta_0^2}p^{2(u+a)\mathsf{FP}}\frac{1}{p^{\frac{u(k+1)}{2}}} \leq 1,$$

where the last inequality follows from condition (4.12). In other words we have $\Pi(\mathcal{D}_k|z) \geq 1-(\zeta/10)$, which by Remark 3.2 implies (3.8). We then conclude from Theorem 3.1 that

$$\mathsf{SpecGap}_\zeta(K) \geq \frac{\kappa}{1+8\mathsf{m}_1}, \tag{5.34}$$

where κ and m_1 are defined using $\mathcal{D}_k$. We bound these terms as in Theorem 4.2 with some important simplifications due the facts that all models here belong to $\mathcal{D}_k$. In particular, since $\mathcal{D}_k \subseteq \Delta_s$, we readily have

$$\mathsf{m}_1 \leq 8k. \tag{5.35}$$

Similarly, the lower bound on κ also simplifies. Because $\delta^{(1)}$ and $\delta^{(2)}$ can differ only at a component j such that $\delta_{\star j} = 0$ (a non-important variable), we see that only the lower bound (5.29) applies. Hence κ can be taken as

$$\kappa = \frac{1}{2} p^{-\frac{2\rho_0}{n}\frac{\varrho}{\lambda_1^2}}. \tag{5.36}$$

The theorem follows from the same calculations as in the proof of Theorem 4.2. □

Appendix A. Some technical results.

We make use of the following standard Gaussian deviation bound.

Lemma A.1. *Let $Z \sim \boldsymbol{N}(0, I_m)$, and $u_1, \ldots, u_N$ be vectors of $\mathbb{R}^m$. Then for all $x \geq 0$,*

$$\mathbb{P}\left[\max_{1\leq j\leq N} |\langle u_j, Z\rangle| > \max_{1\leq j\leq N} \|u_j\|_2 \sqrt{2(x + \log(N))}\right] \leq \frac{2}{e^x}.$$

Lemma A.2. *Suppose that $X \in \mathbb{R}^{n\times p}$ is a random matrix with i.i.d. standard Normal entries. Given an integer s, and positive constants σ, γ and ρ, set*

$$\mathcal{C}_0 \stackrel{\text{def}}{=} \max_{\delta\in\Delta:\ \|\delta\|_0\leq s}\ \max_{i\neq j,\ \delta_j=0} \left| X_j' \left(I_n + \frac{1}{\sigma^2\rho_1} X_\delta X_\delta' + \frac{1}{\sigma^2\rho_0} X_{\delta^c} X_{\delta^c}' \right) X_i \right|.$$

Then there exist some universal finite constants c_0, a, A such that for $n \geq As^2\log(p)$, the following two statements hold with probability at least $1 - \frac{a}{p}$: for $\rho_0^{-1} > 0$ taken small enough and

$$\sigma^2 s\rho_1 \leq c_0\sqrt{n\log(p)}, \tag{A.1}$$

it holds that

$$\mathcal{C}_0 \leq 2c_0\sqrt{n\log(p)}, \quad \text{and} \tag{A.2}$$

$$\min_{\delta:\ \|\delta\|_0\leq s}\ \inf\left\{\frac{u'(X_{\delta^c}' L_\delta^{-1} X_{\delta^c})u}{n\|u\|_2^2},\ u\in\mathbb{R}^{p-s},\ 0 < \|\mathsf{supp}(u)\|_0 \leq s\right\} \geq \frac{1}{32}.$$

Proof. For a matrix $M \in \mathbb{R}^{n\times p}$ we set

$$v(M, s) \stackrel{\text{def}}{=} \inf\left\{\frac{u'(M'M)u}{n\|u\|_2^2}\ u \neq 0, \|u\|_0 \leq s\right\},$$

and for $\kappa_0 = 1/64$ and $c_0 = 8$, we define

$$\mathcal{E} \stackrel{\text{def}}{=} \left\{M \in \mathbb{R}^{n\times p}:\ v(M, s) \geq \kappa_0,\ \max_{1\leq j\leq p} \|M_j\|_2 \leq 2\sqrt{n},\right.$$
$$\left.\min_{1\leq j\leq p} \|M_j\|_2 \geq \sqrt{\frac{n}{2}},\quad \text{and}\quad \max_{j\neq k} |\langle M_j, M_k\rangle| \leq c_0\sqrt{n\log(p)}\right\}.$$

By Theorem 1 of [Raskutti et al.(2010)], Lemma 1-(4.2) of [Laurent and Massart(2000)], and standard Gaussian deviation bounds, we can find universal constants a, A, such that for $n \geq As\log(p)$, we have $\mathbb{P}(X \notin \mathcal{E}) \leq \frac{a}{p}$. So to obtained the statement of the lemma, it suffices to consider some arbitrary element $X \in \mathcal{E}$ and show that (A.2) holds.

Fix $\delta \in \Delta$ such that $\|\delta\|_0 \leq s$. We set $M_\delta \stackrel{\text{def}}{=} I_n + \frac{1}{\sigma^2\rho_1} X_\delta X_\delta'$, so that $L_\delta = M_\delta + \frac{1}{\sigma^2\rho_0} X_{\delta^c} X_{\delta^c}'$. The Woodbury identity gives

$$X_j' L_\delta^{-1} X_k = X_j' M_\delta^{-1} X_k - \frac{1}{\sigma^2\rho_0} X_j' M_\delta^{-1} X_{\delta^c} \left(I_{\|\delta^c\|_0} + \frac{1}{\sigma^2\rho_0} X_{\delta^c}' M_\delta^{-1} X_{\delta^c} \right)^{-1} X_{\delta^c}' M_\delta^{-1} X_k. \tag{A.3}$$

If $C_1 = \max_\ell X_\ell' M_\delta^{-1} X_\ell$, and $C_0 = \max_{\ell \neq j,\, \delta_j = 0} |X_j' M_\delta^{-1} X_\ell|$, then we deduce easily from (A.3) that for all $j \neq k$ such that $\delta_j = 0$,

$$|X_j' L_\delta^{-1} X_k| \leq C_0 + \frac{1}{\sigma^2\rho_0}\left(C_1^2 + pC_0^2\right). \tag{A.4}$$

In order to proceed, we need to bound the term $X_j M_\delta^{-1} X_k$. Easily, for $X \in \mathcal{E}$, we have

$$X_j' M_\delta^{-1} X_j \leq \|X_j\|_2^2 \leq 4n.$$

Another application of the Woodbury identity gives

$$M_\delta^{-1} = I_n - \frac{1}{\sigma^2\rho_1} X_\delta \left(I_{\|\delta\|_0} + \frac{1}{\sigma^2\rho} X_\delta' X_\delta \right)^{-1} X_\delta'. \tag{A.5}$$

Therefore, for $k \neq j$

$$X_j' M_\delta^{-1} X_k = X_j' X_k - \frac{1}{\sigma^2\rho_1} X_j' X_\delta \left(I_{\|\delta\|_0} + \frac{1}{\sigma^2\rho} X_\delta' X_\delta \right)^{-1} X_\delta' X_k.$$

Using $X \in \mathcal{E}$, we educe for $j \neq k$, and $\delta_j = 0$,

$$\begin{aligned} \frac{1}{\sigma^2\rho_1} \left| X_j' X_\delta \left(I_{\|\delta\|_0} + \frac{1}{\sigma^2\rho} X_\delta' X_\delta \right)^{-1} X_\delta' X_k \right| &\leq \frac{1}{\kappa_0 n} \|X_\delta' X_k\|_2 \|X_\delta' X_j\|_2 \\ &\leq \frac{c_0^2 s \log(p) + c_0\sqrt{s\log(p)}}{\kappa_0} \leq c_0\sqrt{n\log(p)}, \end{aligned}$$

for $n \geq As^2\log(p)$, for some constant A. It follows that

$$|X_j' M_\delta^{-1} X_k| \leq 2c_0\sqrt{n\log(p)}.$$

We combine this with (A.4) to obtain that for $j \neq k$ such that $\delta_j = 0$,

$$|X_j' L_\delta^{-1} X_k| \leq 3c_0\sqrt{n\log(p)} \left(1 + \frac{1}{\sigma^2\rho_0} pc_0\sqrt{n\log(p)} \right) + 16\frac{1}{\sigma^2\rho_0} n^2 \leq 8c_0\sqrt{n\log(p)}, \tag{A.6}$$

for ρ_0 large enough. (A.6) says that $\mathcal{C}_0 \leq 8c_0\sqrt{n\log(p)}$, for $X \in \mathcal{E}$, as claimed.

For j such that $\delta_j = 0$, (A.5) gives

$$
\begin{aligned}
X_j' M_\delta^{-1} X_j &= \|X_j\|_2^2 - \frac{1}{\sigma^2\rho_1} X_j' X_\delta \left(I_{\|\delta\|_0} + \frac{1}{\sigma^2\rho_1} X_\delta' X_\delta \right)^{-1} X_\delta' X_j \\
&\geq \|X_j\|_2^2 - \frac{\|X_\delta' X_j\|_2^2}{n\kappa_0} \\
&\geq \frac{n}{4}, \qquad \text{(A.7)}
\end{aligned}
$$

since $n \geq As\log(p)$, and by taking A large enough ($A \geq 4c_0^2/\kappa_0$). Equation (**??**) then yields

$$
\begin{aligned}
X_j' L_\delta^{-1} X_j \geq X_j' M_\delta^{-1} X_j - \frac{1}{\sigma^2\rho_0} \|X_{\delta^c}' M_\delta^{-1} X_j\|_2^2 \\
= X_j' M_\delta^{-1} X_j - \frac{1}{\sigma^2\rho_0} \left[(X_j' M_\delta^{-1} X_j)^2 + \sum_{k:\ \delta_k = 0, k\neq j} (X_j' M_\delta^{-1} X_k)^2 \right].
\end{aligned}
$$

For $2\rho_0^{-1} \leq \sigma^2$, it follows that

$$
X_j' L_\delta^{-1} X_j \geq \frac{n}{8} - \frac{1}{\sigma^2\rho_0} (p - \|\delta\|_0) \left(4c_0^2 n \log(p)\right),
$$

which together with (A.6) and (A.1) implies that for any $u \in \mathbb{R}^p$ such that $\delta^c \supseteq \mathsf{supp}(u)$, and $\|\mathsf{supp}(u)\|_0 \leq s$, we have

$$
u' X_{\delta^c}' L_\delta^{-1} X_{\delta^c} u \geq \frac{n}{32} \|u\|_2^2, \qquad \blacksquare
$$

as claimed.

Acknowledgements. I'm grateful to Joonha Park for pointing out a mistake in an initial draft of the manuscript.

REFERENCES

[Atchade and Bhattacharyya(2018)] Atchade, Y. and Bhattacharyya, A. (2018). An approach to large-scale Quasi-Bayesian inference with spike-and-slab priors. *arXiv e-prints* arXiv:1803.10282.

[Bakry et al.(2008)] Bakry, D., Barthe, F., Cattiaux, P. and Guillin, A. (2008). A simple proof of the poincaré inequality for a large class of probability measures including the log-concave case. *Electronic Communications in Probability* **13**.

[Bakry et al.(2013)] Bakry, D., Gentil, I. and LeDoux, M. (2013). *Analysis and Geometry of Markov Diffusion Operators*. Springer.

[Bobkov(1999)] Bobkov, S. G. (1999). Isoperimetric and analytic inequalities for log-concave probability measures. *Ann. Probab.* **27** 1903–1921.

[Cattiaux and Guillin(2009)] Cattiaux, P. and Guillin, A. (2009). Trends to equilibrium in total variation distance. *Ann. Inst. H. Poincaré Probab. Statist.* **45** 117–145.

[Diaconis and Stroock(1991)] Diaconis, P. and Stroock, D. (1991). Geometric bounds for eigenvalues of markov chains. *The Annals of Applied Probability* **1** 36–61.

[Dwivedi et al.(2018)] Dwivedi, R., Chen, Y., Wainwright, M. J. and Yu, B. (2018). Log-concave sampling: Metropolis-Hastings algorithms are fast! *arXiv e-prints* arXiv:1801.02309.

[Frieze et al.(1994)] Frieze, A., Kannan, R. and Polson, N. (1994). Sampling from log-concave distributions. *Ann. Appl. Probab.* **4** 812–837.

[Ge et al.(2018)] Ge, R., Lee, H. and Risteski, A. (2018). Simulated Tempering Langevin Monte Carlo II: An Improved Proof using Soft Markov Chain Decomposition. *arXiv e-prints* arXiv:1812.00793.

[George and McCulloch(1997)] George, E. I. and McCulloch, R. E. (1997). Approaches to bayesian variable selection. *Statist. Sinica* **7** 339–373.

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

[Liggett(1991)] Liggett, T. M. (1991). l_2 rates of convergence for attractive reversible nearest particle systems: The critical case. *Ann. Probab.* **19** 935–959.

[Liu(1994)] Liu, J. S. (1994). The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association* **89** 958–966.

[Lovász(1999)] Lovász, L. (1999). Hit-and-run mixes fast. *Math. Program.* **86** 443–461.

[Lovász and Simonovits(1993)] Lovász, L. and Simonovits, M. (1993). Random walks in a convex body and an improved volume algorithm. *Random Structures Algorithms* **4** 359–412.

[Lovász and Vempala(2007)] Lovász, L. and Vempala, S. (2007). The geometry of logconcave functions and sampling algorithms. *Random Structures Algorithms* **30** 307–358.

[Madras and Randall(2002)] Madras, N. and Randall, D. (2002). Markov chain decomposition for convergence rate analysis. *Ann. Appl. Probab.* **12** 581–606.

[Meinshausen and Yu(2009)] Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270.

[Montenegro and Tetali(2006)] Montenegro, R. and Tetali, P. (2006). Mathematical aspects of mixing times in markov chains. *Found. Trends Theor. Comput. Sci.* **1** 237–354.

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

[Raginsky et al.(2017)] Raginsky, M., Rakhlin, A. and Telgarsky, M. (2017). Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, vol. 65 of *Proceedings of Machine Learning Research*. PMLR.

[Raskutti et al.(2010)] Raskutti, G., Wainwright, M. J. and Yu, B. (2010). Restricted eigenvalue properties for correlated gaussian designs. *J. Mach. Learn. Res.* **11** 2241–2259.

[Sinclair(1992)] Sinclair, A. (1992). Improved bounds for mixing rates of markov chains and multicommodity flow. *Combinatorics, Probability and Computing* **1** 351–370.

[Welling and Teh(2011)] Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11, Omnipress, USA.
URL http://dl.acm.org/citation.cfm?id=3104482.3104568

[Yang et al.(2016)] Yang, Y., Wainwright, M. J. and Jordan, M. I. (2016). On the computational complexity of high-dimensional bayesian variable selection. *Ann. Statist.* **44** 2497–2532.