REGULARIZATION AND COMPUTATION WITH HIGH-DIMENSIONAL SPIKE-AND-SLAB POSTERIOR DISTRIBUTIONS

YVES ATCHADÉ AND ANWESHA BHATTACHARYYA

(March 2018)

Abstract. We consider the Bayesian analysis of a high-dimensional statistical model with a spike-and-slab prior, and we study the forward-backward envelope of the posterior distribution – denoted $\hat{\Pi}_{\gamma}$ for some regularization parameter $\gamma > 0$. Viewing $\hat{\Pi}_{\gamma}$ as a pseudo-posterior distribution, we work out a set of sufficient conditions under which it contracts towards the true value of the parameter as $\gamma \downarrow 0$, and $p$ (the dimension of the parameter space) diverges to $\infty$. In linear regression models the contraction rate matches the contraction rate of the true posterior distribution. We also study a practical Markov Chain Monte Carlo (MCMC) algorithm to sample from $\hat{\Pi}_{\gamma}$. In the particular case of the linear regression model, and focusing on models with high signal-to-noise ratios, we show that the mixing time of the MCMC algorithm depends crucially on the coherence of the design matrix, and on the initialization of the Markov chain. In the most favorable cases, we show that the computational complexity of the algorithm scales with the dimension $p$ as $O(pe^{s^*})$, where $s^*$ is the number of non-zeros components of the true parameter. We provide some simulation results to illustrate the theory. Our simulation results also suggest that the proposed algorithm (as well as a version of the Gibbs sampler of Narisetty and He (2014)) mix poorly when poorly initialized, or if the design matrix has high coherence.

1. Introduction

Suppose that we wish to infer a parameter $\theta \in \mathbb{R}^p$ from a random sample $Z \in \mathcal{Z}$, based on the statistical model $Z \sim f_{\theta}(z)dz$, where $f_{\theta}$ is a density on a sample space...
$Z$ equipped with a reference sigma-finite measure $dz$. The log-likelihood function
$$\ell(\theta; z) \defeq \log f_\theta(z), \quad \theta \in \mathbb{R}^p, \quad z \in Z,$$
is assumed to be a jointly measurable function on $\mathbb{R}^p \times Z$. We take a Bayesian approach, and consider a setting where $p$ is very large and it is statistically appealing to perform a variable selection step in the estimation process. This problem has attracted a lot of attention in recent years, and several approaches are available. One of the most effective solution – at least in theory – relies on spike-and-slab priors (Mitchell and Beauchamp (1988); George and McCulloch (1997)), and can be described as follows. For $\delta \in \Delta \defeq \{0, 1\}^p$, let $\mu_\delta(d\theta)$ denote the product measure on $\mathbb{R}^p$ given by
$$\mu_\delta(d\theta) \defeq \prod_{j=1}^p \mu_{\delta_j}(d\theta_j),$$
where $\mu_0(dz)$ is the Dirac mass at 0, and $\mu_1(dz)$ is the Lebesgue measure on $\mathbb{R}$. With $\{\omega_{\delta}, \delta \in \Delta\}$ denoting a prior distribution on $\Delta$, we consider the spike-and-slab prior distribution on $\Delta \times \mathbb{R}^p$ given by
$$\omega_{\delta} \left(\frac{p}{2}\right)^{\|\delta\|_1} e^{-\rho\|\theta\|_1} \mu_\delta(d\theta), \quad (1)$$
for a parameter $\rho > 0$. Given $Z = z$, the resulting posterior distribution on $\Delta \times \mathbb{R}^p$ is
$$\Pi(\delta, d\theta|z) \propto f_\theta(z) \omega_{\delta} \left(\frac{p}{2}\right)^{\|\delta\|_1} e^{-\rho\|\theta\|_1} \mu_\delta(d\theta). \quad (2)$$

The posterior distribution (2) has been recently studied in the high-dimensional regime (see e.g. Castillo and van der Vaart (2012); Castillo et al. (2015); Atchadé (2017)), where it is shown to contract towards the true value of the parameter at an optimal rate. In practice however, this posterior is computationally difficult to handle and typically require specialized MCMC techniques such as reversible jump (Green (2003)), and related methods (Gottardo and Raftery (2008); Schreck et al. (2013)). However these MCMC algorithms are often difficult to design and tune, particularly in high-dimensional problems (see for instance Schreck et al. (2013) and Atchadé (2015) for some numerical comparisons). In the particular case of the Gaussian linear model with a Gaussian slab density (that is we replace the Laplace density by a mean-zero Gaussian density in (1)), it is possible to side-step this computational difficulties by integrating out the parameter $\theta$. One can then explore the resulting

1The use of the Laplace distribution is not fundamental. We use it here partly for mathematical and computational convenience, and partly because of its widespread use in the applications. Much of the results below can also be worked out if the Laplace distribution is replaced by a density $g$ such that $g(0) = 1$, and $\log g$ is Lipschitz – note however that this condition is not satisfied by the Gaussian distribution.
discrete distribution $\delta \mapsto \Pi(\delta | z)$ using standard MCMC algorithms (see for instance George and McCulloch (1997); Bottolo and Richardson (2010); Yang et al. (2016) and the reference therein). However this strategy is typically not widely applicable.

A very popular approximation to $\Pi$ – sometimes also referred to as spike-and-slab – is obtained by replacing the point mass at zero by a Gaussian distribution with a small variance $\gamma$, say, (George and McCulloch (1997); Ishwaran and Rao (2005); Rockova and George (2014); Narisetty and He (2014)). The resulting posterior distribution is

$$
\bar{\Pi}_\gamma(\delta, du| z) \propto \omega_{\delta}(2\pi \gamma)^{\|\delta\|_0} \left( \frac{\rho}{2} \right)^{\|\delta\|_0} e^{-\ell(u;z) - \rho \|u\|_1 - \frac{1}{2\gamma} \|u - u_0\|^2_2} du.
$$

(3)

This approximation was recently studied in Narisetty and He (2014), where linear regression model consistency is established in the high-dimensional setting. Here we consider another approximation scheme. We consider the variational approximation of $\Pi$ proposed in Atchade (2015) – denoted $\Pi_\gamma$ – that is obtained by taking its forward-backward envelop (an envelop function similar to the Moreau envelop). One advantage of working with $\Pi_\gamma$ is that it is computationally and mathematically tractable. Indeed, the definition of $\Pi_\gamma$ leads to tight approximation bounds that we leverage in the analysis. In our numerical experiments we found that $\Pi_\gamma$ and $\bar{\Pi}_\gamma$ behave very similarly, and to some extent we expect the results derived here to hold when applied to $\Pi_\gamma$.

To define $\Pi_\gamma$, we endow the Euclidean space $\mathbb{R}^p$ with its Lebesgue measure denoted $d\theta$ or $du$, etc. For a set $A \subseteq \mathbb{R}^p$, let $\iota_A$ denote its characteristic function defined here as $\iota_A(u) = 0$ if $u \in A$, and $\iota_A(u) = +\infty$ otherwise. Given $\delta \in \Delta$, let $\mathbb{R}^p_\delta \defeq \{ \delta \cdot u, \ u \in \mathbb{R}^p \}$ (where $\delta \cdot u \defeq (u_1 \delta_1, \ldots, u_p \delta_p) \in \mathbb{R}^p$). The basic idea is to replace the function $u \mapsto -\ell(u;z) + \rho \|u\|_1 + \iota_{\mathbb{R}^p_\delta}(u)$ by its forward-backward envelop (Patrinos et al. 2014) defined as

$$
h_\gamma(\delta, \theta; z) \defeq \min_{v \in \mathbb{R}^p_\delta} \left[ -\ell(\theta; z) - \langle \nabla \ell(\theta; z), v - \theta \rangle + \rho \|v\|_1 + \frac{1}{2\gamma} \|v - \theta\|^2_2 \right],
$$

(4)

for some parameter $\gamma > 0$. This envelop function is closely related to the more widely known Moreau envelop (Moreau (1965); Bauschke and Combettes (2011); Parikh and Boyd (2013)). It is easily seen from the definition that

$$
h_\gamma(\delta, \cdot; z) \leq -\ell(\cdot; z) + \rho \|\cdot\|_1 + \iota_{\mathbb{R}^p_\delta}(\cdot),
$$

(5)

and furthermore $h_\gamma(\delta, \cdot; z)$ converges pointwise to $-\ell(\cdot; z) + \rho \|\cdot\|_1 + \iota_{\mathbb{R}^p_\delta}(\cdot)$ as $\gamma \downarrow 0$ (see for instance Patrinos et al. (2014); Atchade (2015)). Assuming that $\int_{\mathbb{R}^p} e^{-h_\gamma(\delta,u;Z)} du$ is finite for all $\delta \in \Delta$, we are then naturally led to the probability distribution on $\Delta \times \mathbb{R}^p$ given by

$$
\Pi_\gamma(\delta, du| Z) \propto \omega_{\delta}(2\pi \gamma)^{\|\delta\|_0} \left( \frac{\rho}{2} \right)^{\|\delta\|_0} e^{-h_\gamma(\delta,u;Z)} du,
$$

(6)
that we take as an approximation of Π. It is not automatic that \( \hat{\Pi}_γ \) is a well-defined probability distribution. This needs to be checked on a case by case basis. Since \( h_γ \) is known to approximate \(-\ell(\cdot; z) + \rho \| \cdot \|_1 + \iota_{\mathbb{R}^p}(\cdot)\), we naturally expect \( \hat{\Pi}_γ \) to behave like Π, for \( γ \) small (Atchadé (2015)).

1.1. Main contributions. The contribution of this work is two-fold. Firstly, viewing \( \hat{\Pi}_γ \) as a pseudo/quasi-posterior distribution, we study its statistical properties as the dimension \( p \) increases. We derive some sufficient conditions under which \( \hat{\Pi}_γ \) is well defined and puts most of its probability mass around the true value of the parameter as \( p \to \infty \). More precisely, Theorem 5 can be used to show that a draw \((\delta, \theta)\) from \( \hat{\Pi}_γ(\cdot|Z) \) produces a typically sparse vector \( \delta \), and a non-sparse vector \( \theta \). Moreover the components of \( \theta \) for which \( \delta_j = 0 \) are typically small \( (O(\sqrt{\gamma})) \), whereas the sparsified version of \( \theta \) (that is \( \theta_\delta = \theta \cdot \delta \)) is typically close to \( \theta_* \), the true value of the parameter.

We also study the contraction rate of \( \hat{\Pi}_γ \), and using the linear regression model as an example, we show that the rate of contraction of \( \hat{\Pi}_γ \) matches that of Π as derived in Castillo et al. (2015). Furthermore we show that \( \hat{\Pi}_γ \) enjoys a Bernstein-von Mises approximation, and again in the particular case of the linear regression model, we recover the Bernstein-von Mises theorem established by Castillo et al. (2015), up to some small difference in the Fisher information matrix due to the approximate nature of \( \hat{\Pi}_γ \).

Practical use of Bayesian procedures typically hinges on the ability to draw samples from the posterior distribution. In the second part of the paper we develop a highly efficient Metropolis-within-Gibbs algorithm to sample from \( \hat{\Pi}_γ \). Building on the posterior contraction properties developed in the first part, we analyze the computational complexity of the algorithm as \( p \to \infty \), focusing on the linear regression case. The behavior of the resulting Markov chain depends heavily on the signal-to-noise ratio of the underlying regression problem, on the initial distribution of the Markov chain, and on the coherence of the design matrix \( X \in \mathbb{R}^{n \times p} \), defined as

\[
\mathcal{C}(X) \overset{\text{def}}{=} \sup_{j: \delta_* \cdot 1 = 1} \frac{1}{\sqrt{n}} \| X'_{\delta_*} X_j \|_2,
\]

where \( X_j \) is the \( j \)-th column of \( X \), and \( X_{\delta_*} \) is the sub-matrix of \( X \) obtained by removing the columns of \( X \) for which \( \delta_* \cdot 1 = 1 \). Figure 1-(a)-(c) show the estimated mixing time of the algorithm (truncated at \( 2 \times 10^4 \)) as function of \( p \) under different scenarios, and illustrate the main conclusions of the paper. For comparison we also present – in dashed lines – the mixing time of a similar algorithm to sample from the weak spike-and-slab posterior distribution (3). We refer the reader to Section 3.4 for a detailed description of the experiment. Figure 1-(a) shows the mixing times in a setting where the design matrix \( X \) has low coherence, and a good initial value is used.
to start the algorithm. In Figure 1(b) the initial value remains the same, but the design matrix $X$ has a higher coherence parameter (see Section 3.4 for the details on how such design matrix is produced). Finally in Figure 1(c), the design matrix is the same as in Figure 1(a) but the initial model $\delta$ used to initialize the algorithm now has 20% false-negatives, but no false-positive. Here most of the mixing times observed are greater than the 20,000 iterations mark.

Two observations stand out from these results. The algorithm seems to mix quickly when the design matrix $X$ has low coherence and the initial distribution of the Markov chain has no false-negatives. The mixing seems to degrade when the coherence increases, and the algorithm seems to mix very poorly when the initial distribution has false-negatives. We show in Theorem 15 a result that partly supports these empirical findings. More precisely, we show that with an initial distribution without false-negatives, the mixing time of the algorithm scales with $p$ as

$$O \left( p \exp \left( c s^2 \left[ 1 + \frac{C(X)}{s^2 \log(p)} \left( \frac{p}{n} \right)^2 \right] \right) \right),$$

where $s^* = \|\theta^*\|_0$, and $c$ an absolute constant.

Our result is similar to the recent work by Yang et al. (2016) which studied a Metropolis-Hastings algorithm to sample from the marginal distribution of $\delta$ in a linear regression model with Gaussian spike-and-slab $g$-prior. They showed that their

\[\text{Figure 1. Estimated mixing times as function of the dimension } p. \text{ (a) low coherence, good initialization. (b) high coherence, good initialization. (c) low-coherence poor initialization.}\]
algorithm has a worst case (wrt the initial distribution) mixing time that scales as $O(s^2(n + s)p \log(p))$. Although the paper does not highlight it, it seems that their result implicitly assumes a low-coherence design matrix $X$, via their assumption $D(s_0)$. This would then be consistent with our finding. We note that the posterior distribution considered by Yang et al. (2016) can be viewed as a “collapsed” version of ours, where the regression parameters are integrated out, and the benefits of collapsing variables in a block update sampling problem is well-known (Liu (1994)). The robustness with respect to the initial distribution in the algorithm of Yang et al. (2016) can perhaps be understood in that sense. However we stress again that the idea of collapsing parameters from the posterior distribution is not always feasible, particularly in generalized linear regression models.

1.2. Outline of the paper. The paper is organized as follows. We study the statistical properties of $\hat{\Pi}_\gamma$ in Section 2. We derive two main theorems. Theorem 5 deals with the contraction and the contraction rate of $\hat{\Pi}_\gamma$ as $p \to \infty$, whereas Theorem 8 studies variable selection and the Bernstein-von Mises theorem. We illustrate these results in the particular case of the linear regression model in Section 2.2, leading to Corollary 10. Since the proofs of the results of Section 2 follow similar techniques as in Castillo et al. (2015) and Atchade (2017), the details are placed in a supplementary document. In Section 3 we study the problem of sampling from $\hat{\Pi}_\gamma$, using a slightly modified version of the MCMC algorithm of Atchadé (2015). In the particular case of the linear regression model, we study the mixing of the proposed algorithm. The main result there is Theorem 15, the proof of which is developed in Section 4, with some technical details gathered in the supplement. Some numerical simulations are detailed in Section 3.4 to illustrate the results.

1.3. Notation. Throughout we equip the Euclidean space $\mathbb{R}^p$ ($p \geq 1$ integer) with its usual Euclidean inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|_2$, its Borel sigma-algebra, and its Lebesgue measure denoted $du$ (or $d\theta$ etc...). All vectors $u \in \mathbb{R}^p$ are column-vectors unless stated otherwise. We also use the following norms on $\mathbb{R}^p$: $\|\theta\|_1 \overset{\text{def}}{=} \sum_{j=1}^p |\theta_j|$, $\|\theta\|_0 \overset{\text{def}}{=} \sum_{j=1}^p 1\{\theta_j > 0\}$, and $\|\theta\|_\infty \overset{\text{def}}{=} \max_{1 \leq j \leq p} |\theta_j|$.

We set $\Delta = \{0, 1\}^p$. For $\theta, \theta' \in \mathbb{R}^p$, $\theta \cdot \theta' \in \mathbb{R}^p$ denotes the component-wise product of $\theta$ and $\theta'$. For $\delta \in \Delta$, we set $\mathbb{R}_\delta^p \overset{\text{def}}{=} \{\theta : \theta \in \mathbb{R}^p\}$, and we write $\theta_\delta$ as a short for $\theta_\delta \cdot \delta$. We define $\delta^c \overset{\text{def}}{=} 1 - \delta$, that is $\delta_j^c = 1 - \delta_j$, $1 \leq j \leq p$. For a matrix $A \in \mathbb{R}^{m \times m}$ and $\delta \in \Delta$, $A_\delta$ (resp. $A_{\delta^c}$) denotes the matrix of $\mathbb{R}^{\|\delta\|_0 \times \|\delta\|_0}$ (resp. $\mathbb{R}^{(m - \|\delta\|_0) \times (m - \|\delta\|_0)}$) obtained by keeping only the rows and columns of $A$ for which $\delta_j = 1$ (resp. $\delta_j = 0$). For $\delta, \delta' \in \Delta$, we write $\delta \supseteq \delta'$ to mean that for any $j \in \{1, \ldots, p\}$, whenever $\delta'_j = 1$, we have $\delta_j = 1$. 

Throughout the paper, \( e \) denotes the Euler number, and \( \binom{m}{q} \) is the combinatorial number \( m!/(q!(m-q)! \). For \( x \in \mathbb{R} \), \( \text{sign}(x) \) is the sign of \( x \) (\( \text{sign}(x) = 1 \) if \( x > 0 \), \( \text{sign}(x) = -1 \) if \( x < 0 \), and \( \text{sign}(x) = 0 \) if \( x = 0 \)).

If \( f(\theta, x) \) is a real-valued function that depends on the parameter \( \theta \) and some other argument \( x \), the notation \( \nabla^{(k)} f(\theta, x) \), where \( k \) is an integer, denotes the \( k \)-th partial derivative of \( f \) with respect to \( \theta \). For \( k = 1 \), we write \( \nabla f(\theta, x) \) instead of \( \nabla^{(1)} f(\theta, x) \).

The total variation metric between two probability measures \( \mu, \nu \) is defined as
\[
\|\mu - \nu\|_{\text{tv}} \overset{\text{def}}{=} \sup_{A \text{ meas.}} (\mu(A) - \nu(A)).
\]

All the asymptotic results in the paper are derived by letting the dimension \( p \) grow to infinity, and we say that a term \( x \in \mathbb{R} \) is an absolute constant if \( x \) does not depend on \( p \).

2. Contraction properties of \( \tilde{\Pi}_\gamma \)

In this section we will establish that under some mild conditions, \( \tilde{\Pi}_\gamma \) is a well-defined probability distribution that has similar posterior contraction properties as the spike-and-slab posterior distribution given in [2]. We make the following assumptions.

**H1.** For all \( z \in \mathbb{Z} \), the function \( \theta \mapsto \ell(\theta; z) \) is concave, twice differentiable, and there exist symmetric positive semidefinite matrices \( S, \bar{S} \in \mathbb{R}^{p \times p} \) that does not depend on \( z \) and \( \theta \), such that for all \( \theta \in \mathbb{R}^p \), and all \( z \in \mathbb{Z} \),
\[
-\bar{S} \preceq \nabla^{(2)} \log f_\theta(z) \preceq -S,
\]
where the notation \( A \preceq B \) means that \( B - A \) is positive semidefinite.

**Remark 1.** We note that one can always take \( \bar{S} \) as the zero matrix, since \( \ell \) is concave. Hence H1 mainly requires that the hessian matrix of the log-likelihood function is lower bounded uniformly in \( \theta \) and \( z \) by the matrix \( -\bar{S} \). Although somewhat restrictive, this assumption is enough to handle linear and logistic regression models. It is unclear whether the results developed below can be extended more broadly beyond H1. \( \square \)

For integer \( s \geq 1 \), we set
\[
\tilde{\kappa}(s) \overset{\text{def}}{=} \sup \left\{ \frac{u'\bar{S}u}{\|u\|^2} : u \in \mathbb{R}^p \setminus \{0\}, \|u\|_0 \leq s \right\},
\]
and
\[
\kappa(s) \overset{\text{def}}{=} \inf \left\{ \frac{u'Su}{\|u\|^2} : u \in \mathbb{R}^p \setminus \{0\}, \|u\|_0 \leq s \right\}.
\]
and we convene that \( \bar{\kappa}(0) = 0, \kappa(0) = +\infty \). For \( s = p \), \( \bar{\kappa}(s) \) is the largest eigenvalue of \( \bar{S} \) that we write as \( \lambda_{\max}(\bar{S}) \).

Following a standard approach in Bayesian asymptotics, we will assume that there exists a true value of the parameter \( \theta^\star \) such that \( Z \sim f_{\theta^\star} \). More precisely we assume the following.

**H2.** There exists \( \theta^\star \in \mathbb{R}^p \setminus \{0\} \) such that \( Z \sim f_{\theta^\star} \), and we set \( s^\star \overset{\text{def}}{=} \|\theta^\star\|_0 \).

Under **H2** we expect \( \theta^\star \) to be close to the maximizer of the log-likelihood \( \theta \mapsto \ell(\theta; Z) \). That is, we expect \( \nabla \ell(\theta^\star; Z) \approx 0 \). Therefore the sets \( \mathcal{E}_c \overset{\text{def}}{=} \{z \in Z : \|\nabla \ell(\theta^\star; z)\|_\infty \leq c/2\} \), for \( c > 0 \) will naturally play an important role in the analysis. Throughout the paper, we write \( \delta^\star \in \Delta \) to denote the sparsity structure of \( \theta^\star \) (that is \( \delta^\star_j = 1\{|\theta^\star,j| > 0\}, j = 1, \ldots, p \)). We will write \( \mathbb{P}^\star \) (resp. \( \mathbb{E}^\star \)) to denote the probability distribution (resp. expectation operator) of the random variable \( Z \in Z \) assumed in **H2**.

As a prior distribution on \( \delta \), we assume that the \( \delta_j \) are independent Bernoulli random variables. More precisely, we assume the following.

**H3.** For some absolute constant \( u > 0 \), setting \( q \overset{\text{def}}{=} \frac{1}{p^{1+u}} \), we assume that \( \omega_\delta = p \prod_{j=1}^p q_j \delta_j (1-q)^{1-\delta_j} \).

**Remark 2.** The prior \( \omega_\delta \) induced by **H3** can be written as \( \omega_\delta = g_{\|\delta\|_0} \frac{1}{\|\delta\|_0!} \), where \( g_s = \binom{p}{s} q^s (1-q)^{p-s} \). It is then easily checked that

\[
    g_s \leq \left( \frac{2}{p^u} \right) g_{s-1}, \quad s = 1, \ldots, p. \tag{7}
\]

Discrete priors \( \{\omega_\delta, \delta \in \Delta\} \) of the form \( \omega_\delta = g_{\|\delta\|_0} \frac{1}{\|\delta\|_0!} \) where \( \{g_s\} \) satisfies conditions of the form (7) were introduced by Castillo and van der Vaart (2012), and shown to work well for high-dimensional problems. □

The ability to recover \( \theta^\star \) depends on the quantity of information available, an idea that we formalized by imposing appropriate restricted strong concavity condition on the log-likelihood \( \ell \) via the function

\[
    \mathcal{L}_\gamma(\delta, \theta; z) \overset{\text{def}}{=} \ell(\theta; z) - \ell(\theta^\star; z) - \langle \nabla \ell(\theta^\star; z), \theta - \theta^\star \rangle + 2\gamma \|\delta \cdot \nabla \ell(\theta^\star; z) - \delta \cdot \nabla \ell(\theta^\star; z)\|_2^2. \tag{8}
\]

The function \( \theta \mapsto \ell(\theta; z) - \ell(\theta^\star; z) - \langle \nabla \ell(\theta^\star; z), \theta - \theta^\star \rangle \) is the well-known Bregman divergence associated to \( \ell \). The additional term \( 2\gamma \|\delta \cdot \nabla \ell(\theta^\star; z) - \delta \cdot \nabla \ell(\theta^\star; z)\|_2^2 \) in
is due to the Moreau envelop approximation, and would typically be small for \( \gamma \) small, and \( \delta \) sparse. For our results to work with a certain generality, we use the concept of a rate function introduced in \cite{Atchade17}. A continuous function \( r : [0, +\infty) \to [0, +\infty) \) is called a rate function if \( r(0) = 0 \), \( r \) is increasing and \( \lim_{x \downarrow 0} r(x)/x = 0 \).

**H4.** There exist \( \gamma > 0, \bar{\rho} > 0, \bar{s} \in \{s_*, \ldots, p\} \), and a rate function \( r_1 \) such that for all \( \delta \in \Delta \) such that \( \|\delta\|_0 \leq \bar{s} \), and for all \( \theta \in \mathbb{R}^p \), we have

\[
\log \mathbb{E}_x \left[ e^{\mathcal{L}_\gamma (\delta; \theta; Z)} 1_{E_\bar{\rho}}(Z) \right] \leq -\frac{1}{2} r_1(\|\theta - \theta_*\|_2).
\]

**Remark 3.** We view H4 as a form of restricted strong convexity assumption on the function \(-\ell_\gamma\) \cite{Negahban12}. This assumption is much weaker than the assumption that \( \ell \) is strongly concave. We will see in Section 2.2 that H4 holds for linear regression models. It can also be shown to hold for logistic regression models, although we do not pursue this here.

**Remark 4.** A lower bound on \( \mathcal{L}_\gamma \) will also be needed below. We note that by Assumption H1, we have

\[
\epsilon \overset{\text{def}}{=} \inf \left\{ x > 0 : r_1(z) \geq 3\bar{\rho}(s_* + \bar{s})^{1/2}z \text{ for all } z \geq x \right\}.
\]

We then set \( \Delta_{\bar{s}} \overset{\text{def}}{=} \{ \delta \in \Delta : \|\delta\|_0 \leq \bar{s} \} \), and

\[
\widehat{B}_{m,M} \overset{\text{def}}{=} \bigcup_{\delta \in \Delta_{\bar{s}}} \left( \{\delta\} \times B_{m,M}^{(\delta)} \right),
\]

where

\[
B_{m,M}^{(\delta)} \overset{\text{def}}{=} \left\{ \theta \in \mathbb{R}^p : \|\theta_\delta - \theta_*\|_2 \leq M\epsilon, \|\theta - \theta_\delta\|_2 \leq 2\sqrt{(m+1)\gamma p}, \right. \\
\left. \quad \text{and } \|\theta - \theta_0\|_\infty \leq 2\sqrt{(m+1)\gamma \log(p)} \right\},
\]

for absolute constants \( m, M \).
Theorem 5. Assume $H_1-H_4$. Take $\rho \in (0, \bar{\rho}]$, and $\gamma > 0$ as in $H_4$ such that
\[ 4\gamma \lambda_{\text{max}}(\bar{S}) \leq 1, \quad \text{and} \quad \gamma \rho^2 = o(\log(p)), \quad \text{as} \quad p \to \infty. \quad (13) \]
Suppose also that $\epsilon$ as defined in (11) is finite, and $\Pi_{\gamma}(\cdot | z)$ is well-defined for $\bar{\Pi}*$-almost all $z \in E_{\bar{\rho}}$, and for all $p$ large enough. Then there exists an absolute constant $A_0$ such that for $p \geq A_0$, for all $m \geq 1, M > 2$, and for any measurable subset $E \subseteq E_{\bar{\rho}}$ of $\mathbb{Z}$,
\[
\mathbb{E}^* \left[ \bar{\Pi}_{\gamma} (\mathbb{B}_{m,M} | Z) \mid Z \in \mathcal{E} \right] \geq 1 - \mathbb{E}^* \left[ \bar{\Pi}_{\gamma}(\|\delta\|_0 > \bar{s}| Z) \mid Z \in \mathcal{E} \right]
- 4e^{\frac{a}{2}}p^{(2+u)s_\star} \left( 1 + \frac{\bar{\kappa}(\bar{s}_\star)}{\rho_2^2} \right)^{\bar{s}_\star} \sum_{j \geq 1} e^{-\frac{a}{2} \rho_1 \left( \frac{jM}{\rho_2} \right)^{\bar{s}_\star}} \left( \frac{4\gamma \rho^2 \bar{s}}{p^m} \exp \left( \frac{a}{2} + 3\gamma \rho^2 \bar{s} + 2\gamma \bar{s}^2 (M \epsilon)^2 \right) \right),
\]
provided that $\mathbb{P}^* (Z \in \mathcal{E}) \geq 1/2$.

Proof. See Section A of the Supplement. \qed

The theorem applies naturally with $\mathcal{E} = \mathcal{E}_{\bar{\rho}}$. The general case $\mathcal{E} \subseteq \mathcal{E}_{\bar{\rho}}$ will be needed below. We note that the probabilities of the events $\{ Z \in \mathcal{E}_{\bar{\rho}} \}$ are relatively easy to control, so one can easily derive an unconditional version of the theorem. Theorem outlines a set of sufficient conditions under which one can guarantee that given the event $Z \in \mathcal{E}$, if $(\delta, \theta) \sim \bar{\Pi}_{\gamma} (\cdot | Z)$ then the sparsified vector $\theta_\delta$ satisfies $\| \theta - \theta_\delta \|_2 \leq M \epsilon$, and $\| \theta - \theta_\delta \|_\infty \leq 2\sqrt{(m+1)\gamma \rho} \quad \text{and} \quad \| \theta - \theta_\delta \|_\infty \leq 2\sqrt{(m+1)\gamma \log(p)}$ with high probability. Controlling both the $\ell^0$ and the $\ell^2$ norm gives us a better control of the term $\theta - \theta_\delta$, which will be needed in the mixing time analysis. To use the theorem one needs to establish by other means that $\bar{\Pi}_{\gamma} (\cdot | z)$ is well-defined and puts vanishing probability on the set $\{ \| \delta \|_0 > \bar{s} \}$. This question is addressed in Lemma 21 under some mild additional assumptions.

The second condition in (13) is readily satisfied in most cases. Hence the result implies that one should choose $\gamma$ as
\[ \gamma = \frac{\gamma_0}{4\lambda_{\text{max}}(S)}, \]
for some user-defined absolute constant $\gamma_0 \in (0, 1]$, provided that this choice of $\gamma$ also satisfies $H_4$.

Remark 6. It is interesting to note that we did not directly rely on any sample size condition in the theorem. Here the amount of information available about $\theta$ is formalized in $H_4$ directly in terms of curvature of the log-likelihood function $\ell$. In models
with independent samples, typically translates into a sample size requirement. We refer the reader to Section 2.2 for an example with linear regression models.

2.1. Model selection and Bernstein-von Mises approximation. With some additional assumptions we show next that the distribution \( \hat{\Pi}_\gamma \) puts overwhelming probability mass around \((\delta_*, \theta_*)\) and satisfies a Bernstein-von Mises approximation. To that end we will assume that there exists an absolute constant \( M > 2 \) such that

\[
\theta_* \overset{\text{def}}{=} \min \{|\theta_{s,j}| : \delta_{s,j} = 1\} > M \epsilon. \tag{14}
\]

Clearly this assumption is unverifiable in practice since \( \theta_* \) is typically not known. However a strong signal assumption such as (14) seems to be needed in one form or the other for model selection (Narisetty and He (2014); Castillo et al. (2015); Yang et al. (2016)). For \( \delta \in \Delta \), we recall that the notation \( \delta \supseteq \delta_* \) means that for all \( 1 \leq j \leq p \), \( \delta_{s,j} = 1 \) implies that \( \delta_j = 1 \). We note that when (14) holds then the set \( \mathcal{B}_{m,M}^{(\delta)} \) is empty if \( \delta \) does not satisfy \( \delta \supseteq \delta_* \). Hence one immediate implication of assumption (14) is that the set \( \tilde{\mathcal{B}}_{m,M} \) reduces to

\[
\tilde{\mathcal{B}}_{m,M} = \bigcup_{\delta \in \mathcal{A}} \left( \{ \delta \} \times \mathcal{B}_{m,M}^{(\delta)} \right),
\]

where

\[
\mathcal{A} \overset{\text{def}}{=} \{ \delta \in \Delta : \delta \supseteq \delta_* \} \quad \text{and} \quad \|\delta\|_0 \leq \bar{s} \}.
\]

In other words when (14) and the assumptions of Theorem 5 hold, \( \Pi_\gamma \) puts most of its mass only on sparse models that contains the true model \( \delta_* \). Therefore correct model selection becomes possible if the prior distribution offsets the natural tendency to overfit. More precisely, for \( \Lambda > 0 \) we introduce the set

\[
\mathcal{E}_{\tilde{\rho}, \Lambda} \overset{\text{def}}{=} \mathcal{E}_{\rho} \bigcap_{k=0}^{\bar{s} - s_*} \left\{ z \in \mathcal{Z} : \sup_{\delta \in \mathcal{A} : \|\delta\|_0 = s_* + k} \ell(\hat{\theta}_\delta(z); z) - \ell(\bar{\theta}_\delta(z); z) \leq k \Lambda \right\}.
\]

Given \( \theta \in \mathbb{R}^p \), and \( \delta \in \Delta \setminus \{0\} \), we write \([\theta]_\delta\) to denote the set of \( \delta \)-selected components of \( \theta \) listed in their order of appearance: \([\theta]_\delta = (\theta_j, \delta_j = 1) \in \mathbb{R}^{|\delta||\delta|_0} \). Conversely, if \( u \in \mathbb{R}^{|\delta||\delta|_0} \), we write \((u, 0)_\delta\) to denote the element of \( \mathbb{R}^p \) such that \((u, 0)_\delta = u \). We define the function \( \ell^{[\delta]}(\cdot; z) : \mathbb{R}^{|\delta||\delta|_0} \to \mathbb{R} \) by \( \ell^{[\delta]}(u; z) \overset{\text{def}}{=} \ell((u, 0)_\delta; z) \). We then introduce

\[
\hat{\theta}_\delta(z) \overset{\text{def}}{=} \operatorname{Argmax}_{u \in \mathbb{R}^{|\delta||\delta|_0}} \left[ \ell^{[\delta]}(u; z) \right], \quad z \in \mathcal{Z}. \tag{15}
\]

When \( \delta = \delta_* \), we sometimes write \( \hat{\theta}_*(z) \) instead of \( \hat{\theta}_{\delta_*}(z) \). At times, to shorten the notation we will omit the data \( z \) and write \( \hat{\theta}_{\delta} \) instead of \( \hat{\theta}_\delta(z) \). Omitting the data \( z \), we write \( \mathcal{I}_{\gamma, \delta} \in \mathbb{R}^{|\delta||\delta|_0 \times |\delta||\delta|_0} \) to denote the matrix of second derivatives of \( u \mapsto \ell^{[\delta]}(u; z) \)
evaluated at $\hat{\theta}_s(z)$. When $\delta = \delta_*$, we simply write $\mathcal{I}_\gamma$. We make the following assumption.

**H5.** The integer $\bar{s}$ in $H\bar{s}$ is such that $\kappa(\bar{s}) > 0$, and the following holds.

1. For all $\delta \in \Delta_{\bar{s}}$, and $z \in \mathcal{E}_\rho$, the estimate $\hat{\theta}_\delta(z)$ is well-defined and satisfies
   \[ \|\hat{\theta}_\delta(z) - [\theta_*]_\delta\|_2 \leq \epsilon, \]
   where $\epsilon$ is as defined in (11).

2. For all $z \in \mathcal{Z}$, all $\delta \in \Delta_{\bar{s}}$, the function $u \mapsto \ell^{[\delta]}(u; z)$ is thrice differentiable on $\mathbb{R}^{\|\delta\|_0}$, and for all $c > 0$,
   \[ \varpi_{2,c} \overset{\text{def}}{=} \sup_{z \in \mathcal{E}_\rho} \sup_{\delta \in \Delta_{\bar{s}}} \sup_{u \in \mathbb{R}^{\|\delta\|_0}} \|\nabla^{(3)}\ell^{[\delta]}(u; z)\|_{\text{op}}, \]
   is finite, where $\|\cdot\|_{\text{op}}$ denotes the operator norm\(^3\).

**Remark 7.** When $H\bar{s}$ holds, $H\bar{s}(1)$ is typically easy to check. Indeed, for all $\delta \in \Delta_{\bar{s}}$, and $z \in \mathcal{E}_\rho$, we have
\[ 0 \geq -\ell^{[\delta]}(\hat{\theta}_\delta; z) + \ell^{[\delta]}([\theta_*]_\delta; z) \geq \left\langle -\nabla \ell^{[\delta]}([\theta_*]_\delta; z), \hat{\theta}_\delta - [\theta_*]_\delta \right\rangle + \frac{\kappa(\bar{s})}{2} \|\hat{\theta}_\delta - [\theta_*]_\delta\|_2^2, \]
where the first inequality uses the definition of $\hat{\theta}_\delta$, whereas the second inequality uses $H\bar{s}$ and the definition of $\kappa(\bar{s})$. Hence for all $\delta \in \Delta_{\bar{s}}$, and $z \in \mathcal{E}_\rho$, we have
\[ \|\hat{\theta}_\delta(z) - [\theta_*]_\delta\|_2 \leq \frac{\bar{s}^{1/2}}{\kappa(\bar{s})}, \] which gives an easy way to checking $H\bar{s}(1)$ by comparing the right hand side of (16) to $\epsilon$. $H\bar{s}(2)$ is hard to check in general since it requires the third derivative of the log-likelihood to be uniformly bounded in $z$ (for all practical purposes). However it trivially holds for linear regression models since in that case $\varpi_{2,c} = 0$. It can also be shown to hold for logistic regression models, although we do not pursue this here. \(\square\)

We introduce the following distribution on $\Delta \times \mathbb{R}^p$,
\[ \Pi^\infty_{\gamma}(\delta, d\theta|z) \overset{\text{def}}{=} \pi e^{-\frac{1}{2}([\theta]_\delta - [\theta_*])'\mathcal{I}_\gamma([\theta]_\delta - [\theta_*]) - \frac{1}{2\gamma}([\theta]_\delta - [\theta_*])'(I_p - \gamma \mathcal{Z})([\theta]_\delta - [\theta_*])} \mathbf{1}_{\{\delta_*\}}(\delta) \mathbf{1}_{B_{m,M}^{(\delta_*)}}(\theta)d\theta. \] (17)

Note that the $\delta$-marginal of $\Pi^\infty_{\gamma}(.|z)$ is the point-mass at $\delta_*$. Furthermore, for $p$ large the restriction on the set $B_{m,M}^{(\delta_*)}$ in $\Pi^\infty_{\gamma}(.|z)$ is inconsequential and can be removed –

\(^3\)If $M$ is a linear operator from $\mathbb{R}^p$ to $\mathbb{R}^{p \times p}$, we define $\|M\|_{\text{op}} = \sup_{\|u\|_2 = 1} \|Mu\|_2$.
by Gaussian concentration. Hence for all practical purposes, a draw $U$ from the \( \theta \)-marginal of \( \tilde{\Pi}^{\infty}(\cdot | z) \) is such that \([U]_{\delta_{1}} \) and \([U]_{\delta_{2}} \) are independent, \([U]_{\delta_{1}} \sim \mathcal{N}(\hat{\theta}_{\star}, \mathcal{I}_{\gamma}^{-1})\), and \([U]_{\delta_{2}} \sim \mathcal{N}(0, \gamma((I_{p} - \gamma S)_{\delta_{2}})^{-1})\). Hence \( U_{j} = O_{p}(\sqrt{\gamma}) \) for all \( j \) such that \( \delta_{\star,j} = 0 \).

**Theorem 8.** Assume $H_{\bar{0}}$, $H_{\bar{\beta}}$ and (14) for some absolute constant $M > 2$. Fix $\gamma > 0$ such that $4 \gamma \lambda_{\max}(\bar{S}) \leq 1$, and $\rho > 0$. Suppose also that for some $\Lambda > 0$,

$$
\frac{2 \rho e^{L}}{p^{m}} \sqrt{\frac{2 \pi}{C_{\Lambda}(s)}} \leq 1, \quad \text{and} \quad (M - 1)e_{C}(s)^{1/2} - s_{\star}^{1/2} \geq 2.
$$

Then there exists absolute constant $A_{0}, C$, such that for all $p \geq A_{0}$, for all $m \geq 4$, and for all $z \in \mathcal{E}_{\bar{\beta}, \Lambda}$ such that $\tilde{\Pi}_{\gamma}(\cdot | z)$ is well-defined we have

$$
\tilde{\Pi}_{\gamma}(\{\delta_{\star}\} \times \mathcal{B}_{m,M}^{(\delta_{\star})}| z) \geq \tilde{\Pi}_{\gamma}(\tilde{\mathcal{B}}_{m,M}| z) - \left( \frac{4 \rho e^{L}}{p^{m}} \sqrt{\frac{2 \pi}{C_{\Lambda}(s)}} \right) e^{2(M+2)\rho s^{1/2}e} e^{(M+1)\rho C_{\Lambda}(s)^{1/2}} e^{3\gamma \bar{s}^{2}M / 2} e^{2\gamma \bar{r}(s)\bar{s}(M\epsilon)^{2}} e^{3\gamma \bar{r}^{2}\bar{s}} e^{ar{s}^{2}/2},
$$

Furthermore,

$$
\left\| \tilde{\Pi}_{\gamma}(\cdot | z) - \tilde{\Pi}_{\gamma}^{\infty}(\cdot | z) \right\|_{1 \nu} \leq \left( 1 - \tilde{\Pi}_{\gamma}(\{\delta_{\star}\} \times \mathcal{B}_{m,M}^{(\delta_{\star})}) \right) + 2 \left( 1 - \tilde{\Pi}_{\gamma}(\tilde{\mathcal{B}}_{m,M}| z) \right)
+ C_{t_{1}} \left( 1 + \rho s^{1/2}e + e^{3\gamma \bar{s}^{2}M / 2} \right) e^{C_{\rho s^{1/2}e} + C_{\rho s^{1/2}e} + 3\gamma \bar{s}^{2}M,}
$$

where

$$
t_{1} \overset{\text{def}}{=} e^{\epsilon^{2} + 3\gamma \bar{r}^{2}\bar{s} + 2\gamma \bar{r}(s)^{2}(M\epsilon)^{2}} - 1 + \frac{1}{p^{m}}.
$$

**Proof.** See Section B of the Supplement. \( \square \)

Theorem 5 and 8 can be combined to derive some simple conditions under which $\tilde{\Pi}_{\gamma}$ put most of its probability mass on subsets of the form \( \{\delta_{\star}\} \times \mathcal{B}_{m,M}^{(\delta_{\star})} \), and satisfies a Bernstein-von Mises approximation when $z \in \mathcal{E}_{\bar{\beta}, \Lambda}$. We refer to Section 2.2 for a linear regression example. Controlling the probability $\mathbb{P}_{\star}(Z \in \mathcal{E}_{\bar{\beta}, \Lambda})$ boils down to controlling the log-likelihood ratios of the model. This is easily done for linear regression models (see Section 2.2). The recent work of Sur et al. (2017) provides some tools that can be used to deal with logistic regression models, however this remains to be explored.

The Bernstein-von Mises approximation implies that for $\gamma$ small, posterior confidence intervals on $\theta$ obtained from $\tilde{\Pi}_{\gamma}$ are approximately equivalent to their corresponding frequentist counterparts knowing $\delta_{\star}$ (oracle confidence interval). In that sense the theorem can be used to show as $p$ grows that the Bayesian inference and the oracle frequentist inference agree.
2.2. Application to high-dimensional linear regression models. We illustrate the results with the linear regression model. Suppose that we have $n$ subjects, and on subject $i$ we observe $(Z_i, x_i) \in \mathbb{R} \times \mathbb{R}^p$, where $x_i$ is non-random, and the following holds.

**H6.** For $i = 1, \ldots, n$, $Z_i \sim N(\langle \theta_*, x_i \rangle, \sigma^2)$ are independent random variables, for a parameter $\theta_* \in \mathbb{R}^p \setminus \{0\}$, and a known absolute constant $\sigma^2 > 0$.

Let $X \in \mathbb{R}^{n \times p}$ be such that the $i$-th row of $X$ is $x_i'$. We write $X_j \in \mathbb{R}^n$ to denote the $j$-th column of $X$. Throughout the paper, we will automatically assume that the matrix $X$ is normalized such that $\|X_j\|_2^2 = n$, $j = 1, \ldots, p$.

For $s \geq 1$, we define

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

and $\bar{v}(s) \overset{\text{def}}{=} \sup \left\{ \frac{u'(X'X)u}{n\|u\|_2^2}, \ u \neq 0, \ \|u\|_0 \leq s \right\}$.

We also introduce

$$\bar{v} = \inf \left\{ \frac{u'(X'X)u}{n\|u\|_2^2}, \ u \neq 0, \ \|u\|_0 \leq \|u_{\delta_c^s}\|_1 \leq 7\|u_{\delta_c^s}\|_1 \right\}.$$

Different behavior can be obtained from $\bar{v}$ depending on the choice of the hyperparameter $\rho$. To keep the discussion short, we focus on the small-rho regime, and we choose $\rho$ and $\bar{\rho}$ as

$$\rho = \frac{4}{\sigma} \sqrt{\frac{n}{\log(p)}}, \quad \text{and} \quad \bar{\rho} = \frac{4}{\sigma} \sqrt{\frac{n}{\log(p)}}. \quad (21)$$

We make the following assumption.

**H7.** As $p \to \infty$, $s_* = o(\log(p))$ and

$$\frac{1}{v} + \frac{1}{\bar{v}(s)} + \bar{v}(\bar{s}) = O(1), \quad (22)$$

where $\bar{s} \overset{\text{def}}{=} s_* + \frac{2}{n}(m_0 + 1 + 2s_*)$, for some absolute constant $m_0 \geq 1$.

**Remark 9.** Assumption H7 can be viewed as a minimal sample size requirement. It imposes that we increase the sample size $n$, as $p$ grows, so as to guarantee that $v$, $\bar{v}(s)$ remain bounded away from 0, and $\bar{v}(\bar{s})$ remains bounded from above. For instance if $X$ is a realization of a random matrix with iid entries it is known that H7 holds with high probability if the sample $n$ grows as $O(\bar{s}\log(p))$ (see e.g. Rudelson and Zhou (2013) and the references therein).
In the next result $\epsilon$ as defined in (11) takes the form

$$\epsilon = \frac{24\sigma}{\mu(s)} \sqrt{\frac{(s_\star + \bar{s}) \log(p)}{n}}.$$ 

And the limiting distribution $\tilde{\Pi}_\gamma^\infty(\cdot|z)$ in the Bernstein-von Mises approximation is such that if $(\delta, U) \sim \tilde{\Pi}_\gamma^\infty(\cdot|z)$, then $\delta = \delta_\star$, $[U]_{\delta_\star} \sim \mathcal{N}((X'_{\delta_\star} X_{\delta_\star})^{-1}X_{\delta_\star} z, \sigma^2(X'_{\delta_\star} X_{\delta_\star})^{-1})$, and $[U]_{\delta_\star} \sim \mathcal{N}(0, \gamma((I_p - \frac{2}{\sigma^2} (X'X)_{\delta_\star}))^{-1})$, with $[U]_{\delta_\star}, [U]_{\delta_\star}$ independent. We deduce the following corollary.

**Corollary 10.** Assume $H_3$, $H_6$, $H_7$, and choose $\rho$ and $\bar{\rho}$ as in (21). Suppose also that we choose $\gamma > 0$ such that $8\sigma^2 \gamma \log(p) \lambda_{\max}(X'X) \leq 1$, and as $p \to \infty$,

$$\gamma n \log(p) = O(1), \quad \text{and} \quad \gamma^2 \left( n \text{Tr}(X'X) + \|X'X\|_F^2 \right) = o(\log(p)). \quad (23)$$

Furthermore, suppose that as $p \to \infty$, $\log(p)/n \to 0$, and

$$\lim_{p \to \infty} \left\{ \min_{j: \delta_\star,j = \delta_\star} |\theta_\star,j| \right\} \sqrt{\frac{n}{(s_\star + \bar{s}) \log(p)}} = +\infty. \quad (24)$$

Then setting $\Lambda = 3 \log(n \wedge p)$, the following holds. For all $m > 1$, all $M > \max(2, \sqrt{\frac{u+2}{3}})$, we have

$$\mathbb{E}_* \left[ \tilde{\Pi}_\gamma \left( \{\delta_\star\} \times B_{m,M}^{(\delta_\star)} | Z \right) \left| Z \in \mathcal{E}_{\bar{\rho},\Lambda} \right. \right] \geq 1 - \frac{1}{p^m_0} - \frac{1}{p^{m-1}} - \frac{1}{p^M s} - \frac{1}{p^{u-3}},$$

and

$$\mathbb{P}_* (Z \notin \mathcal{E}_{\bar{\rho},\Lambda}) \leq \frac{2}{p} + \frac{A_1}{(n \wedge p)^T},$$

for all $p$ large enough, where $A_1$ is some absolute constant. If in addition to the assumptions above, $\gamma s_\star n \log(p) = o(1)$, and $\gamma^2 \left( n \text{Tr}(X'X) + \|X'X\|_F^2 \right) = o(1)$, as $p \to \infty$, then

$$\lim_{p \to \infty} \mathbb{E}_* \left[ \left\| \tilde{\Pi}_\gamma(\cdot|Z) - \tilde{\Pi}_\gamma^\infty(\cdot|Z) \right\|_{1v} \right] = 0.$$ 

**Proof.** See Section C of the Supplement. \hfill \Box

The assumption (24) is a high signal-to-noise ratio assumption. It implies that for any $M > 2$, (14) holds for all $p$ large enough. Assumption (23) can always to satisfied by taking $\gamma$ sufficiently small, or by using a design matrix $X$ with few leading eigenvalues.
3. Markov Chain Monte Carlo Computation

In this section we develop and analyze a MCMC algorithm to sample from $\bar{\Pi}_\gamma$. We shall focus on the $\theta$-marginal of $\bar{\Pi}_\gamma$ given by

$$\bar{\Pi}_\gamma(du|z) \propto \sum_{\delta \in \Delta} \omega_\theta(2\pi \gamma)^{-\frac{d_\|\|}{2}} \left(\rho^{\|\delta\|_0}\right) e^{-h_\gamma(\delta, u; z)} du.$$  

We will abuse the notation and continue to write $\bar{\Pi}_\gamma$ to denote this marginal. The set whose probability we seek will typically make it clear whether we are referring to the joint distribution or its marginals.

3.1. A MCMC sampler for $\bar{\Pi}_\gamma$. We start first with a description of the MCMC sampler. We use a data-augmentation approach where we sample the joint variable $(\delta, \theta)$, and then discard $\delta$. To sample $(\delta, \theta)$ we use a Metropolis-Hasting-within-Gibbs sampler, where we update $\delta$ given $\theta$, then we update the selected component $[\theta]_\delta$ given $(\delta, [\theta]_{\delta'})$, and finally update $[\theta]_{\delta'}$ given $(\delta, [\theta]_{\delta})$. We refer the reader to Tierney (1994); Robert and Casella (2004) for an introduction to basic MCMC algorithms.

To develop the details, we need to introduce some notations. For $\gamma > 0$, $\delta \in \Delta$, $\theta \in \mathbb{R}^p$, we define the proximal map

$$\text{Prox}_\gamma(\delta, \theta) \overset{\text{def}}{=} \text{Argmin}_{v \in \mathbb{R}^p} \left[ \rho\|v\|_1 + \frac{1}{2\gamma}\|v - \theta\|_2^2 \right] = \delta \cdot s_\gamma(\theta),$$

where $s_\gamma(\theta) = (s_\gamma(\theta_1), \ldots, s_\gamma(\theta_p)) \in \mathbb{R}^p$ is the soft-thresholding operation applied to $\theta$ with

$$s_\gamma(x) \overset{\text{def}}{=} \text{sign}(x) (|x| - \gamma \rho)_+, \quad x \in \mathbb{R},$$

where $\text{sign}(a)$ is the sign of $a$, and $a_+ = \text{max}(a, 0)$. With this definition, $h_\gamma$ in (1) can be rewritten as

$$h_\gamma(\delta, \theta; z) = -\ell(\theta; z) - \langle \nabla \ell(\theta; z), J_\gamma(\delta, \theta) - \theta \rangle + \rho\|J_\gamma(\delta, \theta)\|_1$$

$$+ \frac{1}{2\gamma}\|J_\gamma(\delta, \theta) - \theta\|_2^2$$

where

$$J_\gamma(\delta, \theta) \overset{\text{def}}{=} \text{Prox}_\gamma(\delta, \theta + \gamma \nabla \ell(\theta; z)) = \delta \cdot s_\gamma(\theta + \gamma \nabla \ell(\theta; z)).$$

(27)
This alternative expression shows that the function \( h_\gamma(\delta, \theta; z) \) is easy to evaluate. Furthermore, plugging (26) in (3) shows that given \( \theta \), the components of \( \delta \) are conditionally independent Bernoulli random variables:

\[
\tilde{\Pi}_\gamma(\delta|\theta, z) = \prod_{j=1}^{P} [p_{\gamma,j}(\theta)]^{\delta_j} [1 - p_{\gamma,j}(\theta)]^{1-\delta_j},
\]

where \( p_{\gamma,j}(\theta) \equiv \frac{q}{1-q} \frac{\sqrt{2\pi}\gamma e^{r_{\gamma,j}}}{1 + \frac{q}{1-q} \frac{\sqrt{2\pi}\gamma e^{r_{\gamma,j}}}} \), \( j = 1, \ldots, p \). (28)

where \( r_{\gamma,j} \equiv r_{\gamma}(\theta_j + \gamma \nabla_j \ell(\theta; z)) \), where \( \nabla_j \ell(\theta; z) \) denotes the \( j \)-th partial derivative of \( \ell(\theta; z) \) with respect to \( \theta \), and evaluated at \( \theta \), and for \( x \in \mathbb{R} \),

\[
\quad \quad \quad \quad \quad \quad r_{\gamma}(x) \equiv -\frac{1}{2\gamma} s_{\gamma}(x)^2 + \frac{1}{\gamma} x s_{\gamma}(x) - \rho|s_{\gamma}(x)| = \begin{cases} \frac{1}{2\gamma}(|x| - \gamma \rho)^2 & \text{if } |x| > \gamma \rho \\ 0 & \text{otherwise.} \end{cases} \tag{29}
\]

It follows from the above that drawing samples from the conditional distribution of \( \delta \) given \( \theta, z \) is easily achieved.

Now consider the conditional distribution \( \tilde{\Pi}_\gamma(\theta|\delta, z) \propto e^{-h_\gamma(\delta, \theta; z)} \). Given \( \delta \), we partition \( \theta \) into \( \theta = ([\theta]_\delta, [\theta]_{\delta^c}) \), where \( [\theta]_\delta \) groups the components of \( \theta \) for which \( \delta_j = 1 \), and \( [\theta]_{\delta^c} \) groups the remaining components. We propose a Metropolis-within-Gibbs MCMC scheme whereby we first update \([\theta]_\delta \) using a Random Walk Metropolis scheme while keeping \([\theta]_{\delta^c} \) fixed, and then we update \([\theta]_{\delta^c} \) using an Independence Metropolis-Hastings scheme while keeping \([\theta]_\delta \) fixed. Again we refer the reader to Tierney (1994) for an introduction to these basic MCMC algorithms. To give more details, it is enough to consider the case where \( 0 < ||\delta||_0 < p \). We update the component \([\theta]_\delta \) using a Random Walk Metropolis with a Gaussian proposal \( \mathcal{N}(0, \tau_{\delta}^2 I_{||\delta||_1}) \) for some scale parameter \( \tau_{\delta}^2 > 0 \) (we give more detail on the choice of \( \tau_{\delta}^2 \) below), while keeping \( \delta \) and \([\theta]_{\delta^c} \) fixed. To update the component \([\theta]_{\delta^c} \), we build an approximation \( \tilde{h}_\gamma \) of \( h_\gamma \) as follows. First, in view of (27), we propose to approximate \( J_\gamma(\delta, \theta) \) by

\[
\tilde{J}_\gamma(\delta, \theta) \equiv \delta \cdot s_{\gamma}(\theta + \gamma \nabla \ell(\theta_{\delta}; z)),
\]

and we note that \( \tilde{J}_\gamma(\delta, \theta) \) does not actually depend on \([\theta]_{\delta^c} \). Using Taylor expansion we also approximate \( \ell(\theta; z) \) and \( \nabla \ell(\theta; z) \) by \( \tilde{\ell}(\theta; z) \) and \( \tilde{\nabla} \ell(\theta; z) \) respectively, where

\[
\tilde{\ell}(\theta; z) \equiv \ell(\theta_{\delta}; z) + \langle \nabla \ell(\theta_{\delta}; z), \theta - \theta_{\delta} \rangle + \frac{1}{2}(\theta - \theta_{\delta})' \nabla^{(2)} \ell(\theta_{\delta}; z)(\theta - \theta_{\delta}),
\]

and

\[
\tilde{\nabla} \ell(\theta; z) \equiv \nabla \ell(\theta_{\delta}; z) + \nabla^{(2)} \ell(\theta_{\delta}; z)(\theta - \theta_{\delta}).
\]

We then propose to approximate \( h_\gamma(\delta, \theta; z) \) by replacing \( J_\gamma(\delta, \theta) \) by \( \tilde{J}_\gamma(\delta, \theta) \), \( \ell(\theta; z) \) by \( \tilde{\ell}(\theta; z) \), and \( \nabla \ell(\theta; z) \) by \( \tilde{\nabla} \ell(\theta; z) \) in the expression of \( h_\gamma \) given in (26). This leads
to
\[
\tilde{h}_\gamma(\delta, \theta; z) \overset{\text{def}}{=} -\tilde{\ell}(\theta; z) - \langle \tilde{\nabla} \ell(\theta; z), \tilde{J}_\gamma(\delta, \theta) - \theta \rangle \\
+ \rho \| \tilde{J}_\gamma(\delta, \theta) \|_1 + \frac{1}{2\gamma} \| \tilde{J}_\gamma(\delta, \theta) - \theta \|_2^2.
\] (30)

**Remark 11.** It will be important to have in mind that in linear regression models, \(\theta \mapsto \ell(\theta; z)\) is quadratic so that \(\tilde{\ell}(\theta; z) = \ell(\theta; z)\), and \(\tilde{\nabla} \ell(\theta; z) = \nabla \ell(\theta; z)\). Hence in that case the approximation \(\tilde{h}\) involves only replacing \(J\) by \(\tilde{J}\).

\[
\Sigma_{\delta, \theta} \overset{\text{def}}{=} \left( I_p + \gamma \nabla^2 \ell(\theta_\delta; z) \right)_{\delta c}^{-1},
\]

and
\[
m_{\delta, \theta} \overset{\text{def}}{=} \Sigma_{\delta, \theta} \left[ (I_p + \gamma \nabla^2 \ell(\theta_\delta; z)) (\tilde{J}_\gamma(\delta, \theta) - \theta_\delta) \right]_{\delta c}.
\] (31)

We set
\[
\Sigma_{\delta, \theta} \text{ and } m_{\delta, \theta} \text{ depend on } \theta \text{ only through } [\theta]_{\delta c}. \text{ We note also that under } H_1 \text{ the matrix } \Sigma_{\delta, \theta} \text{ is symmetric positive definite whenever } \gamma \lambda_{\text{max}}(\bar{S}) < 1. \text{ With some easy algebra it can be shown that } \tilde{h}_\gamma(\delta, \theta; z) \text{ can be written as }
\]
\[
\tilde{h}_\gamma(\delta, \theta; z) = \frac{1}{2\gamma} (\theta]_{\delta c} - m_{\delta, \theta})^t \Sigma_{\delta, \theta}^{-1} (\theta]_{\delta c} - m_{\delta, \theta}) + \text{const},
\]
where the term \text{const} does not depend on \([\theta]_{\delta c}\) (but does depend on \([\theta]_{\delta}\)). Hence, given \(\delta\) and \([\theta]_{\delta}\), we update \([\theta]_{\delta c}\) using an Independence Metropolis-Hastings algorithm with proposal \(N(m_{\delta, \theta}, \gamma \Sigma_{\delta, \theta})\).

Given \(\delta \in \Delta\), let us call \(\hat{K}_\delta\) the transition kernel on \(\mathbb{R}^p\) of the combined \(\theta\)-move that we just described. Let us call \(\hat{K}_\delta\) the kernel obtained by reversing the order of the updates (we update \([\theta]_{\delta c}\) first using the independence sampler described above, followed by a random walk Metropolis update of \([\theta]_{\delta}\)). For the purpose of having a reversible kernel we introduce
\[
K_\delta(\theta, \cdot) \overset{\text{def}}{=} \frac{1}{2} \hat{K}_\delta(\theta, \cdot) + \frac{1}{2} \hat{K}_\delta(\cdot, \theta).
\]
The proposed MCMC algorithm to sample from \(\hat{\Pi}_\gamma\) is as follows.

**Algorithm 1.** For some initial distribution \(\nu_0\) on \(\mathbb{R}^p\), draw \(u_0 \sim \nu_0\). Given \(u_0, \ldots, u_n\) for some \(n \geq 0\), draw independently \(D_{n+1} \sim \text{Ber}(0.5)\).

(1) If \(D_{n+1} = 0\), set \(u_{n+1} = u_n\).

(2) If \(D_{n+1} = 1\),
   (a) Draw \(\tilde{\delta} \sim \hat{\Pi}_\gamma(\cdot | u_n, z)\) as given in (28), and
   (b) draw \(u_{n+1} \sim K_\delta(u_n, \cdot)\).
Remark 12. The use of the kernel $K_\delta$ instead of $\hat{K}_\delta$ (or $\tilde{K}_\delta$) does not increase the computational cost, and insures reversibility, which is needed in our theory. The introduction of the indicator variable $D_n$ implies that half of the time the chain does not move: we have a lazy Markov chain, which is also needed in our theory. These tricks are not used in practice, and for the numerical illustrations presented below we only implemented the kernel $\hat{K}_\delta$.

The indicator variables $\delta$ discarded in Algorithm 1 are important in practice for the variable selection problem, and are usually collected along the iterations. Here we focus the analysis on the continuous variables $u_n \in \mathbb{R}^p$. Obviously we do not lose anything, since given $u_n$ exact sampling of $\delta$ is possible as discussed above. In other words the mixing of the joint process $\{(\delta_n, u_n), n \geq 0\}$ is driven by the mixing of the marginal $\{u_n, n \geq 0\}$. □

3.1.1. Initialization. The choice of the initial distribution $\nu_0$ plays a crucial role for fast mixing. Given $z \in \mathbb{R}^n$, and a model $\delta \in \Delta$, let $\hat{\theta}_\delta = (X_\delta X_\delta')^{-1} X_\delta' z$ denote the ordinary least squares estimate based on the selected variables of $\delta$. Let us call $\nu(\delta)(\cdot|z)$ the Gaussian distribution on $\mathbb{R}^p$ such that if $(U_1, \ldots, U_p) \sim \nu(\delta)(\cdot|z)$, then $[U]_\delta \sim N(\hat{\theta}_\delta, \sigma^2(X_\delta X_\delta')^{-1})$, and (independently) $U_j \overset{i.i.d.}{\sim} N(0, \gamma)$ for all $j$ such that $\delta_j = 0$.

We propose to take the initial distribution $\nu_0$ as $\nu(\delta^{(0)})(\cdot|z)$ for some initial estimate $\delta^{(0)}$ of $\delta^\star$. Perhaps the most natural choice of $\delta^{(0)}$ is the lasso estimate (Tibshirani (1996); Bickel et al. (2009)). In a strong signal-to-noise ratio setting the lasso is known to recover $\delta^\star$ with high probability (Meinshausen and Yu (2009)). However in practice lasso estimates can perform poorly. So it is important to understand the mixing of the MCMC sampler when $\delta^{(0)}$ is close but not exactly equal to $\delta^\star$.

3.1.2. Computational cost. The computational cost of Algorithm 1 is dominated by the cost of sampling from the Gaussian distribution $N(m_\delta, \Sigma_\delta)$, which itself is dominated by the Cholesky decomposition of $\Sigma_\delta$. Hence each iteration of Algorithm 1 in general has a cost that scales with $p$ as $O(p^3)$. However in some cases, a faster implementation is possible along the lines of an algorithm proposed in Bhattacharya et al. (2016). Suppose that the Hessian matrix $\nabla^2 \ell(\theta; z)$ can be written as

$$\nabla^2 \ell(\theta; z) = -X'W_\theta X,$$  \((32)\)

where $X \in \mathbb{R}^{n \times p}$, and $W_\theta \in \mathbb{R}^{n \times n}$ is a non-singular diagonal matrix. This is the case for instance for linear or logistic regression models. Then $\Sigma_\delta = (I_p - ||\delta||_0 - \gamma X_\delta' W_\delta X_\delta)^{-1}$.
where \( X_{\delta^c} \in \mathbb{R}^{n \times (p-\lVert \delta \rVert_0)} \) is the sub-matrix of \( X \) obtained by selecting the columns for which \( \delta_j = 0 \). By the Woodbury formula we then have
\[
\Sigma_{\delta, \theta} = I_{p-\lVert \delta \rVert_0} + \gamma X_{\delta^c}' \left( W_\theta^{-1} - \gamma X_{\delta^c} X_{\delta^c}' \right)^{-1} X_{\delta^c}.
\]
Therefore if \( C_\theta' C_\theta = W_\theta^{-1} - \gamma X_{\delta^c} X_{\delta^c}' \) is the Cholesky decomposition of the \( \mathbb{R}^{n \times n} \) matrix \( W_\theta^{-1} - \gamma X_{\delta^c} X_{\delta^c}' \), we can sample from \( N(0, \gamma \Sigma_{\delta, \theta}) \) by drawing \( Z \sim N(0, I_{p-\lVert \delta \rVert_0}), U \sim N(0, I_n) \), independently and returning
\[
\sqrt{\gamma} Z + \gamma X_{\delta^c}' C_\theta^{-1} U.
\]
It is also easy to see that the Cholesky factor \( C_\theta \) can also be exploited to compute \( m_{\delta, \theta} \). The per-iteration computational cost of this approach is \( O(n^3 + p^2 n) \). Hence in the particular case where (32) holds, the per-iteration computational cost of the algorithm is \( O(p^2 \min(n, p)) \), which matches other state of the art algorithms for high-dimensional regression (Bhattacharya et al. (2016)).

3.2. Mixing time of Markov chains. Our objective in this section is to provide some qualitative bounds on the mixing time of Algorithm 1. Particularly, we wish to understand how this mixing time depends on the dimension \( p \). We follow the conductance approach using the framework of Lovász and Simonovits (1993). However this theory cannot be directly applied since the target distribution of interest is a mixture of log-concave densities, and hence is not log-concave. Our main contribution is the idea that one can invoke the contraction properties of \( \tilde{\Pi}_\gamma \) to essentially reduce the mixing of \( \tilde{\Pi}_\gamma \) to the mixing of \( K_{\delta^c} \) – the Markov kernel that samples from the dominant component of \( \tilde{\Pi}_\gamma \). This latter problem can then be handled by the standard theory of Lovász and Simonovits (1993).

We start with a general overview of the technique using some generic notation; the specific application to \( \tilde{\Pi}_\gamma \) is presented in Section 3.3. Let \( \pi \) be a probability measure on \( \mathbb{R}^p \) that is absolutely continuous with respect to the \( \mathbb{R}^p \)-Lebesgue measure \( d\theta \) such that
\[
\pi(d\theta) \propto e^{-h(\theta)} d\theta, \quad \theta \in \mathbb{R}^p,
\]
for a measurable function \( h : \mathbb{R}^p \to [0, \infty) \). We will abuse notation and write \( \pi \) to denote both \( \pi \) and its density. Let \( P \) be a Markov kernel on \( \mathbb{R}^p \). For any integer \( n \geq 1 \), \( P^n \) denotes the Markov kernel defined recursively as \( P^1 = P \), and
\[
P^n(x, \cdot) \overset{\text{def}}{=} \int P^{n-1}(x, dz) P(z, \cdot).
\]
For a probability measure \( \mu \), the product \( \mu P \) is the probability measure defined as
\[
\mu P(\cdot) \overset{\text{def}}{=} \int \mu(dz) P(z, \cdot).
\]
We say that $P$ is reversible with respect to $\pi$ if for all measurable sets $A, B \subseteq \mathbb{R}^p$:

$$\int_A \pi(d\theta)P(\theta, B) = \int_B \pi(d\theta)P(\theta, A).$$

Reversibility of $P$ with respect to $\pi$ implies that $P$ has invariant distribution $\pi$. We say that a Markov kernel $P$ is lazy if $P(\theta, \{\theta\}) \geq 1/2$ for all $\theta \in \mathbb{R}^p$. For $\zeta \in (0, 1/2)$, the $\zeta$-conductance of the Markov kernel $P$ as introduced by [Lovász and Simonovits (1990)] is defined as

$$\Phi_{\zeta}(P) \overset{\text{def}}{=} \inf \left\{ \frac{\int_A \pi(d\theta)P(\theta, A^c)}{\min(\pi(A) - \zeta, \pi(A^c) - \zeta)} : A \text{ meas., } \zeta < \pi(A) < 1 - \zeta \right\}.$$

The case $\zeta = 0$ corresponds to the usual conductance. The conductance measures how rapidly a Markov chain moves around the space if started from its stationary distribution. In practice most MCMC algorithms are started from some initial distribution $\nu_0$ that is not the stationary distribution. In high-dimensional problems – due to the curse of dimensionality and the concentration of measure phenomenon – the choice of the initial distribution becomes crucial. A fundamental result by [Sinclair and Jerrum (1989)] relates the mixing time of the Markov chain to the conductance of $P$ and the properties of the initial distribution $\nu_0$. More details can also be found in [Dyer et al. (1991); Lovász and Vempala (2007); Belloni and Chernozhukov (2009)] and the references therein. Here we will use the generalization provided by Corollary 1.5-(2) of [Lovász and Simonovits (1993)], which can be stated as follows.

**Theorem 13.** Suppose that $P$ is lazy, and has invariant distribution $\pi$, and fix $\zeta \in (0, 1/2)$. For any probability measure $\nu_0$, and any integer $K \geq 1$, we have

$$\|\nu_0 P^K - \pi\|_{tv} \leq H_{\zeta} \left( 1 + \frac{1}{\zeta} e^{-K \frac{\Phi_{\zeta}(P)}{2}} \right),$$

where

$$H_{\zeta} \overset{\text{def}}{=} \sup_{A: \pi(A) \leq \zeta} |\nu_0(A) - \pi(A)|.$$

Using this result boils down to lower bounding the $\zeta$-conductance $\Phi_{\zeta}(P)$, and upper bounding $H_{\zeta}$. We follow the approach of [Lovász and Vempala (2007)] which consists in studying the restriction of $P$ to some well-chosen subset of $\mathbb{R}^p$. More precisely, if $\Theta \subseteq \mathbb{R}^p$ is a non-empty measurable subset such that $\pi(\Theta) > 0$, the $\Theta$-conductance of $P$ is

$$\Phi_{\Theta}(P) \overset{\text{def}}{=} \inf \left\{ \frac{\int_B \pi(d\theta)P(\theta, B^c \cap \Theta)}{\min(\pi(B), \pi(B^c \cap \Theta))} : B \subseteq \Theta, B \text{ meas., } 0 < \pi(B) < \pi(\Theta) \right\}.$$

We note that if $\pi(\Theta) \geq 1 - \zeta$, then $\Phi_{\zeta}(P) \geq \Phi_{\Theta}(P)$. To see this, let $A$ measurable be such that $\zeta < \pi(A) < 1 - \zeta$. Then $\pi(A \cap \Theta) = \pi(A) - \pi(A \cap \Theta^c) \geq \pi(A) - \zeta > 0$. 


Similarly, \( \pi(A^c \cap \Theta) \geq \pi(A^c) - \zeta = 1 - \zeta - \pi(A) > 0 \). We conclude that
\[
\frac{\int_A \pi(d\theta)P(\theta, A^c)}{\min(\pi(A) - \zeta, \pi(A^c) - \zeta)} \geq \frac{\int_{A^c} \pi(d\theta)P(\theta, A^c \cap \Theta)}{\min(\pi(A \cap \Theta), \pi(A^c \cap \Theta))} \geq \Phi(\Theta).
\]
Hence we can reduce the problem to lower bounding \( \Phi(\Theta) \) for a well-chosen subset \( \Theta \). The next result builds on Theorem 2 of [Belloni and Chernozhukov (2009)] and provides an approach to lower-bound \( \Phi(\Theta) \). The result itself is of some independent interest, since it can be applied more widely. The set up is as follows. With \( \pi \) and \( P \) as above, suppose that there exists a sub-Markov kernel \( \{q(x, \cdot), x \in \mathbb{R}^p\} \) (meaning that \( (x, y) \mapsto q(x, y) \) is measurable, and \( \int_{\mathbb{R}^p} q(x, y)dy \leq 1 \) for all \( x \in \mathbb{R}^p \)) such that
\[
P(\theta, du) \geq q(\theta, u)du, \quad \theta \in \mathbb{R}^p.
\]
(34)

Given \( u, v \in \mathbb{R}^p \), set
\[q_{u,v}(\theta) \coloneqq \min(q(u, \theta), q(v, \theta)), \quad \theta \in \mathbb{R}^p.\]
Let \( \|\cdot\| \) denote some arbitrary pseudo-norm on \( \mathbb{R}^p \) such that \( \|\theta\| \leq \|\theta\|_2 \) for all \( \theta \).

**Theorem 14.** Suppose that \( P \) is a transition kernel that is reversible with respect to \( \pi \) as given in (33), such that (34) holds. Suppose also that the following holds.

1. There exists a nonempty convex set \( \Theta \subset \mathbb{R}^p \) with finite diameter \( \text{diam}(\Theta) \coloneqq \max_{u, v \in \Theta} \|u - v\|_2 \), such that \( h \) is convex on \( \Theta \).
2. There exist \( r > 0, \alpha > 0 \), such that for all \( u, v \in \Theta \) that satisfy \( \|u - v\| \leq r \), we have \( \int_{\Theta} q_{u,v}(\theta)d\theta \geq \alpha \).

Then
\[
\Phi(\Theta) \geq \frac{\alpha}{4} \min \left[ 1, \frac{2r}{\text{diam}(\Theta)} \right].
\]
**Proof.** See Section 4.1.

### 3.3. Application to the kernel \( \tilde{P}_\gamma \) for linear regression models.

We analyze the mixing of Algorithm 1. For simplicity we focus on linear regression models. The transition kernel of the \( \mathbb{R}^p \)-valued Markov chain \( \{u_n, n \geq 0\} \) defined by Algorithm 1 is
\[
\tilde{P}_\gamma(u, dv) \coloneqq \frac{1}{2} \delta_u(dv) + \frac{1}{2} \sum_{\omega \in \Delta} \tilde{P}_\gamma(\omega|u, z)K_\omega(u, dv).
\]
(35)

The incoherence of the design matrix \( X \) plays a role. We define
\[
C(X) \coloneqq \sup_{j: \delta_{i,j} = 1} \sup_{u \in \mathbb{R}^{p-s_+}: \|u\|_2 = 1} \frac{1}{\sqrt{n}} |\langle X_j, X_{\delta_i}u \rangle| = \sup_{j: \delta_{i,j} = 1} \frac{1}{\sqrt{n}} \|X_{\delta_i}'X_j\|_2,
\]
(36)
where \( X_{\delta_i} \in \mathbb{R}^{n \times (p-s_+)} \) is the submatrix of \( X \) corresponding to columns \( j \) for which \( \delta_{i,j} = 0 \). For a random matrix with iid Gaussian entries, \( C(X) \approx \sqrt{p} \). With the same notations as in Section 2.2, we have the following result.
Theorem 15. Assume $H_3$, $H_6$, $H_7$, choose $\rho, \bar{\rho}$ as in (21), and choose $\gamma = \frac{\gamma_0}{n \log(p)}$ for some absolute constant $\gamma_0 > 0$ such that (23) and (24) hold. Furthermore, as $p \to \infty$, suppose that

$$s_3^{3/2}C(X)\sqrt{n \log(p)} = o(1).$$

Suppose that we initialize Algorithm 1 as described in Section 3.1.1 with $\nu_0 = \nu(\delta(0)(|z)$ for $\delta(0) \supseteq \delta_*$ such that $FP \overset{\text{def}}{=} \|\delta(0)\|_0 - s_* = O(1)$, as $p \to \infty$. Fix $\zeta_0 \in (0, 1/2)$. Then we can find absolute constants $C_0, C_1, C_2, C_3$ (that depend on $\zeta_0$) such that if we scale the step-size of the Random Walk Metropolis update of Algorithm 1 as $\tau_\delta = \frac{C_0}{\|\delta\|_0^{1/2}\sqrt{n \log(p)}}$, the following holds. For all $p \geq C_1$, and all integer $K$ such that

$$K \geq C_2 (1 + FP) p \exp\left(C_3 s_*^2 \left[1 + \left(\frac{C(X)}{s^{1/2} \log(p) \sqrt{p n}}\right)^2\right]\right),$$

we have

$$E_* \|\nu_0 P_K^\gamma - \Pi_\gamma(\cdot|Z)\|_{tv} \leq 3\zeta_0,$$

provided that the constants $u, m_0$ in $H_3$ and $H_7$ are taken large enough.

Proof. See Section 4.2. \hfill \square

The theorem suggests that the mixing of the algorithm has a complex dependence on the dimension $p$. However, if the coherence of the design matrix $X$ is small and $p$ is not too large compared to $n$, then the theorem suggests that the mixing time is essentially linear in $p$. We note also that the bound degrades for large values of the false-positive number $FP$. However, since we have assumed in the theorem that $FP = O(1)$, it is not possible to get a clear read of the dependence on $FP$.

The strong signal-to-noise ratio (24) plays a crucial role in the analysis, and our method of proof breaks down if this assumption does not hold. More flexible techniques such as the space decomposition approach (Madras and Randall (2002); Guan and Krone (2007); Woodard et al. (2009)), or perhaps the coupling approach of Mangoubi and Smith (2017) might be more successful in relaxing this assumption.

3.4. Numerical illustrations. We illustrate some of the conclusions above with the following simulation study. We consider a linear regression model with Gaussian noise $N(0, \sigma^2)$, where $\sigma^2$ is set to 1. We experiment with sample size $n = 200$, and dimension $p \in \{500, 1000, \cdots, 5000\}$. We fix the number of non-zero coefficients to $s_* = 10$, and $\delta_*$ is given by

$$\delta_* = (1, \cdots, 1, 0, \cdots, 0).$$
The non-zero coefficients of $\theta_*$ are uniformly drawn from $(-a - 1, -a) \cup (a, a + 1)$, where

$$a = \text{SNR} \sqrt{\frac{s_\star \log(p)}{n}}, \quad \text{where} \quad \text{SNR} \in \{0.5, 4\}.$$  

We expect $\text{SNR} = 4$ to bring us closer to satisfying (24). We consider two different design matrices. A standard design $X \in \mathbb{R}^{n \times p}$ with i.i.d. entries $N(0, 1)$. For such matrices $C(X)$ is of order $\sqrt{p}$. We generate a high coherence design matrix as follows. First we generate $X_0 \in \mathbb{R}^{n \times p}$ with i.i.d. standard Gaussian entries. We standardize $X_0$ to satisfy (20), and form its singular value decomposition $X_0 = \sum_{k=1}^{n} S_k U_k V_k'$. We then define

$$X = \sum_{k=1}^{K} S_k U_k V_k',$$

for some integer $K$. In the extreme case where $K = 1$ the coherence is of order $C(X) \approx \sqrt{np}$. We set $K = 30 + (p - 2000)/100$ (we use $K = 20$ for $p \leq 1000$).

We choose the prior as in H3, with $u = 5$, $\rho$ as in (21), and set $\gamma$ as

$$\gamma = \frac{\gamma_0 \sigma^2}{\lambda_{\max}(X'X)},$$

for a tuning parameter $\gamma_0 = 0.2$. We use the initial distribution $\nu_0 = \nu^{(\delta^{(0)})(\cdot | z)}$ described in Section 3.1.1 for some initial value $\delta^{(0)} \in \Delta$. Obviously, we ideally want to set $\delta^{(0)} = \delta_*$. We will allow for some errors in specifying $\delta^{(0)}$, and we consider three choices, (a) a good initialization where $\delta^{(0)}$ has no false-negative and 10% false-positive, (b) a poor initialization where $\delta^{(0)}$ has no false-positive but 20% false-negative, (c) a lasso initialization where $\delta^{(0)}$ is taken as the lasso estimate computed using MATLAB default cross-validation set up. The lasso initialization corresponds of course to the initialization one would typically use in practice.

Since the ideal value of the step-size $\tau$ in the Random Walk Metropolis step of Algorithm 1 is not known, we use an adaptive Random Walk Metropolis algorithm (Atchadé and Rosenthal (2005)) to adaptively select $\tau$ such that the acceptance probability of the Markov chain is approximately 30%.

To monitor the mixing, we compute the sensitivity and the precision at iteration $k$ as

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

We note that (24) is actually overly strong. What is needed is (14) for some absolute constant $M > 2$ that can be taken large.
And we empirically measure the mixing time of the algorithm as the first time $k$ where both $\text{SEN}_k$ and $\text{PREC}_k$ reach 1, truncated to $2 \times 10^4$ – that is we stop any run that has not mixed after $2 \times 10^4$ iterations. In the high signal-to-noise regime ($\text{SNR} = 4$) this definition makes sense, since in that case we know that with high probability most of the probability mass of $\hat{\Pi}_\gamma$ is concentrated on $\delta_*$. In the weak signal-to-noise ratio regime this definition is not appropriate since in this case the distribution $\hat{\Pi}_\gamma$ can have a non-negligible probability of omitting some of the non-zero coefficients. In this case we amend the definition and set the mixing time as the first time where $\text{SEN} \geq \alpha_{\text{SEN}}$, and $\text{PREC} = 1$, where $\alpha_{\text{SEN}}$ is set by running a long preliminary version of the algorithm.

For comparison we also show the results for a similar Metropolis-within-Gibbs algorithm to sample from the weak spike-and-slab posterior distribution $\bar{\Pi}_\gamma$ given in (3). This distribution differs from the posterior distribution analyzed by Narisetty and He (2014) only in the fact that we have used here a Laplace slab density, instead of the Gaussian density used by Narisetty and He (2014). One can sample similarly from $\bar{\Pi}_\gamma$ with the same strategy described above for $\hat{\Pi}_\gamma$, with the additional simplification that the conditional distribution of $[\theta]_{\delta^c}$ given $\delta, [\theta]_\delta$ has a Gaussian closed form. The resulting sampler is similar to the Gibbs sampler implemented in Narisetty and He (2014).

All the results presented are based 45 independent MCMC replications. Figure 1 presented in the introduction shows the behavior of the mixing time as function of the dimension $p$, with $\text{SNR} = 4$, under different initialization and design matrix coherence. The simulation results confirm the conclusion of Theorem 15 that the mixing time is roughly linear in $p$ when the algorithm is well initialized and the coherence of the design matrix is low. We also observe that the two algorithms (for $\hat{\Pi}_\gamma$ and $\bar{\Pi}_\gamma$) behave similarly.

We also explore the behavior of the algorithms under the lasso initialization. Figure 2-3 show the boxplots of the mixing times under different scenarios for $p = 500$ and $p = 2,000$.

4. Proofs of Theorem 14 and Theorem 15

4.1. Proof of Theorem 14. The proof is similar to the proof of Theorem 2 of Belloni and Chernozhukov (2009), or Theorem 3.2 of Lovász and Simonovits (1993). It is based on well-known iso-perimetric inequalities for log-concave densities. We will use the following version taken from Vempala (2005) Theorem 4.2.
Figure 2. Boxplots of estimated mixing times (first row) and relative error (second row) of sampling from $\tilde{\Pi}_\gamma$ (denoted MA) and $\bar{\Pi}_\gamma$ (denoted SS), under different configurations of signal-to-noise ratio and matrix coherence. Dimension $p = 500$.

Figure 3. Boxplots of estimated mixing times (first row) and relative error (second row) of sampling from $\tilde{\Pi}_\gamma$ (denoted MA) and $\bar{\Pi}_\gamma$ (denoted SS), under different configurations of signal-to-noise ratio and matrix coherence. Dimension $p = 2000$.

Lemma 16. Let $\Theta$ be a convex subset of $\mathbb{R}^p$, and $h: \Theta \rightarrow \mathbb{R}$ a convex function. Let $\Theta = S_1 \cup S_2 \cup S_3$ be a partition of $\Theta$ into nonempty measurable components such that $d = \inf_{x_1 \in S_1, x_2 \in S_2} \|x_1 - x_2\|_2 > 0$. Then

$$\int_{S_3} e^{-h(\theta)} d\theta \geq \frac{2d}{\text{diam}(\Theta)} \min \left[ \int_{S_1} e^{-h(\theta)} d\theta, \int_{S_2} e^{-h(\theta)} d\theta \right].$$
Remark that (34) implies that for all \( u, v \in \mathbb{R}^p \), and for \( z \in \{ u, v \} \), \( A \mapsto P(z, A) - \int_A q_{u,v}(x)dx \) is a non-negative measure on \( \mathbb{R}^p \). Hence for any measurable subset \( A \) of \( \Theta \)

\[
P(u, A) - P(v, A) \leq P(u, A) - \int_A q_{u,v}(x)dx \leq P(u, \Theta) - \int_\Theta q_{u,v}(x)dx.
\]

A similar bound holds for \( P(v, A) - P(u, A) \), leading to the following result.

**Lemma 17.** If (34) holds, then for all \( u, v \in \mathbb{R}^p \),

\[
\sup_{A \subseteq \Theta} |P(u, A) - P(v, A)| \leq P(u, \Theta) \lor P(v, \Theta) - \int_\Theta q_{u,v}(x)dx,
\]

where \( a \lor b \) def = max(a, b).

With Lemma 16 and Lemma 17 in place, the proof of Theorem 14 can be done as follows. Fix \( A \subseteq \Theta \) such that \( 0 < \pi(A) < 1 \). Define

\[
S_1 \overset{\text{def}}{=} \{ \theta \in A : P(\theta, A^c \cap \Theta) < \frac{\alpha}{2} \}, \quad S_2 \overset{\text{def}}{=} \{ \theta \in A^c \cap \Theta : P(\theta, A) < \frac{\alpha}{2} \},
\]

and \( S_3 = \Theta \setminus (S_1 \cup S_2) \). Hence we have a partition \( \Theta = S_1 \cup S_2 \cup S_3 \) of \( \Theta \). If \( \pi(S_1) \leq \pi(A)/2 \) then

\[
\int_A \pi(d\theta)P(\theta, A^c \cap \Theta) \geq \frac{\alpha}{2} \pi(A \setminus S_1) \geq \frac{\alpha}{4} \pi(A). \tag{38}
\]

Similarly, if \( \pi(S_2) \leq \pi(A^c \cap \Theta)/2 \), then

\[
\int_{A^c \cap \Theta} \pi(d\theta)P(\theta, A) \geq \frac{\alpha}{2} \pi((A^c \cap \Theta) \setminus S_2) \geq \frac{\alpha}{4} \pi(A^c \cap \Theta). \tag{39}
\]

Now suppose that \( \pi(S_1) > \pi(A)/2 \) and \( \pi(S_2) > \pi(A^c \cap \Theta)/2 \). Then by reversibility

\[
\int_A \pi(d\theta)P(\theta, A^c \cap \Theta) = \frac{1}{2} \int_A \pi(d\theta)P(\theta, A^c \cap \Theta) + \frac{1}{2} \int_{A^c \cap \Theta} \pi(d\theta)P(\theta, A)
\geq \frac{\alpha}{4} (\pi(A \setminus S_1) + \pi((A^c \cap \Theta) \setminus S_2))
= \frac{\alpha}{4} \pi(S_3). \tag{40}
\]

Fix \( \theta, \vartheta \in \Theta \) such that \( \| \theta - \vartheta \| \leq r \). Therefore, by assumption we have \( \int_{\Theta} \phi_{u,v}(\theta)\ d\theta \geq \alpha \). Without any loss of generality suppose that \( P(\theta, \Theta) \geq P(\vartheta, \Theta) \). It follows from Lemma 17 that

\[
P(\theta, \Theta) - \alpha \geq \sup_{B \subseteq \Theta} |P(\theta, B) - P(\vartheta, B)| \geq P(\theta, \Theta) - P(\theta, A^c \cap \Theta) - P(\vartheta, A),
\]

where the last inequality follows by setting \( B = A = \Theta \setminus (A^c \cap \Theta) \). Hence \( P(\theta, A^c \cap \Theta) \geq \alpha - P(\vartheta, A) \). This means that if \( \| \theta - \vartheta \| \leq r \) and \( \vartheta \in S_2 \) we necessarily have \( \theta \notin S_1 \).
Hence \( d = \inf_{\theta_1 \in S_1, \theta_2 \in S_2} \| \theta_1 - \theta_2 \|_2 \geq \inf_{\theta_1 \in S_1, \theta_2 \in S_2} \| \theta_1 - \theta_2 \| \geq r \). By Lemma 16,
\[
\pi(S_3) \geq \frac{2r}{\text{diam}(\Theta)} \min(\pi(S_1), \pi(S_2)).
\]
Combining this with with (40), (38) and (39) we conclude that
\[
\Phi_{\Theta}(P) \geq \frac{\alpha}{4} \min \left( 1, \frac{2r}{\text{diam}(\Theta)} \right),
\]
as claimed.

\[\square\]

4.2. Proof of Theorem 15. The first step of the proof is a lower-bound on the conductance of \( \tilde{P}_\gamma \) that we derive in the next lemma. Given \( m \geq 1, M > 2, \zeta \in (0, 1/2) \), we define
\[
\mathcal{E}_\rho(\zeta, m, M) \overset{\text{def}}{=} \mathcal{E}_\rho \cap \left\{ z \in \mathbb{R}^n : \tilde{\Pi}_\gamma(\{\delta_s\} \times B_{m,M}(\delta_s)|z) \geq 1 - \zeta \right\}.
\]

**Lemma 18.** Assume \( H^6, H^7 \) and choose \( \gamma = \frac{\gamma_0}{n \log(p)} \) for some absolute constant \( \gamma_0 > 0 \) such that \( 4(\gamma/\sigma^2)\lambda_{\max}(X'X) \leq 1 \). Furthermore, as \( p \to \infty \), suppose that
\[
\frac{s_3^{3/2}C(X)}{\sqrt{n} \log(p)} = o(1).
\]
Take \( \rho \in (0, \bar{\rho}] \) where \( \bar{\rho} \) is as in (21). Fix \( \zeta \in (0, 1/2), m \geq 5, M > 2 \) arbitrary. Then we can find finite absolute constants \( C_0, C_1, C_2, C_3 \geq 1 \) that do not depend on \( \zeta \) such that, setting the step-size \( \tau_3 \) of the Random Walk Metropolis updates of Algorithm 7 as
\[
\tau_3 = \frac{C_0}{\|\delta\|_0^{1/2} \sqrt{n} \log(p)},
\]
the following holds. For all \( p \geq C_1 \) and all \( z \in \mathcal{E}_\rho(\zeta, m, M) \), we have
\[
\Phi_{\zeta}(\tilde{P}_\gamma) \geq \frac{C_2}{\sqrt{\rho}} e^{-C_3 s^2 \left[ 1 + \left( \frac{C(X)}{\sqrt{n} \log(p)} \sqrt{\frac{\pi}{2}} \right)^2 \right]}.
\]

**Proof.** Fix \( m \geq 5, M > 2, \zeta \in (0, 1/2), \) and \( z \in \mathcal{E}_\rho(\zeta, m, M) \) arbitrary. To shorten notation, we write \( B(\delta) \) for \( B_{m,M}(\delta) \), and \( \tilde{\mathcal{E}} \) for \( \mathcal{E}_\rho(\zeta, m, M) \). Since \( z \in \mathcal{E} \), we have
\[
1 - \tilde{\Pi}_\gamma(\{\delta_s\} \times B_{m,M}(\delta_s)|z) \leq \zeta.
\]
Let \( A \) be a measurable set such that \( \zeta < \tilde{\Pi}_\gamma(A|z) < 1 - \zeta \). We wish to lower bound the quantity \( \int_A \tilde{\Pi}_\gamma(d\theta|z)\tilde{P}_\gamma(\theta, A^c) / \min(\tilde{\Pi}_\gamma(A|z) - \zeta, \tilde{\Pi}_\gamma(A^c|z) - \zeta) \). Given
the expression of $\bar{P}_\gamma$ and $\bar{\Pi}_\gamma$ given in \cite{35} and \cite{25} respectively, we have

$$
\int_A \bar{\Pi}_\gamma(d\theta|z)\bar{P}_\gamma(\theta, A^c) \geq \frac{1}{2} \sum_\delta \int_A \bar{\Pi}_\gamma(d\theta|z)\bar{\Pi}_\gamma(\delta|\theta, z)K_\delta(\theta, A^c),
$$

$$
= \frac{1}{2} \sum_\delta \bar{\Pi}_\gamma(\delta|z) \int_A \bar{\Pi}_\gamma(d\theta|\delta, z)K_\delta(\theta, A^c),
$$

$$
\geq \frac{1}{2} \bar{\Pi}_\gamma(\delta_*|z) \int_{A \cap B(\delta_*)} \bar{\Pi}_\gamma(d\theta|\delta_*, z)K_{\delta_*}(\theta, A^c \cap B(\delta_*)).
$$

Then using the definition of $\Phi_{B(\delta)}(K_\delta)$ (the $B(\delta)$-conductance of $K_\delta$), we get

$$
\int_A \bar{\Pi}_\gamma(d\theta|z)\bar{P}_\gamma(\theta, A^c) \geq \frac{1}{2} \bar{\Pi}_\gamma(\delta_*|z)\Phi_{B(\delta_*)}(K_{\delta_*})
$$

$$
\times \min \left( \bar{\Pi}_\gamma(A_{\delta_*}|\delta_*, z), \bar{\Pi}_\gamma(A_{\delta_*}^c|\delta_*, z) \right),
$$

(42)

where $A_\delta \overset{\text{def}}{=} A \cap B(\delta)$ and $A_\delta^c \overset{\text{def}}{=} A^c \cap B(\delta)$. On the other hand, since $1 - \bar{\Pi}_\gamma(\{\delta_*\} \times B(\delta_*)|z) \leq \zeta$, we conclude that:

$$
\bar{\Pi}_\gamma(A|z) \leq \bar{\Pi}_\gamma(\{\delta_*\} \times A_{\delta_*}|z) + 1 - \bar{\Pi}_\gamma(\{\delta_*\} \times B(\delta_*)|z),
$$

$$
\leq \bar{\Pi}_\gamma(\{\delta_*\} \times A_{\delta_*}|z) + \zeta.
$$

Hence $\bar{\Pi}_\gamma(A|z) - \zeta \leq \bar{\Pi}_\gamma(\delta_*|z)\bar{\Pi}_\gamma(A_{\delta_*}|\delta_*, z)$. A similar bound holds for $\bar{\Pi}_\gamma(A^c|z) - \zeta$. We combine these with (42) to get

$$
\frac{\int_A \bar{\Pi}_\gamma(d\theta|z)\bar{P}_\gamma(\theta, A^c)}{\min (\bar{\Pi}_\gamma(A|z) - \zeta, \bar{\Pi}_\gamma(A^c|z) - \zeta)} \geq \frac{1}{2} \Phi_{B(\delta_*)}(K_{\delta_*}).
$$

Since the right-hand side of the last display is independent of $A$ we conclude that:

$$
\Phi_{\zeta}(\bar{P}_\gamma) \geq \frac{1}{2} \Phi_{B(\delta_*)}(K_{\delta_*}).
$$

(43)

To lower-bound $\Phi_{B(\delta_*)}(K_{\delta_*})$, we shall apply Theorem 14. To save some notation, in the remaining of the proof we will write $\delta$ for $\delta_*$, and $s$ for $s_*$. To lower bound $\Phi_{B(\delta)}(K_\delta)$, we will apply Theorem 14 with $\Theta$ taken as $B(\delta)$, and $\|\theta\| \overset{\text{def}}{=} \|\theta\|_2$.

Let us check that all the assumptions are satisfied. Clearly, $K_\delta$ is reversible with respect to $\bar{\Pi}_\gamma(d\theta|\delta, z) \propto e^{-h_\gamma(\delta, \theta; z)}d\theta$, and $\theta \mapsto h_\gamma(\delta, \theta; z)$ is convex \cite{2014} Theorem 2.3).

Let us first recall our notations. For $1 \leq i \leq p$, $u \in \mathbb{R}^i$, $v \in \mathbb{R}^{p-i}$, and $\delta \in \Delta$ such that $\|\delta\|_0 = i$, we write $(u, v)_\delta$ to denote the vector of $\mathbb{R}^p$, $\theta$ say, such that $[\theta]_\delta = u$, and $[\theta]_{\delta^c} = v$. When $v = (0, \ldots, 0) \in \mathbb{R}^{p-i}$, we write $(u, 0)_\delta$. If the structure $\delta$ is understood, we shall simply write $(u, v)$.

For $\tau^2 > 0$ and $u \in \mathbb{R}^s$, let $Q_{u, \tau^2}(u, \cdot)$ denote the density of the Gaussian distribution $N(u, \tau^2 I_s)$ on $\mathbb{R}^s$, and let $G_{\gamma, \delta}(u, \cdot)$ denote the density of the Gaussian distribution
Hence $K$ are accepted:

$K_\delta(\theta, A) \geq \frac{1}{2} \int_A Q_{s_\delta}^{-2}(\theta) \min \left[ 1, \frac{e^{-h_\gamma(\delta, ([u_\delta, [\theta]_{\delta^c}] + z) - h_\gamma(\delta, \theta; z)}}{e^{-h_\gamma(\delta, ([\theta]_{\delta^c}; [u_\delta, \theta]_{\delta^c}] + z)}} \right] \, du.

Hence $K_\delta$ satisfies (34) with

$$q(\theta, u) \overset{\text{def}}{=} \frac{1}{2} Q_{s_\delta}^{-2}(\theta) \min \left[ 1, \frac{e^{-h_\gamma(\delta, ([u_\delta, [\theta]_{\delta^c}] + z)}}{e^{-h_\gamma(\delta, ([\theta]_{\delta^c}; [u_\delta, \theta]_{\delta^c}] + z)}} \right] \times G_{\gamma, \delta}([u_\delta, [\theta]_{\delta^c}]) \min \left[ 1, \frac{e^{-h_\gamma(\delta, ([u_\delta, [\theta]_{\delta^c}] + z)}}{e^{-h_\gamma(\delta, ([\theta]_{\delta^c}; [u_\delta, \theta]_{\delta^c}] + z)}} \right].$$

We show in Lemma 27 and Lemma 28 of the supplement that we can find a finite absolute constant $C_0$ such that for all $p$ large enough

$$\sup_{z \in E_\rho} \sup_{\theta \in B^{(\delta)}} \left| \hat{h}_\gamma(\delta, \theta; z) - h_\gamma(\delta, \theta; z) \right| \leq C_0 R_1,$$

and

$$\sup_{z \in E_\rho} |h_\gamma(\delta, \theta_1; z) - h_\gamma(\delta, \theta_2; z)| \leq C_0 R_2 \|\theta_2 - \theta_1\| + C_0 R_3$$

for all $\theta_1, \theta_2 \in B^{(\delta)}$, such that $[\theta_1]_{\delta^c} = [\theta_2]_{\delta^c}$, where

$$R_1 \overset{\text{def}}{=} s^\gamma \left( \rho + C(X) \sqrt{(m + 1) \gamma np} \right)^2,$$

$$R_2 \overset{\text{def}}{=} s^{1/2} \left( \rho + n M \epsilon + C(X) \gamma n s^{1/2} \sqrt{(m + 1) \gamma np} \right)^2,$$

and

$$R_3 = \gamma s(\rho + n M \epsilon + C(X) \sqrt{(m + 1) \gamma np}) \left( \rho + n M \epsilon + C(X) \sqrt{(m + 1) \gamma np} \right).$$

Hence for $z \in E_\rho$, and $([u_\delta, [\theta]_{\delta^c}]), ([u_\delta, [u_{\delta^c}]]) \in B^{(\delta)}$, we have

$$\min \left[ 1, \frac{e^{-h_\gamma(\delta, ([u_\delta, [\theta]_{\delta^c}] + z)}}{e^{-h_\gamma(\delta, ([\theta]_{\delta^c}; [u_\delta, \theta]_{\delta^c}] + z)}} \right] G_{\gamma, \delta}([u_\delta, [\theta]_{\delta^c}]) \min \left[ 1, \frac{e^{-h_\gamma(\delta, ([u_\delta, [\theta]_{\delta^c}] + z)}}{e^{-h_\gamma(\delta, ([\theta]_{\delta^c}; [u_\delta, \theta]_{\delta^c}] + z)}} \right] \geq e^{-2C_0 R_1}.$$
It follows that for all $p$ large enough, $z \in \mathcal{E}_{\beta}$, and $\theta_1, \theta_2 \in \mathcal{B}(\delta)$,

$$
\int_{\mathbb{R}^p} \min(q(\theta_1, u), q(\theta_2, u)) \, du \geq \frac{e^{-C_0(2R_1 + R_3)}}{2} \times \int_{\mathcal{B}(\delta)} \min \left( Q_{s, \tau^2}(|\theta_1|_\delta, [u]_\delta)e^{-C_0R_2\|u - \theta_1\|}, Q_{s, \tau^2}(|\theta_2|_\delta, [u]_\delta)e^{-C_0R_2\|u - \theta_2\|} \right) 
\times G_{\gamma, \delta}([u]_\delta, [u]_{\delta^c}) \, du.
$$

To proceed we define

$$
\mathcal{V}_1 \eqdef \left\{ x \in \mathbb{R}^s : \|x - [\theta_1]_\delta\|_2 \leq M\epsilon, \|x - [\theta_1]_\delta\|_2 \leq \frac{\sqrt{2\pi}}{320} \frac{4M\epsilon}{s^{1/2}\log(p)} \right\},
$$

and

$$
\mathcal{V}_2 \eqdef \left\{ v \in \mathbb{R}^{p-s} : \|v\|_2 \leq \epsilon_1, \|v\|_\infty \leq \epsilon_2 \right\},
$$

where $\epsilon_1 = 2\sqrt{(1 + m)\gamma}$, $\epsilon_2 = 2\sqrt{(m + 1)\gamma \log(p)}$. We note that $\mathcal{V} \eqdef \{(u, v) : u \in \mathcal{V}_1, v \in \mathcal{V}_2\} \subset \mathcal{B}(\delta)$, so that it follows from the last display that

$$
\int_{\mathbb{R}^p} \min(q(\theta_1, u), q(\theta_2, u)) \, du \geq \frac{e^{-C_0(2R_1 + R_3)}}{2} \times \frac{e^{-C_0R_2\sqrt{\pi}}}{320} \frac{4M\epsilon}{s^{1/2}\log(p)} 
\times \left[ \inf_{x \in \mathcal{V}_1} G_{\gamma, \delta}(x, \mathcal{V}_2) \right] \int_{\mathcal{V}_1} \min \left( Q_{s, \tau^2}(|\theta_1|_\delta, x), Q_{s, \tau^2}(|\theta_2|_\delta, x) \right) \, dx. \quad (44)
$$

We recall that $G_{\gamma, \delta}(x, \cdot)$ is the density of the Gaussian distribution $\mathcal{N}(m_{\delta, (x, 0)}, \gamma \Sigma_{\delta, (x, 0)})$ on $\mathbb{R}^{p-s}$ where $\Sigma_{\delta, (x, 0)} = (I_{p-s} - \frac{\gamma}{\sigma^2} X_{\delta^c} X_{\delta^c})^{-1}$, and

$$
m_{\delta, (x, 0)} = -\frac{\gamma}{\sigma^2} (I_{p-s} - \frac{\gamma}{\sigma^2} X_{\delta^c} X_{\delta^c})^{-1} X_{\delta^c} X_\delta (|\tilde{J}_\delta(\delta, (x, 0))|_\delta - x).
$$

We need to lower bound the term $G_{\gamma, \delta}(x, \mathcal{V}_2) = \mathcal{N}(m_{\delta, (x, 0)}, \gamma \Sigma_{\delta, (x, 0)})(\mathcal{V}_2)$. It suffices to show that $\|m_{\delta, (x, 0)}\|_2 \leq \sqrt{\gamma p}$ and $\|m_{\delta, (x, 0)}\|_\infty \leq \sqrt{\gamma}$. Indeed if these inequalities hold, and if $(U_1, \ldots, U_{p-s}) \iid \mathcal{N}(0, 1)$, we have

$$
\mathbb{P} \left( \|m_{\delta, (x, 0)} + \gamma^{1/2} \Sigma_{\delta, (x, 0)}^{1/2} U\|_2 > \epsilon_1 \right) \\
\leq \mathbb{P} \left( U^T \Sigma_{\delta, (x, 0)} U > \frac{1}{\gamma} (\epsilon_1 - \sqrt{\gamma p})^2 \right) \leq \exp \left[ -\frac{2\sqrt{(m + 1)p} - \sqrt{p} - \sqrt{\text{Tr}(\Sigma_{\delta, (x, 0)})}}{2\lambda_{\max}(\Sigma_{\delta, (x, 0)})} \right],
$$

where the last inequality uses Lemma 25 in the supplement. Similarly,

$$
\mathbb{P} \left( \|m_{\delta, (x, 0)} + \gamma^{1/2} \Sigma_{\delta, (x, 0)}^{1/2} U\|_\infty > \epsilon_1 \right) \leq 2 \exp \left[ \log(p) - \frac{(m + 1)\log(p)}{2\lambda_{\max}(\Sigma_{\delta, (x, 0)})} \right],
$$

and
Noting that the largest eigenvalue of $\Sigma_{\delta,(x,0)}$ is bounded from above by $4/3$ (since $4\gamma \lambda_{\text{max}}(X'X) \leq \sigma^2$), we can easily conclude that

$$\inf_{x \in V_1} G_{\gamma,\delta}(x,V_2) \geq \frac{1}{2}, \quad (45)$$

for all $p$ large enough, and $m \geq 5$. It remains to show that $\|m_{\delta,(x,0)}\|_2 \leq \sqrt{7}$ and $\|m_{\delta,(x,0)}\|_\infty \leq \sqrt{7}$. For any $\theta$, $(m_{\delta,\theta})_j = -\frac{\gamma}{\sqrt{p}} \langle \hat{J}_\gamma(\delta, \theta) - \theta_j \rangle_\delta X_\delta^t X_\delta (\Sigma_{\delta,\theta} e_j)$, where $e_j$ denotes the $j$-th standard unit vector. Clearly, $\|\Sigma_{\delta,\theta} e_j\|_2 \leq 4/3$. Hence using the definition of the incoherence parameter $C(X)$,

$$\|m_{\delta,\theta}\|_\infty \leq \frac{4}{3} \frac{\gamma}{\sigma^2} C(X) \sqrt{n} \|\hat{J}_\gamma(\delta, \theta) - \theta_\delta\|_1.$$  

Noting that for $\theta = (x, 0) \in B(\delta)$, $\|\hat{J}_\gamma(\delta, \theta) - \theta_\delta\|_1 \leq s \gamma \left( \frac{3\rho}{2} + \frac{n\sqrt{\hat{V}(s)} M\epsilon}{\sigma^2} \right) \leq C s \gamma (\rho + n M\epsilon) = O(s \gamma n \epsilon)$, we easily get

$$\|m_{\delta,(x,0)}\|_\infty = o(\sqrt{\gamma}), \quad \text{and} \quad \|m_{\delta,(x,0)}\|_2 \leq \sqrt{p} \|m_{\delta,(x,0)}\|_\infty = o(\sqrt{p}),$$

since $s C(X) (\gamma n)^{3/2} \epsilon = o(1)$, by assumption (37). Therefore, (45) and (44) imply that

$$\int_{\mathbb{R}^p} \min(q(\theta_1, u), q(\theta_2, u)) du \geq \frac{e^{C_0(2R_1+R_3)}}{16} \frac{e^{-C_0 R_2 \frac{\sqrt{2\pi}}{320} \frac{4 M\epsilon}{s^{1/2} \log(p)}}}{4} \int_{V_1} \min\left( Q_{s,\tau_\delta}([\theta_1]_\delta, x), Q_{s,\tau_\delta}([\theta_2]_\delta, x) \right) dx.$$  

We lower bound the integral on the right-hand side of the last display using Lemma 29 in the supplement that we applied with $d \leftarrow s$, $R \leftarrow M\epsilon$, $\sigma \leftarrow \tau_\delta = \frac{\sqrt{2\pi}}{320} \frac{M\epsilon}{s \log(p)}$, and $r \leftarrow 4\tau_\delta \sqrt{s} = \frac{\sqrt{2\pi}}{320} \frac{4 M\epsilon}{s^{1/2} \log(p)}$. The lemma implies that for $\|\theta_1 - \theta_2\| \leq \tau_\delta/4$, we have

$$\int_{\mathbb{R}^p} \min(q(\theta_1, u), q(\theta_2, u)) du \geq \frac{e^{C_0(2R_1+R_3)}}{16} \frac{e^{-C_0 R_2 \frac{\sqrt{2\pi}}{320} \frac{4 M\epsilon}{s^{1/2} \log(p)}}}{4} \int_{V_1} \min\left( Q_{s,\tau_\delta}([\theta_1]_\delta, x), Q_{s,\tau_\delta}([\theta_2]_\delta, x) \right) dx.$$  

Hence $K_\delta$ satisfies all the assumption of Theorem 14 and we conclude that

$$\Phi_{B(\delta)}(K_\delta) \geq \frac{1}{64} e^{-2C_0(2R_1+R_3)} e^{-C_0 R_2 \frac{\sqrt{2\pi}}{320} \frac{4 M\epsilon}{s^{1/2} \log(p)}} \times \min\left( 1, \frac{\sqrt{2\pi}}{2 \times 320} \frac{M\epsilon}{s \log(p)} \left( \frac{2\sqrt{(m+1)\gamma p} + M\epsilon}{2\sqrt{(m+1)\gamma p} + M\epsilon} \right) \right).$$

Using $\gamma = \frac{\gamma_0}{n \log(p)}$, we check that for some absolute constant $C$,

$$R_1 \leq C \frac{sp}{n} \left( \frac{C(X)}{\log(p)} \right)^2, \quad R_3 \leq C \left( s^2 + s^{3/2} \sqrt{\frac{p}{n \log(p)}} \right).$$
and
\[
\frac{R_2 M \epsilon}{s^{1/2} \log(p)} \leq C s \left( 1 + \sqrt{\frac{p}{n} \frac{C(X)}{(\log(p))^{3/2}}} \right).
\]
Hence, we have
\[
R_1 + R_3 + \frac{R_2 M \epsilon}{s^{1/2} \log(p)} = O \left( s^2 \left[ 1 + \left( \frac{C(X)}{s^{1/2} \log(p)} \sqrt{\frac{p}{n}} \right)^2 \right] \right),
\]
and
\[
\frac{M \epsilon}{s \log(p)} \left( 2 \sqrt{(m+1) \gamma p} + M \epsilon \right) \geq \frac{C}{1 + \sqrt{m+1} s \log(p) + \sqrt{ps}},
\]
and it follows that for all \( p \) large enough,
\[
\Phi_{B(\delta)}(K_\delta) \geq C_1 \frac{1}{\sqrt{p}} e^{-C_2 s^2 \left[ 1 + \sqrt{\frac{C(X)}{s^{1/2} \log(p)} \sqrt{\frac{p}{n}}} \right]^2},
\]
(46)
for all \( p \) large enough, for absolute constants \( C_1, C_2 \). The result then follows by combining (46), and (43). \( \square \)

4.2.1. **Proof of Theorem 15.** Throughout the proof \( C \) denotes a generic universal constant, whose actual value may change from one appearance to the next. To shorten notation, we will also write \( \delta \) to denote \( \delta^{(0)} \) (the initialization of Algorithm 1 as described in Section 3.1.1). Fix \( \zeta_0 \in (0, 1/2) \). Set
\[
m_1 = \max \left( 0, \frac{8 \sigma^2}{\bar{s}(s) \gamma_0} \log \left( \frac{8FP}{\zeta_0} \right) \right), \quad \text{and} \quad \zeta = \frac{\zeta_0}{16pC_0(1+m_1+FP)},
\]
(47)
where \( C_0 \geq 1 \) is an absolute constant that we specify later. Let \( m \geq 5 \) and \( M > \max(2, \sqrt{\frac{u+2}{3}}) \) arbitrary, and set
\[
\mathcal{E}_{\bar{\rho}}(\zeta, m, M) \overset{\text{def}}{=} \mathcal{E}_{\bar{\rho}} \cap \left\{ z \in \mathbb{R}^n : \tilde{\Pi}_\lambda(B_{m,M} | z) \geq 1 - \zeta \right\}.
\]
By Markov’s inequality
\[
\mathbb{P}_* (Z \notin \mathcal{E}_{\bar{\rho}}(\zeta, m, M)) \leq \mathbb{P}_* \left[ 1 - \tilde{\Pi}_\lambda(B_{m,M} | Z) > \zeta | Z \in \mathcal{E}_{\bar{\rho},\lambda} \right] + \mathbb{P}_* (Z \notin \mathcal{E}_{\bar{\rho},\lambda}),
\]
\[
\leq \frac{1}{\zeta} \mathbb{E}_* \left[ 1 - \tilde{\Pi}_\lambda(B_{m,M} | Z) | Z \in \mathcal{E}_{\bar{\rho},\lambda} \right] + \mathbb{P}_* (Z \notin \mathcal{E}_{\bar{\rho},\lambda}).
\]
Therefore by Corollary 10 we can choose absolute constants \( m \geq 5, M > \max(2, \sqrt{\frac{u+2}{3}}) \) (that depend only on \( C_0, m_1, FP \) and \( \zeta_0 \)) such that if \( u \) and \( m_0 \) are also large enough, then we have
\[
\mathbb{P}_* (Z \notin \mathcal{E}_{\bar{\rho}}(\zeta, m, M)) \leq \frac{\zeta_0}{2},
\]
(48)
for all \( p \) large enough. By Theorem 13 and Lemma 18 there exist absolute constants \( C_1, C_2, C_3, C_4 \) that does not depend on \( \zeta \) such that: for all \( p \geq C_1 \), all \( z \in \hat{\mathcal{E}}_{\bar{\rho}}(\zeta, m, M) \),
choosing the step-size of the Random Walk Metropolis as \( \tau = \frac{C_2}{\|\delta\|^2 \sqrt{n \log(p)}} \) and choosing integer an \( K \) satisfying

\[
K \geq C_3 p \log \left( \frac{1}{\zeta} \right) \exp \left( C_4 s^2 \left[ 1 + \left( \frac{C(X)}{s^{1/2} \log(p)} \sqrt{\frac{p}{n}} \right)^2 \right] \right),
\]

we have

\[
\| \nu_0 \tilde{P}_\gamma^K - \Pi_\gamma(\cdot | z) \|_{tv} \leq 2 \sup_{A: \Pi_\gamma(A | z) \leq \zeta} | \nu_0(A) - \Pi_\gamma(A | z) |.
\]

Using this with (48), it follows that

\[
\mathbb{E}_* \left[ \| \nu_0 \tilde{P}_\gamma - \Pi_\gamma(\cdot | Z) \|_{tv} \right] \leq 2 \mathbb{E}_* \left[ \sup_{A: \Pi_\gamma(A | Z) \leq \zeta} | \nu_0(A) - \Pi_\gamma(A | z) | \mathbf{1}_{\mathcal{E}_\rho(\zeta, m, M)}(Z) \right] + \frac{\zeta_0}{2} \tag{49}
\]

Therefore to finish the proof it suffices to upper bound the first term on the right-hand side of the last display. We recall that \( \nu_0 = \nu^{(\delta)}(\cdot | z) \) (where here \( \delta \) is a short for \( \delta^{(0)} \)), and we split the term \( \Pi_\gamma(A | z) - \nu_0 \) as

\[
\left( \Pi_\gamma(A | z) - \nu^{(\delta)}_{B_1^{(\delta)^*}}(A | z) \right) + \left( \nu^{(\delta)}_{B_1^{(\delta)^*}}(A | z) - \nu^{(\delta)}(A | z) \right),
\]

where \( B_1^{(\delta)^*} \) is a short for \( B_{m_1, M}^{(\delta)} \). For any measurable set \( A \) of \( \mathbb{R}^p \) such that \( \Pi_\gamma(A | z) \leq \zeta \), if \( \Pi_\gamma(A | z) \geq \nu^{(\delta)}_{B_1^{(\delta)^*}}(A | z) \), then

\[
\left| \Pi_\gamma(A | z) - \nu^{(\delta)}_{B_1^{(\delta)^*}}(A | z) \right| \leq \zeta.
\]

We note that for \( z \in \mathcal{E}_\rho(\zeta, m, M) \),

\[
\Pi_\gamma(A | z) \geq \Pi_\gamma(\{ \delta^* \} + A \cap B_1^{(\delta)^*} | z) = \Pi_\gamma(\{ \delta^* \} + B_1^{(\delta)^*} | z) \frac{\Pi_\gamma(A \cap B_1^{(\delta)^*} | \delta^*, z)}{\Pi_\gamma(B_1^{(\delta)^*} | \delta^*, z)} \geq (1 - \zeta) \frac{\Pi_\gamma(A \cap B_1^{(\delta)^*} | \delta^*, z)}{\Pi_\gamma(B_1^{(\delta)^*} | \delta^*, z)},
\]
Using this we see that if $\bar{\Pi}_\gamma(A|z) \leq \zeta$, then 
\[ \frac{\bar{\Pi}_\gamma(A \cap B_1^{(\delta_*)}|\delta_*, z)}{\bar{\Pi}_\gamma(B_1^{(\delta_*)}|\delta_*, z)} \leq \zeta/(1 - \zeta) \leq 2\zeta, \]
so that if $\bar{\Pi}_\gamma(A|z) < \nu_1^{(\delta_*)}(A|z)$ then
\[ \left| \bar{\Pi}_\gamma(A|z) - \nu_1^{(\delta_*)}(A|z) \right| \leq \frac{\nu_1^{(\delta_*)}(A|z)}{\bar{\Pi}_\gamma(B_1^{(\delta_*)}|\delta_*, z)} - \frac{\bar{\Pi}_\gamma(A \cap B_1^{(\delta_*)}|\delta_*, z)}{\bar{\Pi}_\gamma(B_1^{(\delta_*)}|\delta_*, z)} + \zeta, \]
\[ \leq \left| \nu_1^{(\delta_*)}(A|z) - \frac{\bar{\Pi}_\gamma(A \cap B_1^{(\delta_*)}|\delta_*, z)}{\bar{\Pi}_\gamma(B_1^{(\delta_*)}|\delta_*, z)} \right| + \zeta. \]

We conclude that for $z \in \tilde{E}_\rho(\zeta, m, M)$
\[ \sup_{A: \bar{\Pi}_\gamma(A|z) \leq \zeta} \left| \bar{\Pi}_\gamma(A|z) - \nu_1^{(\delta_*)}(A|z) \right| \leq 2\zeta \left( 1 + \sup_{\bar{\Pi}_\gamma(A \cap B_1^{(\delta_*)}|\delta_*, z) > 0} \frac{\nu_1^{(\delta_*)}(A|z)}{\bar{\Pi}_\gamma(B_1^{(\delta_*)}|\delta_*, z)} \frac{\nu_1^{(\delta_*)}(A|z)}{\nu_1^{(\delta_*)}(B_1^{(\delta_*)}|\delta_*, z)} \right). \] (50)

Given $A \subseteq B_1^{(\delta_*)}$, we write
\[ \frac{\nu_1^{(\delta_*)}(A|z)}{\nu_1^{(\delta_*)}(B_1^{(\delta_*)}|\delta_*, z)} = \frac{\nu_1^{(\delta_*)}(A|\delta_*, z)}{\bar{\Pi}_\gamma(B_1^{(\delta_*)}|\delta_*, z)} \times \frac{\nu_1^{(\delta_*)}(A|\delta_*, z)}{\nu_1^{(\delta_*)}(B_1^{(\delta_*)}|\delta_*, z)}. \] (51)

Define
\[ g_\gamma(\delta, \theta; z) \overset{\text{def}}{=} \frac{1}{2} [\theta|\delta - \hat{\theta}_\delta]T_{\gamma, \delta}([\theta|\delta - \hat{\theta}_\delta] + \frac{1}{2\gamma}(\theta - \theta_\delta)(\theta - \theta_\delta)). \]

The first ratio on the right hand side of (51) can be written as
\[ \frac{\nu_1^{(\delta_*)}(A|z)}{\nu_1^{(\delta_*)}(B_1^{(\delta_*)}|\delta_*, z)} \overset{\text{def}}{=} \int_A e^{-g_\gamma(\delta, \theta; z)} d\theta \]
\[ \overset{\text{def}}{=} \int_A e^{-h_\gamma(\delta, \theta; z) - \ell(\theta, z) + \rho\|\theta_\delta\|_1} d\theta \]
\[ \times \frac{\int_{B_1^{(\delta_*)}} e^{-h_\gamma(\delta, \theta; z) - \ell(\hat{\theta}_\delta; z) + \rho\|\hat{\theta}_\delta\|_1} d\theta}{\int_{B_1^{(\delta_*)}} e^{-g_\gamma(\delta, \theta; z)} d\theta}. \] (52)

By Lemma [19] in the supplement, we have
\[ h_\gamma(\delta_*, \theta; z) = -\ell(\hat{\theta}_\delta; z) + \rho\|\hat{\theta}_\delta\|_1 + \frac{1}{2\gamma}(\theta - \theta_\delta)(\theta - \theta_\delta) - R, \]
where the remainder $R$ satisfies
\[ 0 \leq R \leq \frac{1}{2} (\theta - \theta_\delta)^T S(\theta - \theta_\delta) + \frac{\gamma}{2}\|\delta_* \cdot \nabla \ell(\theta; z) - \rho\text{sign}(\theta_\delta)\|_2^2. \] (53)
Finally we use the assumption
\[\gamma \leq 1\]
since
\[\rho \leq 1\] and
\[\gamma, \gamma \rho s \leq 1\].

For
\[\ell(\theta; z)\]
is quadratic and
\[\delta_{\Lambda} \cdot \nabla \ell \big(\hat{\theta}_{\delta_{\Lambda}}; z\big) = 0\], we have
\[\ell(\hat{\theta}_{\delta_{\Lambda}}; z) = \ell(\delta_{\Lambda}; z) + \frac{1}{2} \left(\delta_{\Lambda} - \hat{\theta}_{\delta_{\Lambda}}\right)^{T} \mathcal{I}_{\delta_{\Lambda}} (\delta_{\Lambda} - \hat{\theta}_{\delta_{\Lambda}}) = 0\].

Hence
\[-h(\delta, \theta; z) = \ell(\hat{\theta}_{\delta_{\Lambda}}; z) + \rho \|\hat{\theta}_{\delta_{\Lambda}}\|_{1} + g_{\gamma}(\delta, \theta; z) = \rho \left(\|\hat{\theta}_{\delta_{\Lambda}}\|_{1} - \|\hat{\theta}_{\delta_{\Lambda}}\|_{1}\right) + R.\]

For
\[z \in E_{\rho}, \text{ and } \theta \in B_{1}(\delta_{\Lambda}), \|\hat{\theta}_{\delta_{\Lambda}}\|_{1} - \|\hat{\theta}_{\delta_{\Lambda}}\|_{1} \leq s_{\gamma}^{1/2}(M + 1)\epsilon\].

Using this and the fact that
\[R \geq 0\],
we see that the first term on the right-hand side of (52) is upper bounded by
\[e^{(M+1)\rho s_{\gamma}^{1/2} \epsilon} \leq e^{(p)}\],
since
\[\rho s_{\gamma}^{1/2} \epsilon = o(\log(p))\] by assumption. By proceeding as in the calculations leading to (64) in the supplement, we can show that
\[-h(\delta, \theta; z) = \ell(\hat{\theta}_{\delta_{\Lambda}}; z) + \rho \|\hat{\theta}_{\delta_{\Lambda}}\|_{1} + g_{\gamma}(\delta, \theta; z) = \rho \left(\|\hat{\theta}_{\delta_{\Lambda}}\|_{1} - \|\hat{\theta}_{\delta_{\Lambda}}\|_{1}\right) + R.\]

For
\[z \in E_{\rho}, \text{ and } \theta \in B_{1}(\delta_{\Lambda}), \|\hat{\theta}_{\delta_{\Lambda}}\|_{1} - \|\hat{\theta}_{\delta_{\Lambda}}\|_{1} \leq s_{\gamma}^{1/2}(M + 1)\epsilon\].

Using this and the fact that
\[R \geq 0\],
we see that the first term on the right-hand side of (52) is upper bounded by
\[e^{(M+1)\rho s_{\gamma}^{1/2} \epsilon} \leq e^{(p)}\],
since
\[\rho s_{\gamma}^{1/2} \epsilon = o(\log(p))\] by assumption. By proceeding as in the calculations leading to (64) in the supplement, we can show that
\[-h(\delta, \theta; z) = \ell(\hat{\theta}_{\delta_{\Lambda}}; z) + \rho \|\hat{\theta}_{\delta_{\Lambda}}\|_{1} + g_{\gamma}(\delta, \theta; z) = \rho \left(\|\hat{\theta}_{\delta_{\Lambda}}\|_{1} - \|\hat{\theta}_{\delta_{\Lambda}}\|_{1}\right) + R.\]

For
\[z \in E_{\rho}, \text{ and } \theta \in B_{1}(\delta_{\Lambda}), \|\hat{\theta}_{\delta_{\Lambda}}\|_{1} - \|\hat{\theta}_{\delta_{\Lambda}}\|_{1} \leq s_{\gamma}^{1/2}(M + 1)\epsilon\].

Using this and the fact that
\[R \geq 0\],
we see that the first term on the right-hand side of (52) is upper bounded by
for all \( p \) large enough. The second term on the right hand side of (51) can be written as

\[
\frac{\nu^{(\delta)}(A|z)}{\nu^{(\delta)}(B_1^{(\delta_*)}|z)} = \int_A e^{-g_\gamma(\delta,\theta;z)} d\theta \times \frac{\int_{B_1^{(\delta_*)}} e^{-g_\gamma(\delta_*,\theta;z)} d\theta}{\int_{B_1^{(\delta_*)}} e^{-g_\gamma(\delta,\theta;z)} d\theta}.
\]

We claim that

\[
\sup_{z \in \mathcal{E}_\rho} \sup_{\theta \in B_1^{(\delta_*)}} |g_\gamma(\delta,\theta;z) - g_\gamma(\delta_*,\theta;z)| \leq C(m_1 + \text{PF}) \log(p),
\]

for some absolute constant \( C \geq 1 \) (that does not depend on \( m, M \)). (54) together with the equation right before it then implies that the second term on the right hand side of (51) is upper bounded by \( p^{C(m_1 + \text{PF})} \). Therefore we can conclude that (50) becomes

\[
\sup_{A: \Pi_0(A|z) \leq \zeta} \left| \tilde{\Pi}_{\gamma}(A|z) - \nu^{(\delta)}_{B_1^{(\delta_*)}}(A|z) \right| 1_{\mathcal{E}_\rho(\zeta/m,M)}(z) \leq 4\zeta p^{C(1+m_1+\text{PF})}.
\]

Hence taking \( C_0 \) as the constant \( C \) of the last display in the expression of \( \zeta \), we conclude from (55) and (49) that for all \( p \) large enough

\[
E_* \left[ \|\nu_0 \tilde{P}_\gamma^K - \tilde{\Pi}_{\gamma}(-|Z)\|_{\text{tv}} \right] \leq \zeta_0
\]

\[
+ 2E_* \left[ \|\nu^{(\delta)}_{B_1^{(\delta_*)}}(-|Z) - \nu^{(\delta)}(-|Z)\|_{\text{tv}} 1_{\mathcal{E}_\rho(\zeta/m,M)}(Z) \right].
\]

Noting that \( \nu^{(\delta)}_{B_1^{(\delta_*)}}(-|z) \) is the restriction of \( \nu^{(\delta)}(-|z) \) to \( B_1^{(\delta_*)} \), we have

\[
\|\nu^{(\delta)}_{B_1^{(\delta_*)}}(-|z) - \nu^{(\delta)}(-|z)\|_{\text{tv}} \leq 1 - \nu^{(\delta)}(B_1^{(\delta_*)}|z),
\]

and

\[
1 - \nu^{(\delta)}(B_1^{(\delta_*)}|z) \leq \mathbb{P} \left( \|\hat{\theta}_{\delta_0} + \mathcal{I}_{\delta_0,\gamma}^{-1/2} V\|_{\delta_0} - |\theta_0|_{\delta_0} 2 > M\epsilon \right)
\]

\[
+ \mathbb{P} \left( \|\hat{\theta}_{\delta_0} + \mathcal{I}_{\delta_0,\gamma}^{-1/2} V\|_{\delta_0-\delta} 2 > 2(m_1 + 1)\gamma p \right)
\]

\[
+ \mathbb{P} \left( \|\hat{\theta}_{\delta_0} + \mathcal{I}_{\delta_0,\gamma}^{-1/2} V\|_{\delta_0-\delta} 2 > 4(m_1 + 1)\gamma \log(p) \right)
\]

\[
+ \mathbb{P} \left( \gamma \|W\|_2^2 > 2(m_1 + 1)\gamma p \right)
\]

\[
+ \mathbb{P} \left( \gamma \|W\|_2 < 4(m_1 + 1)\gamma \log(p) \right),
\]

where \( V = (V_1, \ldots, V_s) \overset{iid}{\sim} \mathcal{N}(0, 1) \), and \( W = (W_1, \ldots, W_{p-s_0-1}) \overset{iid}{\sim} \mathcal{N}(0, 1) \). For \( z \in \mathcal{E}_\rho \) (which implies that \( \|\hat{\theta}_0 - |\theta_0|\|_2 \leq \epsilon \), as seen in (16)), given that \( \epsilon = \frac{\sigma}{\nu(\delta)} \sqrt{72(m_0 + 1)} \frac{\delta \log(p)}{n} \), and noting that the largest eigenvalue of \( \mathcal{I}_{\gamma,\delta}^{-1} \) is \( \sigma^2/(\nu(\delta)) \),
we can then use standard Gaussian deviation bounds to conclude that the sum of the first and last two terms on the right hand side of (58) is upper bounded by

\[
P\left(\|V\|_2 > \frac{(M - 1)\epsilon}{\sqrt{\frac{\sigma^2}{nv(\delta)}}}\right) + P\left(\|W\|_2 > \sqrt{2(1 + m_1)p}\right)
\]

\[
+ P\left(\|W\|_\infty > 2\sqrt{(1 + m_1)\log(p)}\right) \leq \frac{1}{p^{\rho_\epsilon(m_0 + 1)}} + e^{-m_1p/4} + \frac{1}{p^{m_1}} \leq \zeta_0,
\]

for all \(p\) large enough, and \(m_1 \geq 1\). Suppose that for \(z \in \mathcal{E}_\rho(m, M, \gamma), ||\hat{\theta}_\delta||_\infty \leq \sqrt{(m_1 + 1)\gamma \log(p)}\) (which implies that \(||\hat{\theta}_\delta||_2 \leq F\gamma^{1/2}(m_1 + 1)\gamma \log(p))\). Therefore the sum of the second and third terms on the right hand side of (58) is upper bounded by

\[
P\left(\|Z^{1/2}V\|_2 > \frac{1}{2}(m_1 + 1)\gamma p\right) + P\left(\|Z^{-1/2}V\|_\infty > \sqrt{(m_1 + 1)\gamma \log(p)}\right)
\]

\[
\leq \exp\left\{ -\frac{1}{2} \left( \frac{v(s)}{\sigma} \sqrt{\frac{1}{2}(m_1 + 1)\gamma np - \sqrt{FP}} \right)^2 \right\}
\]

\[
+ 2\exp\left(\log(FP) - \frac{1}{2}(m_1 + 1)\frac{v(s)}{2\sigma^2} \gamma n \log(p)\right) \leq e^{-FP} + 2FP e^{-\frac{m_1 v(s)\gamma n}{2\sigma^2}} \leq \frac{\zeta_0}{4},
\]

for all \(p\) large enough, given the choice of \(m_1 + 1 \geq \frac{2\sigma^2}{2v(s)} \gamma n \log\left(\frac{4FP}{\zeta_0}\right)\). We combine this bound with (58), (57), and (56) to conclude that

\[
\mathbb{E}\left[\|\nu_0 \tilde{P}_\gamma - \Pi_\gamma(\cdot|z)\|_4\right] \leq 2\zeta_0
\]

\[
+ 2P\left[||\hat{\theta}_\delta(Z)||_\infty > \sqrt{(m_1 + 1)\gamma \log(p)}\right].
\] (59)

We note that \(\hat{\theta}_\delta(Z)\) is the ordinary least square estimate in the regression model 

\[Z = X_\delta \theta + \sigma E, \text{ where } E \sim \mathcal{N}(0, I_n).\] It follow that \(\hat{\theta}_\delta(Z)\) \(\sim \mathcal{N}(0, \sigma^2 Q),\) where

\[Q = (X'_{\delta-\delta} (I_{s_\epsilon} - X_{\delta_s} (X'_{\delta_s} X_{\delta_s})^{-1} X'_{\delta_s}) X_{\delta-\delta})^{-1}.
\]

We also note that for any unit vector \(u \in \mathbb{R}^{||\delta||_0-s_\epsilon}\), we have

\[u'Q^{-1}u \geq u'X'_{\delta-\delta} X_{\delta-\delta} u - \frac{1}{nv(s_\epsilon)} \|X'_{\delta_s} X_{\delta-\delta} u\|^2 \geq \frac{\nu v(s)}{2}, \]

for all \(p\) large enough, since \(C(X)s_\epsilon/\sqrt{n} = o(1)\). Therefore the largest eigenvalue of \(Q\) is upper bounded by \(2/\sqrt{\nu v(s)}\). As a result, and since \(\gamma n \log(p) = \gamma_0\), and \(FP = O(1),\)
by Gaussian tail bounds (Lemma 25), we have
\[
P_\ast\left( \|\hat{\theta}_\delta(Z)\|_{\delta-\delta_*} \leq \sqrt{(m_1 + 1)\gamma \log(p)} \right)
\leq \exp\left\{ \log(\text{FP}) - \frac{(m_1 + 1)\nu(s)\gamma n \log(p)}{8\sigma^2} \right\} \leq \frac{\zeta_0}{2},
\]
for all \( p \) large enough. Hence (59) becomes
\[
\mathbb{E}_\ast \left[ \|\nu_0 \hat{P}^K_\gamma - \tilde{\Pi}_\gamma(\cdot|z)\|_{\text{tv}} \right] \leq 3\zeta_0,
\]
as claimed. To complete the proof, it remains to establish (54). To that end, we set
\[
diff(\theta) \overset{\text{def}}{=} g_\gamma(\delta, \theta; z) - g_\gamma(\delta_*, \theta; z).
\]
Since \( \delta \supseteq \delta_* \), we have
\[
2\diff(\theta) = ((\theta)_\delta - \hat{\theta}_\delta)'I_{\gamma,\delta}((\theta)_\delta - \hat{\theta}_\delta) - ((\theta)_{\delta_*} - \hat{\theta}_*)'I_{\gamma}((\theta)_{\delta_*} - \hat{\theta}_*)
\]
\[
+ \frac{1}{\gamma}[(\theta)_{\delta-\delta_*} - (\theta)_{\delta-\delta_*}].
\]
We split the term ((\theta)_\delta - \hat{\theta}_\delta)'I_{\gamma,\delta}((\theta)_\delta - \hat{\theta}_\delta) - ((\theta)_{\delta_*} - \hat{\theta}_*)'I_{\gamma}((\theta)_{\delta_*} - \hat{\theta}_*) as
\[
((\theta)_{\delta} - (\theta)_{\delta_*})'I_{\gamma,\delta}((\theta)_{\delta} - \hat{\theta}_\delta) - ((\theta)_{\delta_*} - (\theta)_{\delta_*})'I_{\gamma}((\theta)_{\delta_*} - \hat{\theta}_*)
\]
\[
+ 2(\theta)_{\delta} - (\theta)_{\delta_*})'I_{\gamma,\delta}((\theta)_{\delta} - \hat{\theta}_\delta) - 2((\theta)_{\delta_*} - (\theta)_{\delta_*})'I_{\gamma}((\theta)_{\delta_*} - \hat{\theta}_*)
\]
\[
+ (\theta)_{\delta} - (\theta)_{\delta_*})'I_{\gamma,\delta}((\theta)_{\delta} - (\theta)_{\delta_*}) - (\hat{\theta}_\delta - (\theta)_{\delta_*})'I_{\gamma}((\theta)_{\delta_*} - (\theta)_{\delta_*}).
\]
We calculate that
\[
((\theta)_{\delta} - (\theta)_{\delta_*})'I_{\gamma,\delta}((\theta)_{\delta} - (\theta)_{\delta_*}) - ((\theta)_{\delta_*} - (\theta)_{\delta_*})'I_{\gamma}((\theta)_{\delta_*} - (\theta)_{\delta_*})
\]
\[
= 2[\theta - (\theta)_{\delta_*}]' \left( \frac{1}{\sigma^2} X'_{\delta - \delta_*} X_{\delta - \delta_*} \right) [\theta]_{\delta - \delta_*}
\]
\[
+ ([\theta]_{\delta - \delta_*}' \left( \frac{1}{\sigma^2} X'_{\delta - \delta_*} X_{\delta - \delta_*} \right) [\theta]_{\delta - \delta_*}.
\]
Since \( \hat{\theta}_\delta = (X'_{\delta} X_{\delta})^{-1} X'_{\delta} z \), we get \( \theta - (\theta)_{\delta_*} = (X'_{\delta} X_{\delta})^{-1} X'_{\delta} (z - X\theta) \), for all \( \delta \supseteq \delta_* \). We use this to calculate that
\[
((\theta)_{\delta} - (\theta)_{\delta_*})'I_{\gamma,\delta}((\theta)_{\delta} - (\theta)_{\delta_*}) - ((\theta)_{\delta_*} - (\theta)_{\delta_*})'I_{\gamma}((\theta)_{\delta_*} - (\theta)_{\delta_*})
\]
\[
= -\frac{1}{\sigma^2} [\theta]_{\delta - \delta_*}' X'_{\delta - \delta_*} (z - X\theta),
\]
and
\[
(\hat{\theta}_\delta - (\theta)_{\delta_*})'I_{\gamma,\delta}(\hat{\theta}_\delta - (\theta)_{\delta_*}) - (\hat{\theta}_\delta - (\theta)_{\delta_*})'I_{\gamma}(\hat{\theta}_\delta - (\theta)_{\delta_*})
\]
\[
= \frac{1}{\sigma^2} (z - X\theta)' X_{\delta - \delta_*} (X'_{\delta' - \delta_*} X_{\delta' - \delta_*})^{-1} X'_{\delta - \delta_*} (z - X\theta).
All together, we have

\[
\text{diff}(\theta) = \frac{1}{\sigma^2} (z - X\theta_*)'X_{\delta-\delta_*} (X_{\delta-\delta_*}'X_{\delta-\delta_*})^{-1}X_{\delta-\delta_*}'(z - X\theta_*) \\
+ \frac{2}{\sigma^2} [\theta - \theta_*]' \delta_{\delta-\delta_*} (X_{\delta-\delta_*}' X_{\delta-\delta_*}) \left[\theta \right]_{\delta-\delta_*} - \frac{2}{\sigma^2} [\theta]_{\delta-\delta_*} X_{\delta-\delta_*}' (z - X\theta_*) \\
- \frac{1}{\gamma} [\theta]_{\delta-\delta_*}' \left( I_{\parallel \delta_{0-s_*}} - \frac{\gamma}{\sigma^2} X_{\delta-\delta_*}' X_{\delta-\delta_*} \right) [\theta]_{\delta-\delta_*}.
\]

For $\theta \in B^{(\delta_*)}$, and $z \in \mathcal{E}_{\rho}$, we have

\[
\frac{1}{\sigma^2} (z - X\theta_*)'X_{\delta-\delta_*} (X_{\delta-\delta_*}'X_{\delta-\delta_*})^{-1}X_{\delta-\delta_*}'(z - X\theta_*) \leq \frac{8(m_0 + 1)FP}{v(s)} \log(p).
\]

We check that for $\theta \in B^{(\delta_*)}$, and $z \in \mathcal{E}_{\rho}$,

\[
\left| \frac{2}{\sigma^2} [\theta - \theta_*]' \delta_{\delta-\delta_*} (X_{\delta-\delta_*}' X_{\delta-\delta_*}) \left[\theta \right]_{\delta-\delta_*} \right| \leq CFP^{1/2} \frac{\mu_0 s_*}{\sqrt{n \log(p)}} \log(p) = o(\log(p)),
\]

\[
\left| \frac{2}{\sigma^2} [\theta]_{\delta-\delta_*}' X_{\delta-\delta_*}'(z - X\theta_*) \right| \leq \frac{2}{\sigma} \frac{FP}{\sqrt{\log(p)}} \sqrt{8(m_0 + 1)(m + 1) \log(p)} = o(\log(p)),
\]

and

\[
0 \leq \frac{1}{\gamma} [\theta]_{\delta-\delta_*}' \left( I_{\parallel \delta_{0-s_*}} - \frac{\gamma}{\sigma^2} X_{\delta-\delta_*}' X_{\delta-\delta_*} \right) [\theta]_{\delta-\delta_*} \leq 4(m_1 + 1) \log(p).
\]

We deduce easily (54), and this complete the proof of the theorem. □
Appendix A. Proof of Theorem 5

A.1. Some Preliminary results. Throughout the proof, \( \theta_* \) is the true value of the parameter as introduced in \( \Delta \). \( \delta_* \) denotes its sparsity structure, and \( s_* = \| \theta_* \|_0 \). The first lemma is taken from Atchadé (2015) Lemma 2 and gives an approximation of the function \( h_\gamma(\delta, \theta; z) \) for \( \gamma \) small.

Lemma 19. Assume \( H_\Delta \) and let \( h_\gamma \) as in (4). For all \( \delta \in \Delta, \gamma > 0, u \in \mathbb{R}^p, \) and \( z \in \mathbb{Z} \), we have

\[
-\ell(u_\delta; z) + \rho \| u_\delta \|_1 + \frac{1}{2\gamma} \| u - u_\delta \|_2^2 - \frac{1}{2}(u - u_\delta)'S(u - u_\delta) \geq h_\gamma(\delta, u; z)
\]

where \( R_\gamma(\delta, u; z) \) satisfies

\[
0 \leq R_\gamma(\delta, u; z) \leq \frac{\gamma}{2} \| \delta \cdot \nabla \ell(u; z) - \rho \text{sign}(u_\delta) \|_2^2.
\]

Proof. Using the definition of \( h_\gamma \) in (4), and the definition of \( S \) in \( H_\Delta \) we have

\[
h_\gamma(\delta, u; z) \leq -\ell(u; z) - \langle \nabla \ell(u; z), u_\delta - u \rangle + \frac{1}{2\gamma} \| u - u_\delta \|_2^2 + \rho \| u_\delta \|_1
\]

\[
\leq -\ell(u_\delta; z) + \rho \| u_\delta \|_1 - \frac{1}{2}(u - u_\delta)'S(u - u_\delta) + \frac{1}{2\gamma} \| u - u_\delta \|_2^2,
\]

which is the first inequality. To prove the second, we note that for any \( v \in \mathbb{R}^p \),

\[
-\ell(u; z) - \langle \nabla \ell(u; z), v - u \rangle = -\ell(u_\delta; z) - \langle \nabla \ell(u; z), v - u_\delta \rangle + \mathcal{F}(u|u_\delta),
\]

where \( \mathcal{F}(u|u_\delta) \) def = \( -[\ell(u; z) - \ell(u_\delta; z) - \langle \nabla \ell(u_\delta; z), u - u_\delta \rangle + \langle \nabla \ell(u; z) - \nabla \ell(u_\delta; z), u - u_\delta \rangle] \). By Taylor expansion with integral remainder and \( H_\Delta \) we have

\[
\mathcal{F}(u|u_\delta) = (u - u_\delta)' \left[ \int_0^1 t \nabla (2) \ell(u + t(u - u_\delta)) dt \right] (u - u_\delta) \geq -\frac{1}{2}(u - u_\delta)'S(u - u_\delta).
\]

Hence for all \( u \in \mathbb{R}^p \) and all \( v \in \mathbb{R}^p \),

\[
\ell(u; z) - \langle \nabla \ell(u; z), v - u \rangle \geq -\ell(u_\delta; z) - \langle \nabla \ell(u; z), v - u_\delta \rangle - \frac{1}{2}(u - u_\delta)'S(u - u_\delta).
\]

By convexity of the \( \ell^1 \)-norm, \( \| v \|_1 \geq \| u_\delta \|_1 + \langle \text{sign}(u_\delta), v - u_\delta \rangle \). We combine the last two inequalities to conclude that

\[
\ell(u; z) - \langle \nabla \ell(u; z), v - u \rangle + \rho \| v \|_1 + \frac{1}{2\gamma} \| v - u \|_2^2 \geq -\ell(u_\delta; z) + \rho \| u_\delta \|_1 + \langle \rho \text{sign}(u_\delta) - \nabla \ell(u; z), v - u_\delta \rangle + \frac{1}{2\gamma} \| v - u_\delta \|_2^2 + \frac{1}{2\gamma} \| u - u_\delta \|_2^2 - \frac{1}{2}(u - u_\delta)'S(u - u_\delta).
\]
The second inequality follows by noting that \( \langle \rho \text{sign}(u_\delta) - \nabla \ell(u; z), v - u_\delta \rangle + \frac{1}{2\gamma} \| v - u_\delta \|^2 \geq -\frac{1}{2\gamma} \| \text{sign}(u_\delta) - \nabla \ell(u; z) \|^2 \).

The next result gives a lower bound on the normalizing constant of \( \tilde{\Pi}_\gamma \).

**Lemma 20.** Assume H1-H2. For \( \gamma > 0, z \in \mathcal{Z} \), let \( \tilde{\mathcal{C}}_\gamma(z) \) denote the normalizing constant of \( \tilde{\Pi}_\gamma \). If \( \gamma \lambda_{\max}(\bar{S}) < 1 \), then we have

\[
\sqrt{\det(|I_p - \gamma \bar{S}|_{\delta_j})} (2\pi\gamma)^{-\frac{p}{2}} \tilde{\mathcal{C}}_\gamma(z) \geq \omega_{\delta_\gamma} e^{\ell(\theta_{\gamma}; z)} e^{-\rho\|\theta_{\gamma}\|_1} \left( \frac{\rho^2}{\bar{\kappa}(s_*) + \rho^2} \right)^{s_*},
\]

where for a matrix \( A \in \mathbb{R}^{p \times p} \), and \( \delta \in \Delta \), the notation \( [A]_{\delta_\epsilon} \) denote the sub-matrix of \( A \) obtained after removing the rows and columns of \( A \) which \( \delta_j = 1 \).

**Proof.** By definition we have

\[
\tilde{\mathcal{C}}_\gamma(z) = \sum_{\delta \in \Delta} \omega_{\delta} (2\pi\gamma)^{-\|\delta\|_0} \left( \frac{2}{2} \right)^{-\|\delta\|_0} \int_{\mathbb{R}^p} e^{-h_{\gamma}(\delta, u; z)} du 
\geq \omega_{\delta_\gamma} (2\pi\gamma)^{-\|\delta_\gamma\|_0} \left( \frac{2}{2} \right)^{-\|\delta_\gamma\|_0} \int_{\mathbb{R}^p} e^{-h_{\gamma}(\delta, u; z)} du.
\]

By the first inequality of Lemma 19, \( \mathcal{C}_\gamma(z) \geq \omega_{\delta_\gamma} (2\pi\gamma)^{-\|\delta_\gamma\|_0} \left( \frac{2}{2} \right)^{-\|\delta_\gamma\|_0} \int_{\mathbb{R}^p} e^{\ell(u_{\gamma}; z)} e^{-\rho\|u_{\gamma}\|_1} e^{-\frac{1}{2\gamma}(u_{\gamma} - u_\delta)(I_p - \gamma \bar{S})(u_{\gamma} - u_\delta) du.}
\]

The integrand in the last display is a multiplicatively separable function of \( [u]_{\delta_\gamma} \) and \( [u]_{\delta_\epsilon} \). Integrating out \( [u]_{\delta_\epsilon} \) then yields

\[
\tilde{\mathcal{C}}_\gamma(z) \geq \frac{(2\pi\gamma)^{\frac{p}{2}}}{\sqrt{\det(|I_p - \gamma \bar{S}|_{\delta_j})}} \omega_{\delta_\gamma} \left( \frac{2}{2} \right)^{-\|\delta_\gamma\|_0} \int_{\mathbb{R}^p} e^{\ell(u; z)} e^{-\rho\|u\|_1} \mu_{\delta_\gamma}(du).
\]

The lower bound on the term \( \omega_{\delta_\gamma} \left( \frac{2}{2} \right)^{-\|\delta_\gamma\|_0} \int_{\mathbb{R}^p} e^{\ell(u; z)} e^{-\rho\|u\|_1} \mu_{\delta_\gamma}(du) \) established in Atchade (2017) Lemma 11 is then employed to deduce (61).

**Lemma 21.** Assume H1-H2. Let \( \bar{p} > 0, \rho \in (0, \bar{p}] \), and \( \gamma > 0 \) be such that \( 4\gamma \lambda_{\max}(\bar{S}) \leq 1 \). If for all \( \delta \in \Delta \), and \( \theta \in \mathbb{R}^p \),

\[
\log \mathbb{E}_* \left[ e^{\left( 1 - \frac{\gamma}{2} \right) \langle \nabla \ell(\theta_{\gamma}; Z), \theta_{\gamma} \rangle + \mathcal{L}_\gamma(\delta, \theta; Z) \rangle \mathcal{I}_{\bar{p}_{\gamma}}(Z) \right] 
\leq -\frac{1}{2} \bar{r}_0(\| \theta - \theta_{\gamma} \|_2) \mathcal{I}_{\{ \| \theta - \theta_{\gamma} \|_{s_*} \leq \bar{p}(\| \theta - \theta_{\gamma} \|_{s_*})_{1}, \| \theta - \theta_{\gamma} \|_1 \}}(\theta),
\]

for some rate function \( \bar{r}_0 \), then \( \tilde{\Pi}_\gamma(\cdot; z) \) is well-defined for \( \mathbb{P}_* \)-almost all \( z \in \mathcal{E}_{\bar{p}} \). Furthermore

\[
\mathbb{E}_* \left[ \mathcal{I}_{\bar{p}_{\gamma}}(Z) \tilde{\Pi}_\gamma(\| \delta \|_0 > s_* + \eta| Z) \right] \leq \frac{1}{p^{\bar{m}_0}},
\]
for all \( m_0 \geq 1 \), where

\[
\eta \overset{\text{def}}{=} \frac{2}{u} \left[ m_0 + 2s_\ast + \frac{a_0}{2\log(p)} + \frac{s_\ast \log \left( 1 + \frac{\hat{\kappa}(s_\ast)}{\rho^2} \right)}{\log(p)} \right. \\
+ \left. \frac{\gamma}{2\log(p)} \left( Tr(S - \bar{S}) \right) + \frac{2\gamma^2}{\log(p)} \left( \lambda_{max}(\bar{S}) Tr(\bar{S}) + 4\|S\|_F^2 \right) \right],
\]

and

\[
a_0 \overset{\text{def}}{=} -\min_{x > 0} \left[ r_0(x) - 4\rho s_\ast^{1/2} x \right].
\]

**Proof.** That \( \Pi_\gamma(z) \) is well-defined is equivalent to the statement that its normalizing constant, denoted \( C_\gamma(z) \), is finite. Hence it suffices to establish that \( C_\gamma(z) \) is finite for \( \mathbb{P}_\ast \)-almost all \( z \) in \( \mathcal{E}_p \). By using the second inequality of Lemma 19 we have

\[
\frac{2(2\pi\gamma)^{\frac{p}{2}}C_\gamma(z)}{e^{\ell(\theta,z) - \rho\|\theta_\ast\|_1}} = \sum_{\delta \in \Delta} \omega_\delta \left( \frac{\rho}{2} \right)^{\|\theta_0\|_0} \left( \frac{1}{2\pi\gamma} \right)^{\frac{p-\|\theta_0\|_0}{2}} \int_{\mathbb{R}^p} \frac{e^{-h_\gamma(\delta,\theta;z)}}{e^{\ell(\theta,z) - \rho\|\theta_\ast\|_1}} d\theta
\]

\[
\leq \sum_{\delta \in \Delta} \omega_\delta \left( \frac{\rho}{2} \right)^{\|\theta_0\|_0} \left( \frac{1}{2\pi\gamma} \right)^{\frac{p-\|\theta_0\|_0}{2}} \int_{\mathbb{R}^p} \frac{e^{\ell(\theta_0;z) - \rho\|\theta_\ast\|_1}}{e^{\ell(\theta,z) - \rho\|\theta_\ast\|_1}} e^{-\frac{1}{\gamma}(\theta - \theta_\delta)'(I_p - \gamma S)(\theta - \theta_\delta) + R_\gamma(\delta,\theta;z) + 3\gamma \rho^2 \|\theta_0\|_0} d\theta.
\]

For \( z \in \mathcal{E}_p \), and \( \delta \in \Delta \), we use the convexity of the square norm and \( \Pi_\gamma \) to bound the term \( R_\gamma(\delta,\theta;z) \) (given in Lemma 19) as

\[
R_\gamma(\delta,\theta;z) \leq \frac{\gamma}{2} \| \delta \cdot \nabla \ell(\theta;z) - \delta \cdot \nabla \ell(\theta_\delta;z) \|_2 \leq 2\| \delta \cdot \nabla \ell(\theta_\delta;z) \|_2^2 + 2\gamma \kappa(\|\delta\|_0)(\theta - \theta_\delta)'(\theta - \theta_\delta) + 3\gamma \rho^2 \|\delta\|_0.
\]

It follows that for \( z \in \mathcal{E}_p \),

\[
-\frac{1}{2\gamma}(\theta - \theta_\delta)'(I_p - \gamma S)(\theta - \theta_\delta) + P_\gamma(\delta,\theta;z) \leq -\frac{1}{2\gamma}(\theta - \theta_\delta)'A_\delta(\theta - \theta_\delta)
\]

\[
+ 2\gamma \| \delta \cdot \nabla \ell(\theta_\delta;z) - \delta \cdot \nabla \ell(\theta_\ast;z) \|_2^2 + 3\gamma \rho^2 \|\theta_0\|_0.
\]

where \( A_\delta \overset{\text{def}}{=} I_p - \gamma(1 + 4\gamma \kappa(\|\delta\|_0)\bar{S}) \). If \( \delta = (1, \ldots, 1) \), \[64\] still holds with \( A_\delta = 0 \). Note also that if \( \delta \neq (1, \ldots, 1) \), the matrix \( A_\delta \) is positive definite under the assumption \( 4\gamma \lambda_{max}(\bar{S}) \leq 1 \). We recall the notation \( \mathcal{L}_\gamma(\delta,\theta;z) \) introduced in \[8\] of the main manuscript, and use it to write

\[
\ell(\theta_\delta;z) - \ell(\theta_\ast;z) + 2\gamma \| \delta \cdot \nabla \ell(\theta_\delta;z) - \delta \cdot \nabla \ell(\theta_\ast;z) \|_2^2 = \langle \nabla \ell(\theta_\ast;z), \theta - \theta_\ast \rangle + \mathcal{L}_\gamma(\delta,\theta_\delta;z)
\]

\[
= \frac{\rho}{\rho} \langle \nabla \ell(\theta_\ast;z), \theta_\delta - \theta_\ast \rangle + \left( 1 - \frac{\rho}{\rho} \right) \langle \nabla \ell(\theta_\ast;z), \theta_\delta - \theta_\ast \rangle + \mathcal{L}_\gamma(\delta,\theta_\delta;z). \]

\[65\]
Since \( \| \nabla \ell(\theta_*; z) \|_\infty \leq \rho/2 \), for \( z \in \mathcal{E}_\rho \), we have \( \frac{\rho}{2} \| (\nabla \ell(\theta_*; z), \theta_\delta - \theta_*) \| \leq (\rho/2) \| \theta_\delta - \theta_* \|_1 \). Using this, and accounting for all the terms, we obtain

\[
\frac{(2\pi\gamma)^\frac{3}{2} c_\rho(z)}{e^{\ell(\theta_*; z) - \rho}\|\theta_*\|_1} 1_{\mathcal{E}_\rho}(z) \leq \sum_{\delta \in \Delta} \omega_\delta \left( \frac{\rho}{2} \right)^{\|\delta\|_1} \left( \frac{1}{2\pi\gamma} \right)^{\frac{p-1}{2}} \\
\times e^{3\gamma\|\delta\|_0} \delta^2 1_{\mathcal{E}_\rho}(z) \int_{\mathbb{R}^p} e^{d(\theta_\delta) + \frac{1}{2\pi}(\nabla \ell(\theta_*; z), \theta_\delta - \theta_*) + C_\gamma(\delta, \theta_*; z)} e^{-\frac{1}{\rho}(\theta_\delta - \theta_*)' A_\delta (\theta_\delta - \theta_*)} d\theta,
\]

where \( d(\theta) \overset{\text{def}}{=} \frac{\rho}{2}\| \theta - \theta_* \|_1 - \rho (\| \theta \|_1 - \| \theta_* \|_1) \). Taking the expectation on both sides and using Fubini’s theorem and (62) gives

\[
\mathbb{E}_* \left[ \frac{(2\pi\gamma)^\frac{3}{2} c_\rho(Z)}{e^{\ell(\theta_*; z) - \rho}\|\theta_*\|_1} 1_{\mathcal{E}_\rho}(Z) \right] \leq \sum_{\delta \in \Delta} \omega_\delta \left( \frac{\rho}{2} \right)^{\|\delta\|_1} \left( \frac{1}{2\pi\gamma} \right)^{\frac{p-1}{2}} \\
\times e^{3\gamma\|\delta\|_0} \delta^2 \int_{\mathbb{R}^p} e^{d(\theta_\delta) - \frac{1}{\rho}(\theta_\delta - \theta_*)' A_\delta (\theta_\delta - \theta_*)} d\theta,
\]

All the integrals on the right-hand side of the last inequality are finite since the matrices \( A_\delta \) are symmetric positive definite, and \( d(\theta) \sim -\frac{\rho}{2}\| \theta \|_1 \), for \( \| \theta \|_1 \) large. Hence for \( \mathbb{P}_* \)-almost all \( z \in \mathcal{E}_\rho \), \( \tilde{C}_\gamma(z) \) is finite as claimed.

To establish the second part of the lemma, we first note that for any \( z \in Z \), and any measurable subset \( B \) of \( \Delta \times \mathbb{R}^p \), using (61), the second inequality of Lemma 19 (64), and similar calculations as above, we get

\[
\Pi_\gamma(B|z) \leq \sqrt{\det \left( [I_p - \gamma \mathbf{L}]_{\delta_0} \right)} \left( 1 + \frac{\bar{k}(s_\delta)}{\rho^2} \right)^{s_\delta} \sum_{\delta \in \Delta} \omega_\delta \left( \frac{\rho}{2} \right)^{\|\delta\|_1} \\
\times e^{3\gamma\|\delta\|_0} \delta^2 \int_{B(\delta)} e^{-\rho}\|\delta\|_1 e^{\ell(\theta_\delta; z)} e^{2\gamma\| \delta \|_0} \| \nabla \ell(\theta_*; z) - \delta \nabla \ell(\theta_*; z) \|_2 e^{-\frac{1}{\rho}(\theta_\delta - \theta_*)' A_\delta (\theta_\delta - \theta_*)} d\theta,
\]

where \( B(\delta) \overset{\text{def}}{=} \{ \theta \in \mathbb{R}^p : (\delta, \theta) \in B \} \). For some arbitrary \( \eta > 0 \), Let \( \mathcal{A}_\delta \overset{\text{def}}{=} \{ \delta \in \Delta : \| \delta \|_0 > s_\delta + \eta \} \), and \( \tilde{A}_\delta = \mathcal{A}_\delta \times \mathbb{R}^p \). It follows from the last display (applied to \( B = \tilde{A}_\delta \times \mathbb{R}^p \)), together with (65) and Fubini’s theorem that for \( z \in \mathcal{E}_\rho \),

\[
\mathbb{E}_* \left[ \Pi_\gamma(\tilde{A}_\delta|Z) 1_{\mathcal{E}_\rho}(Z) \right] \leq \sqrt{\det \left( [I_p - \gamma \mathbf{L}]_{\delta_0} \right)} \left( 1 + \frac{\bar{k}(s_\delta)}{\rho^2} \right)^{s_\delta} \sum_{\delta \in \mathcal{A}_\delta} \omega_\delta \left( \frac{\rho}{2} \right)^{\|\delta\|_1} \\
\times \sqrt{\det([A_\delta]_{\delta}^\top)} \int_{\mathbb{R}^p} e^{d(\theta)} e^{\ell(\theta_\delta; z) - \delta \nabla \ell(\theta_*; z)} e^{2\gamma\| \delta \|_0} \| \nabla \ell(\theta_*; z) - \delta \nabla \ell(\theta_*; z) \|_2 e^{-\frac{1}{\rho}(\theta_\delta - \theta_*)' A_\delta (\theta_\delta - \theta_*)} d\theta,
\]

where \( \det([A_\delta]_{\delta}^\top) \) is taken as 1 if \( \delta \equiv 1 \). We claim that for all \( \theta \in \mathbb{R}^p \),

\[
d(\theta) + \log \mathbb{E}_* \left[ e^{\left( 1 - \frac{\rho}{2} \right)(\nabla \ell(\theta_*; z), \theta_\delta - \theta_*) + C_\gamma(\delta, \theta_*; z)} 1_{\mathcal{E}_\rho}(Z) \right] \leq -\frac{\rho}{4}\| \theta - \theta_* \|_1 + \frac{a_\delta}{2},
\]
where \(a_0 = -\min_{x>0} \left[ r_0(x) - 4\rho s_*^{1/2} \right] > 0\). Obviously this claim is true if \(\theta = \theta_*\).

Suppose now that \(\theta \neq \theta_*\). First using the notation \(\delta^c \overset{\text{def}}{=} 1 - \delta\), we note that

\[
d(\theta) = \frac{\rho}{2} \| \delta_* \cdot (\theta - \theta_*) \|_1 + \frac{\rho}{2} \| \delta^c_* \cdot \theta \|_1 - \rho \| \delta_* \cdot \theta \|_1 - \rho \| \delta^c_* \cdot \theta \|_1 + \rho \| \theta_* \|_1.
\]

If \(\| \delta^c_* \cdot (\theta - \theta_*) \|_1 > 7 \| \delta_* \cdot (\theta - \theta_*) \|_1\), then

\[
d(\theta) \leq -\frac{\rho}{4} \| \delta^c_* \cdot (\theta - \theta_*) \|_1 - \frac{7\rho}{4} \| \delta_* \cdot (\theta - \theta_*) \|_1 + \frac{3\rho}{2} \| \delta_* \cdot (\theta - \theta_*) \|_1 \leq -\frac{\rho}{4} \| \theta - \theta_* \|_1.
\]

This bound together with (62) shows that the claim (68) holds true when \(\| \delta^c_* \cdot (\theta - \theta_*) \|_1 > 7 \| \delta_* \cdot (\theta - \theta_*) \|_1\). Now, if \(\| \delta^c_* \cdot (\theta - \theta_*) \|_1 \leq 7 \| \delta_* \cdot (\theta - \theta_*) \|_1\), then again by (62), (68) is upper bounded by

\[
-\frac{\rho}{2} \| \delta^c_* \cdot (\theta - \theta_*) \|_1 + \frac{3\rho}{2} \| \delta_* \cdot (\theta - \theta_*) \|_1 - \frac{1}{2} r_0(\| \theta - \theta_* \|_2) \leq -\frac{\rho}{2} \| \theta - \theta_* \|_1 - \frac{1}{2} \left[ r_0(\| \theta - \theta_* \|_2) - 4\rho s_*^{1/2} \| \theta - \theta_* \|_2 \right] \leq -\frac{\rho}{2} \| \theta - \theta_* \|_1 + \frac{a_0}{2},
\]

which also gives (68). We can then use (68) to deduce that

\[
\int_{\mathbb{R}^p} e^{d(\theta)} \mathbb{E}_\theta \left[ e^{\left( \frac{1}{2} \left( \nabla^2 \varphi(\theta), \delta - \varphi \right) + L_{\gamma}(\delta, \theta) \right)} \mathbf{1}_{\mathcal{E}_\theta}(Z) \right] \mu_\delta(d\theta) \leq e^{a_0/2} \int_{\mathbb{R}^p} e^{-\frac{\rho}{2} \| \theta - \theta_* \|_1} \mu_\delta(d\theta) \leq e^{a_0/2} \left( \frac{8}{\rho} \right)^{\| \theta \|_0}, \quad (69)
\]

and using this in (67) we conclude that

\[
\mathbb{E}_\theta \left[ \overline{\Pi}_{\gamma}(A_\theta^c | Z) \mathbf{1}_{\mathcal{E}_\theta}(Z) \right] \leq e^{a_0/2} \left( 1 + \frac{\bar{\kappa}(s_*)}{\rho^2} \right)^{s_*} \sum_{\delta \in A_\theta^c} \frac{\omega_{\delta^c}^2 4^\| \theta \|_0 e^{2\gamma^2 \| \theta \|_0} \sqrt{\det \left( [I_p - \gamma S^c]_{\delta^c} \right)} }{\sqrt{\det( [A_\delta]_{\delta^c} )}}.
\]

We claim that for all \(\delta \in \Delta\),

\[
\sqrt{\det \left( [I_p - s_* - \gamma S]_{\delta^c} \right)} \overline{\det ( [A_\theta]_{\delta^c} )} \leq \exp \left( s_* + \frac{\gamma}{2} \text{Tr}(S - S) + 2\gamma^2 \left( \bar{\kappa}(\| \theta \|_0) \text{Tr}(S) + 4\| S \|_F^2 \right) \right). \quad (70)
\]

To show this, we write

\[
\frac{\sqrt{\det \left( [I_p - s_* - \gamma S]_{\delta^c} \right)} \overline{\det ( [A_\theta]_{\delta^c} )}}{\sqrt{\det ( [A_\delta]_{\delta^c} )}} = \sqrt{\frac{\det ( [I_p - s_* - \gamma S]_{\delta^c} )}{\det ( [I_p]_{\delta^c} - \gamma S^c_{\delta^c} )}} \sqrt{\frac{\det ( [I_p]_{\delta^c} - \gamma S^c_{\delta^c} )}{\det ( [A_\delta]_{\delta^c} )}}.
\]
The first term on the right-hand of the last display can be further written as
\[
\frac{\sqrt{\det (I_p - \gamma S)_{\delta^c}}}{\det (I_p - \gamma S)} \frac{\det (I_p - \gamma S)}{\det (I_p - \gamma S)_{\delta^c}} \leq \left( \frac{4}{3} \right)^{\frac{\gamma \kappa}{2}},
\]
where the inequality follows from Lemma 26 and the fact that all the eigenvalues of the matrix \( I_p - \gamma S \) are between 3/4 and 1. If \( \lambda_j \) (resp. \( \lambda_j \)) denote the eigenvalues of \([\bar{S}]_{\delta^c} \) (resp. \([S]_{\delta^c} \)), we have
\[
\frac{\sqrt{\det (I_p - \|\delta\|_0 - \gamma [S]_{\delta^c})}}{\sqrt{\det ([A]_{\delta^c})}} = \exp \left[ \frac{1}{2} \sum_{j=1}^{p} \left( \log(1 - \gamma \lambda_j) - \log(1 - \gamma (1 + 4 \gamma \kappa(\|\delta\|_0) \bar{\lambda}_j)) \right) \right].
\]
Since \( 4 \gamma \lambda_{\text{max}}(\bar{S}) \leq 1, \gamma \bar{\lambda}_j \leq 1/4, \) and \( \gamma (1 + 4 \gamma \kappa(\bar{s}) \bar{\lambda}_j) \leq 1/2 \) for all \( 1 \leq j \leq p - \|\delta\|_0 \).
Furthermore, the function \( \log \) satisfies \( \log(1 - x) \leq -x, \) and \( \log(1 - x) \geq -x - 4x^2 \) for \( x \in [0, 1/2] \). We deduce that
\[
\frac{\sqrt{\det (I_p - \|\delta\|_0 - \gamma [S]_{\delta^c})}}{\sqrt{\det ([A]_{\delta^c})}} \leq e^{2 \text{Tr}(\bar{S} - \bar{S}) + 2 \gamma^2 (\kappa(\|\delta\|_0) \text{Tr}(\bar{S}) + 4\|\bar{S}\|^2)}, \quad (71)
\]
These last two results together establishes (70). On the other hand, using H3 we have
\[
\sum_{\delta \in \mathcal{A}_1} \sum_{\|\delta\|_0 p \rho^2} e^{3\gamma \|\delta\|_0 \rho^2} = \sum_{j=s^* + 1}^{p} \binom{p}{j} \left( \frac{q}{1 - q} \right)^{j - s^*} (4e^{3\gamma \rho^2})^j \\
\leq \left( \binom{p}{s^*} (4e^{3\gamma \rho^2})^{s^*} \sum_{j=s^* + 1}^{p} \left( \frac{8e^{3\gamma \rho^2}}{p^u} \right)^{j - s^*},
\]
using the fact that \( \frac{q}{1 - q} = \frac{p^{r+1}}{p^{r+1}} \leq \frac{2}{p^{r+1}} \) for \( p \geq 2, \) and \( \binom{p}{j} \leq \binom{p}{s^*} (p/s^*)^j \). Hence for \( p \) large enough so that \( 8e^{3\gamma \rho^2} \leq 1/2, \) we get
\[
\sum_{\delta \in \mathcal{A}_1} \sum_{\|\delta\|_0 p \rho^2} e^{3\gamma \|\delta\|_0 \rho^2} \leq \left( \binom{p}{s^*} (4e^{3\gamma \rho^2})^{s^*} \left( \frac{8e^{3\gamma \rho^2}}{p^u} \right)^{\eta} \leq e^{\frac{3}{2} s^* \log(p) - \frac{u}{2} \log(p)},
\]
for all \( p \) large enough, where here we use again the assumption that \( \gamma \rho^2 = o(\log(p)), \) and \( \log(\binom{p}{s^*}) \leq s^* \log(p) \). Hence we conclude that
\[
\mathbb{E}_p \left[ \Pi_{\gamma} (\tilde{\mathcal{A}}_1 | Z) \mathbf{1}_{\mathcal{E}_0} (Z) \right] \leq e^{2 \eta} \left( 1 + \frac{\kappa(s^*)}{\rho^2} \right)^{s^*} e^{2s^* \log(p) - \frac{u}{2} \log(p)} \exp \left( \frac{\gamma}{2} \text{Tr}(\bar{S} - \bar{S}) + 2\gamma^2 (\lambda_{\text{max}}(\bar{S}) \text{Tr}(\bar{S}) + 4\|\bar{S}\|^2) \right) \leq \frac{1}{p^{ma}},
\]
by choosing \( \eta \) as in the statement of the lemma. Hence the result. \( \square \)

For linear regression models the previous lemma takes a slightly sharper form.
**Lemma 22.** In the particular case of the linear regression model, for all $\rho > 0$, all $\gamma > 0$ such that $4\gamma\sigma^2\lambda_{\max}(X'X) \leq 1$, and all $z \in \mathbb{R}^n$, $\tilde{\Pi}_\gamma(\cdot | z)$ is well-defined.

**Proof.** Here $\tilde{S} = (1/\sigma^2)X'X$, and as above, we have

$$(2\pi\gamma)^{-\frac{p}{2}} \tilde{C}_\gamma(z) \leq \sum_{\delta \in \Delta} \omega_\delta \left( \frac{p}{2} \right) \frac{||\delta||_0}{(2\pi\gamma)^{\frac{p-||\delta||_0}{2}}} \int_{\mathbb{R}^p} e^{\ell(\theta_\delta; z)} e^{-\rho ||\theta_\delta||_1} e^{-\frac{1}{2\gamma}(\theta_\delta - \theta_0)\gamma S(\theta_\delta - \theta_0) + R_\gamma(\delta, \theta_\delta; z)} d\theta.$$

With a similar argument we bound

$$R_\gamma(\delta, \theta; z) \leq \frac{3}{2} \gamma \rho^2 ||\delta||_0 + \frac{3\gamma}{\sigma^2} \lambda_{\max}(X'X) \frac{1}{2\sigma^2} ||z - X\theta_\delta||_2^2$$

$$+ \frac{3\gamma}{2\sigma^2} \lambda_{\max}(X'X) \frac{1}{\sigma^2} (\theta_\delta - \theta_0)'(X'X)(\theta_\delta - \theta_0).$$

Therefore, if $\frac{4\gamma}{\sigma^2} \lambda_{\max}(X'X) \leq 1$, then

$$\ell(\theta_\delta; z) - \frac{1}{2\gamma}(\theta_\delta - \theta_0)'(I_p - \gamma S)(\theta_\delta - \theta_0) + R_\gamma(\delta, \theta_\delta; z) \leq \frac{3}{2} \gamma \rho^2 ||\delta||_0 + \frac{1}{4} \ell(\theta_\delta; z) - \frac{1}{2\gamma}(\theta_\delta - \theta_0)'M_\delta(\theta_\delta - \theta_0),$$

where $M_\delta = (I_p - \gamma(1 + \frac{3\gamma}{\sigma^2} \lambda_{\max}(X'X)))\tilde{S}$. And since $\ell(\theta_\delta; z) = -\frac{1}{2\sigma^2} ||z - X\theta_\delta||_2^2 \leq 0$, we get

$$(2\pi\gamma)^{-\frac{p}{2}} \tilde{C}_\gamma(z) \leq \sum_{\delta \in \Delta} \omega_\delta \left( \frac{p}{2} \right) \frac{||\delta||_0}{(2\pi\gamma)^{\frac{p-||\delta||_0}{2}}} \frac{3\gamma \rho^2}{2\sigma^2} \int_{\mathbb{R}^p} e^{-\rho ||\theta_\delta||_1} e^{-\frac{1}{2\gamma}(\theta_\delta - \theta_0)'M_\delta(\theta_\delta - \theta_0)} d\theta.$$

All the integrals on the right hand side of the last display are finite since $M_\delta$ is positive definite. This proves the lemma. \qed

The proof of Theorem 5 is based on some classical testing arguments for which we need the following result. This lemma slightly extends Lemma 14 of [Atchade (2017)](http://example.com).

**Lemma 23.** Assume $H_2$ and $H_4$. Define

$$\bar{c} \equiv \inf \left\{ x > 0 : r_1(z) \geq 3\bar{\rho}(s_* + \bar{s})^{1/2}z, \text{ for all } z \geq x \right\}.$$

If $\bar{c} < \infty$, then for any $M > 2$, there exists a measurable function $\phi : \mathcal{Z} \to [0, 1]$ such that

$$\mathbb{E}_* (\phi(Z)) \leq \left( \frac{p}{\bar{s}} \right) 9^8 \sum_{j \geq 1} e^{-\frac{1}{\bar{c}}r_1(jM/\bar{s})}.$$
Furthermore, for all $\delta \in \Delta$ such that $\|\delta\|_0 \leq \bar{s}$, and all $\theta \in \mathbb{R}_\delta^p$ such that $\|\theta - \theta_*\|_2 > jM\bar{\epsilon}$ for some $j \geq 1$, we have

$$
\int_{\mathcal{E}_\delta} (1 - \phi(z)) e^{2\gamma|\delta \cdot \nabla \ell(\theta,z) - \delta \cdot \nabla \ell(\theta_*,z)|^2} f_\theta(z)dz \leq e^{-\frac{1}{4} \gamma (\frac{\delta}{M\|\theta\|_2})}.
$$

**Proof.** If $q_1, q_2$ are two integrable functions on some arbitrary measure space, we define their Hellinger affinity as

$$
\mathcal{H}(q_1, q_2) \overset{\text{def}}{=} \int \sqrt{q_1 q_2}.
$$

We will rely on the following result due to Kleijn and van der Vaart (2006).

**Lemma 24** (Kleijn-Van der Vaart (2006)). Let $p$ a density, $Q$ a family of integrable functions. Then there exists a measurable $[0, 1]$-valued function $\phi$ such that

$$
\sup_{q \in Q} \left[ \int \phi q + \int (1 - \phi) q \right] \leq \sup_{q \in \text{conv}(Q)} \mathcal{H}(p, q),
$$

where $\text{conv}(Q)$ is the convex hull of $Q$.

Fix $\delta \in \Delta$ such that $\|\delta\|_0 \leq \bar{s}$, and $\theta \in \mathbb{R}_\delta^p$. To put ourselves in the setting on Lemma 24, we define

$$
\bar{q}_{\delta,u}(z) \overset{\text{def}}{=} e^{2\gamma|\delta \cdot \nabla \ell(u,z) - \delta \cdot \nabla \ell(\theta_*,z)|^2} f_{\theta}(z) 1_{\mathcal{E}_\delta}(z), \quad u \in \mathbb{R}_\delta^p, \quad z \in Z.
$$

For $z \in \mathcal{E}_\delta$, $u \in \mathbb{R}_\delta^p$, we have

$$
\frac{\bar{q}_{\delta,u}(z)}{f_{\theta}(z)} = e^{(\nabla \ell(\theta_*,z) - \nabla \ell(\theta,z)) u - \theta_*) + \mathcal{L}_\gamma(u,z) 1_{\mathcal{E}_\delta}(z) \leq e^\frac{\bar{\ell}}{2} |u - \theta_*)| + \mathcal{L}_\gamma(\delta, u,z). \quad (72)
$$

Therefore

$$
\int_Z \bar{q}_{\delta,u}(z)dz \leq e^\frac{\bar{\ell}}{2} \bar{\epsilon} \mathbb{E}_* \left[ e^{\mathcal{L}_\gamma(\delta,u,z)} 1_{\mathcal{E}_\delta}(Z) \right] < \infty,
$$

by (4). Now, fix $\eta \geq 2\bar{\epsilon}$, and suppose that $\|\theta - \theta_*\|_2 > \eta$. Let

$$
\mathcal{P}_{\delta,\theta} \overset{\text{def}}{=} \left\{ \bar{q}_{\delta,u} : u \in \mathbb{R}_\delta^p, \|u - \theta\|_2 \leq \frac{\eta}{2} \right\}.
$$

According to Lemma 24, applied with $p = f_{\theta_*}$, and $Q = \mathcal{P}_{\delta,\theta}$, there exists a test function $\phi_{\delta,\theta}$ such that

$$
\sup_{q \in \mathcal{P}_{\delta,\theta}} \left[ \int \phi_{\delta,\theta} f_{\theta_*} + \int (1 - \phi_{\delta,\theta}) q \right] \leq \sup_{q \in \text{conv}(\mathcal{P}_{\delta,\theta})} \mathcal{H}(f_{\theta_*}, q). \quad (73)
$$

Any $q \in \text{conv}(\mathcal{P}_{\delta,\theta})$ can be written as $q = \sum_{j} \alpha_j \bar{q}_{\delta,u_j}$, where $\sum_{j} \alpha_j = 1$, $u_j \in \mathbb{R}_\delta^p$, $\|u_j - \theta\|_2 \leq \eta/2$. Notice that this implies that $\|u_j - \theta_*\|_2 > \eta/2 > \bar{\epsilon}$. Therefore, by
Jensen’s inequality and (72) we get

$$\mathcal{H}(q, f_{\theta_*}) = \int_{\mathcal{E}_p} \left( \sum_j \alpha_j \frac{\tilde{q}_{\delta,u_j}(z)}{f_{\theta_*}(z)} f_{\theta_*}(z) dz \right)$$

$$\leq \left( \sum_j \alpha_j \int_{\mathcal{E}_p} \frac{\tilde{q}_{\delta,u_j}(z)}{f_{\theta_*}(z)} f_{\theta_*}(z) dz \right)$$

$$\leq \sqrt{\sum_j \alpha_j e^{\frac{2}{\tilde{\rho}}} \|u_j - \theta_*\|_2 \|1_{\mathcal{E}_*} e^{\mathbb{L}_s(\delta,u_j;Z)} 1_{\mathcal{E}_p}(Z)\|}$$

$$\leq \sqrt{\sum_j \alpha_j e^{\frac{2}{\tilde{\rho}}} \|u_j - \theta_*\|_2 \|1 - \frac{1}{2} r_1(\|u_j - \theta_*\|_2)\|}.$$  

Since \(\|u_j\|_0 \leq \bar{s}, \|u_j - \theta_*\|_2 > \eta/2 \geq \bar{\epsilon}\), and by the definition of \(\bar{\epsilon}\), we have

$$\frac{\tilde{\rho}}{2} \|u_j - \theta_*\|_1 - \frac{1}{6} r_1(\|u_j - \theta_*\|_2) \leq \frac{\tilde{\rho}(\bar{s} + s_*)^{1/2}}{2} \|u_j - \theta_*\|_2 - \frac{1}{6} r_1(\|u_j - \theta_*\|_2) \leq 0.$$  

Note that if \(\delta \supseteq \delta_*\), then we can replace \(\bar{s} + s_*\) by \(\bar{s}\) in the above. We conclude that for any \(q \in \text{conv}(\mathcal{P}_{\delta,\theta})\),

$$\mathcal{H}(q, f_{\theta_*}) \leq e^{-\frac{1}{2} r_1(\bar{s})}. \quad (74)$$  

Now for \(M > 2\), write \(\bigcup_\delta \{\theta \in \mathbb{R}_p^\delta : \|\theta - \theta_*\|_2 > M \bar{\epsilon}\} \) as \(\bigcup_\delta \bigcup_{j \geq 1} \mathcal{A}_\epsilon(\delta, j)\), where the unions in \(\delta\) are taken over all \(\delta\) such that \(\|\delta\|_0 = \bar{s}\), and

$$\mathcal{A}_\epsilon(\delta, j) \overset{\text{def}}{=} \left\{\theta \in \mathbb{R}_p^\delta : jM \bar{\epsilon} < \|\theta - \theta_*\|_2 \leq (j + 1)M \bar{\epsilon}\right\}.$$  

For \(\mathcal{A}_\epsilon(\delta, j) \neq \emptyset\), let \(\mathcal{S}(\delta, j)\) be a maximally \((jM \bar{\epsilon}/2)\)-separated point in \(\mathcal{A}_\epsilon(\delta, j)\). It is easily checked that the cardinality of \(\mathcal{S}(\delta, j)\) is upper bounded by \(9\|\delta\|_0 = 9^\bar{s}\) (see for instance Example 7.1 for the arguments). For \(\theta_{\delta,jk} \in \mathcal{S}(\delta, j)\), let \(\phi_{\delta,jk}\) denote the test function \(\phi_{\delta,jk}\) obtained above with \(\theta = \theta_{\delta,jk}\) and \(\eta = jM \bar{\epsilon}\). From (73) and (74) \(\phi_{\delta,jk}\) satisfies

$$\sup_{u \in \mathbb{R}_p^\delta, \|u - \theta_{\delta,jk}\|_2 \leq \frac{jM \bar{\epsilon}}{2}} \left[ \mathbb{E}_s(\phi_{\delta,jk}(Z)) + \int_{\mathcal{E}_p} (1 - \phi_{\delta,jk}(z)) \tilde{q}_u(z) dz \right] \leq e^{-\frac{1}{2} r_1(\frac{jM \bar{\epsilon}}{2})}. \quad (75)$$  

Then we set

$$\phi = \sup_{\delta : \|\delta\|_0 = \bar{s}} \sup_{j \geq 1} \sup_{\theta_{\delta,jk} \in \mathcal{S}(\delta, j)} \phi_{\delta,jk}.\quad$$

It then follows that

$$\mathbb{E}_s(\phi(Z)) \leq \sum_\delta \sum_{j \geq 1} \sum_{\theta_{\delta,jk} \in \mathcal{S}(\delta, j)} \mathbb{E}_s(\phi_{\delta,jk}(Z)) \leq \left(\frac{\rho}{s}\right) 9^\bar{s} \sum_{j \geq 1} e^{-\frac{1}{2} r_1(\frac{jM \bar{\epsilon}}{2})}.$$
And if for some $\delta$, such that $\|\delta\|_0 \leq \bar{s}$ and $\theta \in \mathbb{R}^p_\delta$ we have $\|\theta - \theta_*\|_2 > jM\epsilon$, then we can find $\bar{\delta}$ with $\|\bar{\delta}\|_0 = \bar{s}$, such that $\theta \in \mathbb{R}^p_{\bar{\delta}}$ and $\theta$ resides within $(iM\epsilon)/2$ of some point $\theta_{\bar{\delta},ik} \in S(\bar{\delta}, i)$ for some $i \geq j$. Hence, by (75),

$$\int_{E_\rho} (1 - \phi(z))\bar{q}_{\delta,\theta}(z)dz \leq \int_{E_\rho} (1 - \phi_{\bar{\delta},ik}(z))\bar{q}_{\delta,\theta}(z)dz \leq e^{-\frac{1}{2}(iM\epsilon)^2}.$$

This ends the proof. □

We make use of the following Gaussian version of the Hanson-Wright inequality. The result follows directly from deviation bounds for Lipschitz function of Gaussian random variables.

**Lemma 25.** If $Z \sim \mathcal{N}(0, I_m)$, and $A \in \mathbb{R}^{m \times m}$ is a symmetric positive semi-definite matrix, then for all $t \geq \text{Tr}(A)$,

$$\mathbb{P}(X'AX > t) \leq \exp \left[ -\frac{(\sqrt{t} - \sqrt{\text{Tr}(A)})^2}{2\|A\|_2} \right].$$

We will also need the following lemma on determinants of sub-matrices.

**Lemma 26.** If symmetric positive definite matrices $A, M$ and $D \in \mathbb{R}^{q \times q}$ are such that $M = \begin{pmatrix} A & B \\ B' & D \end{pmatrix}$, then

$$\det(A)\lambda_{\min}(M)^{q} \leq \det(M) \leq \det(A)\lambda_{\max}(M)^{q}.$$

**Proof.** This follows from Cauchy’s interlacing property for eigenvalues. See for instance [Horn and Johnson (2012)] Theorem 4.3.17. □

A.2. **Proof of Theorem 5**. Let $\Delta_\delta \overset{\text{def}}{=} \{ \delta \in \Delta : \|\delta\|_0 \leq \bar{s} \}$. We have $\Delta \times \mathbb{R}^p = (\Delta \setminus \Delta_\delta) \times \mathbb{R}^p \cup \mathcal{F}_1 \cup \mathcal{F}_{21} \cup \mathcal{F}_{22} \cup \mathcal{B}_{m,M}$, where

$$\mathcal{F}_1 \overset{\text{def}}{=} \bigcup_{\delta \in \Delta_\delta} \{ \delta \} \times \{ \theta \in \mathbb{R}^p : \|\theta_\delta - \theta_*\|_2 > M\epsilon \},$$

$$\mathcal{F}_{21} \overset{\text{def}}{=} \bigcup_{\delta \in \Delta_\delta} \{ \delta \} \times \left\{ \theta \in \mathbb{R}^p : \|\theta_\delta - \theta_*\|_2 \leq M\epsilon, \text{ and } \|\theta - \theta_\delta\|_2 > 2\sqrt{(m+1)\gamma p} \right\},$$

$$\mathcal{F}_{22} \overset{\text{def}}{=} \bigcup_{\delta \in \Delta_\delta} \{ \delta \} \times \left\{ \theta \in \mathbb{R}^p : \|\theta_\delta - \theta_*\|_2 \leq M\epsilon, \|\theta - \theta_\delta\|_2 \leq 2\sqrt{(m+1)\gamma p}, \text{ and } \|\theta - \theta_\delta\|_\infty > 2\sqrt{(m+1)\gamma \log(p)} \right\}.$$

Hence we can write,

$$\Pi_{\gamma}(\mathcal{B}_{m,M}|Z) = 1 - \Pi_{\gamma}(\|\delta\|_0 > \bar{s}|Z) - \Pi_{\gamma}(\mathcal{F}_1|Z) - \Pi_{\gamma}(\mathcal{F}_{21}|Z) - \Pi_{\gamma}(\mathcal{F}_{22}|Z).$$
Therefore, using the assumption that $P_*(Z \in E_\rho) \geq 1/2$, and the definition of condition probability,

\[
E_* [\tilde{\Pi}_\gamma (B_{m,M} | Z) | Z \in E_\rho] \geq 1 - E_* [\tilde{\Pi}_\gamma (||\delta||_0 > \epsilon | Z) | Z \in E_\rho]
- 2E_* [1_{E_\rho}(Z)\tilde{\Pi}_\gamma (F_1 | Z)] - 2E_* [1_{E_\rho}(Z)\tilde{\Pi}_\gamma (F_{21} | Z)] - 2E_* [1_{E_\rho}(Z)\tilde{\Pi}_\gamma (F_{22} | Z)].
\] (76)

Therefore, to finish the proof it suffices to upper-bounding the terms on the right-hand side of (76).

Let $\phi$ denote the test function asserted by Lemma 23 where $M > 2$ is some arbitrary absolute constant. We can then write

\[
E_* [1_{E_\rho}(Z)\tilde{\Pi}_\gamma (F_1 | Z)] \leq E_* (\phi(Z)) + E_* [1_{E_\rho}(Z) (1 - \phi(Z)) \tilde{\Pi}_\gamma (F_1 | Z)].
\] (77)

Since $(p^2)9^8 \leq e^{8 \log(9p)}$, we have

\[
E_* (\phi(Z)) \leq e^{8 \log(9p)} \sum_{j=1}^{\infty} e^{-\frac{1}{4} \gamma (M^2 + j^2)}.
\] (78)

We note that for any $z \in Z$, and any measurable subset $\tilde{B}$ of $\Delta \times \mathbb{R}^p$, we have

\[
\tilde{\Pi}_\gamma (\tilde{B} | z) = \frac{1}{C_{\gamma} (z)} \sum_{\delta \in \Delta} \omega_\delta (2\pi^2)^\frac{||\delta||_0}{2} \left( \frac{\rho}{2} \right)^{||\delta||_0} \int_{\tilde{B}(\delta)} e^{-h_\gamma (\delta, \theta; z)} d\theta,
\]

where $B(\delta) \overset{\text{def}}{=} \{ \theta \in \mathbb{R}^p : (\delta, \theta) \in B \}$. Then using (61) and the second inequality of Lemma 19, we get

\[
\tilde{\Pi}_\gamma (\tilde{B} | z) \leq \sqrt{\det ([I_p - \gamma S]_{\delta_\gamma}^\gamma)} \left( 1 + \frac{K(s_\gamma)}{\rho^2} \right)^{s_\gamma} \frac{1}{\omega_{\delta_\gamma}} \sum_{\delta \in \Delta} \omega_\delta \left( \frac{1}{2\pi^2} \right)^{p-\frac{||\delta||_0}{2}} \left( \frac{\rho}{2} \right)^{||\delta||_1}
\times \int_{\tilde{B}(\delta)} e^{\ell(\theta; z)} e^{-\rho||\theta||_1} e^{-\frac{1}{2\gamma^2} (\theta - \theta_\delta) (I_p - \gamma S)(\theta - \theta_\delta) + R_\gamma (\delta, \theta; z)} d\theta.
\] (79)

Combined with (64) this yields

\[
\tilde{\Pi}_\gamma (\tilde{B} | z) \leq \sqrt{\det ([I_p - \gamma S]_{\delta_\gamma}^\gamma)} \left( 1 + \frac{K(s_\gamma)}{\rho^2} \right)^{s_\gamma} \frac{1}{\omega_{\delta_\gamma}} \sum_{\delta \in \Delta} \omega_\delta \left( \frac{1}{2\pi^2} \right)^{p-\frac{||\delta||_0}{2}} \left( \frac{\rho}{2} \right)^{||\delta||_1}
\times e^{2\gamma \rho^2 ||\delta||_0} \int_{\tilde{B}(\delta)} e^{\ell(\theta; z)} e^{-\rho||\theta||_1} e^{2\gamma \delta \cdot \nabla \ell(\theta; z) - \delta \cdot \nabla \ell(\theta; z)} ||\delta||_1^2 + \frac{1}{2\gamma^2} (\theta - \theta_\delta) A_{\delta} (\theta - \theta_\delta) d\theta,
\] (80)
where $A_\delta = I_p - \gamma(1 + 4\gamma\tilde{\kappa}(\delta))\tilde{S}$. We apply this with $B = \mathcal{F}_1$ and Fubini's theorem to get:

$$
\mathbb{E}_* \left[ 1_{\mathcal{E}_p}(Z)(1 - \phi(Z))\Pi_1(\mathcal{F}_1 | Z) \right] 
\leq \left( 1 + \frac{\tilde{\kappa}(s_*)}{\rho^2} \right)^{s_*} \sum_{\delta \in \Delta_\delta} \frac{\omega_{\delta}}{w_{\delta_*}} \int_{\mathbb{R}^p} e^{-\rho\|\theta\|_1} \int_{\mathbb{R}^p} e^{-\rho\|\theta\|_1} \mu_\delta(d\theta) 
\leq e^{-\frac{1}{6}r_1(\frac{L_{M_*}}{\rho})} \int_{\mathbb{R}^p} e^{-\rho\|\theta\|_1} \mu_\delta(d\theta),
$$

and it is easily seen that $\int_{\mathbb{R}^p} e^{-\rho\|\theta\|_1} \mu_\delta(d\theta) \leq \left( \frac{2}{\rho} \right)^{\|\delta\|_0}. \tag{81}$ Therefore (81) reduces to

$$
\mathbb{E}_* \left[ 1_{\mathcal{E}_p}(Z)(1 - \phi(Z))\Pi_1(\mathcal{F}_1 | Z) \right] \leq \left( 1 + \frac{\tilde{\kappa}(s_*)}{\rho^2} \right)^{s_*} \sum_{j=1}^{\infty} e^{-\frac{1}{6}r_1(\frac{L_{M_*}}{\rho}) + 8\rho^2\|S\|_0 + \frac{1}{2}} \frac{\omega_{\delta}}{w_{\delta_*}} e^{3\gamma\rho^2\|\delta\|_0} \sqrt{\det(I_p - \gamma[S]\delta^c)} \sqrt{\det([A_{\delta}]_{\delta^c})} \tag{82}
$$

Using (70) and borrowing the definition of $a$ in Equation (10) of the main manuscript, we get

$$
\mathbb{E}_* \left[ 1_{\mathcal{E}_p}(Z)(1 - \phi(Z))\Pi_1(\mathcal{F}_1 | Z) \right] \leq e^{s_* + \frac{1}{2}} \left( 1 + \frac{\tilde{\kappa}(s_*)}{\rho^2} \right)^{s_*} \sum_{\delta \in \Delta_\delta} \frac{\omega_{\delta}}{w_{\delta_*}} e^{3\gamma\rho^2\|\delta\|_0} \sum_{j=1}^{\infty} e^{-\frac{1}{6}r_1(\frac{L_{M_*}}{\rho}) + 8\rho^2\|S\|_0 + \frac{1}{2}} \frac{\omega_{\delta}}{w_{\delta_*}} \left[ \sum_{\delta \in \Delta_\delta} \frac{\omega_{\delta}}{w_{\delta_*}} e^{3\gamma\rho^2\|\delta\|_0} \right].
$$

Using (13) we have $4e^{3\gamma\rho^2} \leq p^{u+1}$, for all $p$ large enough, so that Assumption H3 gives

$$
\sum_{\delta \in \Delta_\delta} \frac{\omega_{\delta}}{w_{\delta_*}} e^{3\gamma\rho^2\|\delta\|_0} \leq \left( \frac{1 - q}{q} \right)^{s_*} \sum_{s=0}^{\infty} \left( \frac{2e^{3\gamma\rho^2}}{p^{u+1}} \right)^s \leq 2(1 - q)^{s_*} p^{s_*(1+u)} \leq 2p^{s_*(1+u)}.
$$
It follows that

$$E_* \left[ 1_{E_p}(Z)(1 - \phi(Z))\Pi_\gamma(F_1|Z) \right]$$

$$\leq 2e^{2\epsilon} \exp \left( (2 + u)s_* \log(p) + s_* \log \left( 1 + \frac{\bar{k}(s_*)}{p^2} \right) \right)$$

$$\times \sum_{j=1}^{\infty} e^{-\frac{j}{2} \epsilon_1(M\epsilon) + 8\delta \bar{\delta}^{1/2}(M\epsilon)}. \quad (83)$$

We set $F_{21}^{(\delta)} \overset{\text{def}}{=} \{ \theta \in \mathbb{R}^p : \|\theta - \theta_*\|_2 \leq \epsilon_1 \}$, with $\epsilon_1 = 2\sqrt{(m + 1)\gamma p}$. From the definition of $\Pi_\gamma$, and using the two inequalities of Lemma 19, we have

$$\Pi_\gamma(F_{21}|Z) = \frac{\sum_{\delta \in \Delta_{\vec{S}}} \omega_\delta \left( \frac{\delta}{2} \right) \|\delta\|_0 \int_{F_{21}^{(\delta)}} e^{-h_\gamma(\delta, u; Z)} du}{\sum_{\delta \in \Delta} \omega_\delta \left( \frac{\delta}{2} \right) \|\delta\|_0 \int_{\mathbb{R}^p} e^{-h_\gamma(\delta, u; Z)} du} \leq \frac{\sum_{\delta \in \Delta_{\vec{S}}} \omega_\delta \left( \frac{\delta}{2} \right) \|\delta\|_0 \int_{F_{21}^{(\delta)}} e^{\ell(u;Z) - \|u\|_1 - \frac{1}{2\gamma} (u - u_\delta)'(I_p - \gamma \bar{S})(u - u_\delta) + R_\gamma(\delta; u; Z)} du}{\sum_{\delta \in \Delta} \omega_\delta \left( \frac{\delta}{2} \right) \|\delta\|_0 \int_{\mathbb{R}^p} e^{\ell(u;Z) - \|u\|_1 - \frac{1}{2\gamma} (u - u_\delta)'(I_p - \gamma \bar{S})(u - u_\delta) du}$$

For $Z \in E_p$, and $\delta \in \Delta_{\vec{S}}$, proceeding as in (64) we have

$$- \frac{1}{2\gamma} (u - u_\delta)'A(u - u_\delta) + R_\gamma(\delta; u; Z) \leq - \frac{1}{2\gamma} (u - u_\delta)'A(u - u_\delta) + 3\gamma \rho^2 \bar{s} + 2\gamma \bar{k}(\bar{s})^2 \|u_\delta - \theta_*\|_2^2. \quad (84)$$

where $A = I_p - \gamma(1 + 4\gamma \bar{k}(\bar{s}))\bar{S}$. It follows that

$$1_{E_p}(Z)\Pi_\gamma(F_{21}|Z) \leq 1_{E_p}(Z)e^{3\gamma \rho^2 \bar{s} + 2\gamma \bar{k}(\bar{s})^2(M\epsilon)^2}$$

$$\times \frac{\sum_{\delta \in \Delta_{\vec{S}}} \omega_\delta \left( \frac{\delta}{2} \right) \|\delta\|_0 \left[ \int_{\mathbb{R}^p} e^{\ell(u;Z) - \|u\|_1} \mu_\delta(du) \right]}{\sqrt{\text{det}(\|A\|_{\vec{S}})}} \frac{1}{\sqrt{\text{det}(I_p - \|\delta\|_0 - \gamma \bar{S}_{\vec{S}})}}. \quad (85)$$

where

$$T_\delta \overset{\text{def}}{=} \frac{\int_{F_{21}^{(\delta)}} e^{-\frac{1}{2\gamma} z'(A\|_{\vec{S}})z} dz}{\int_{\mathbb{R}^p - \|\delta\|_1} e^{-\frac{1}{2\gamma} z'(A\|_{\vec{S}})z} dz},$$

and $F_{21}^{(\delta)} \overset{\text{def}}{=} \{ \theta \in \mathbb{R}^p : \|\theta\|_2 \geq \epsilon_1 \}$. We have seen in (71) that

$$\sqrt{\frac{\text{det}(I_p - \|\delta\|_0 - \gamma \bar{S}_{\vec{S}})}{\text{det}(\|A\|_{\vec{S}})}} \leq e^{\frac{\epsilon_1}{2}}.$$
Therefore, $(85)$ is upper bounded by

$$e^{3\gamma p^2 s + 2\gamma \kappa(s)^2 (M\varepsilon)^2} e^{\frac{a}{2}} \sup_{\delta \in \Delta_2} T_{\delta}.$$  

$T_{\delta}$ is the probability of the set $\tilde{F}_{2}^{(\delta)} \subset \mathbb{R}^{p-\|\delta\|}$ under the Gaussian distribution $\mathcal{N}(0, \gamma ([A]_{\delta^c})^{-1})$. Under the assumption $4\gamma \lambda_{\max}(\tilde{S}) \leq 1$, all the eigenvalues of the matrix $A = I_p - \gamma (1 + 4\gamma \kappa(s)) \tilde{S}$ are between $[1/2, 1]$, and so are the eigenvalues of the sub-matrix $[A]_{\delta^c}$ (by Cauchy’s interlacing property for eigenvalues; see Theorem 4.3.17 of [Horn and Johnson (2012)]). Hence by Lemma 25 we have $T_{\delta} \leq e^{-\frac{m}{4}}$, for $m \geq 1$. Hence,

$$1_{F_{p}}(Z) \Pi_{\gamma}(F_{21}|Z) \leq e^{-\frac{m}{4}} \exp \left(3\gamma p^2 s + 2\gamma \kappa(s)^2 (M\varepsilon)^2 + \frac{a}{2} \right).$$

A similar bound holds for $F_{22}$ with $\frac{1}{p^2}$ instead of $e^{-mp/2}$. This completes the proof.

### Appendix B. Proof of Theorem 8

**Proof.** We split the proof into two parts.

**Part one: Model selection consistency.** To shorten notation, we write $B^{(\delta)}$ for $B_{m,M}^{(\delta)}$, and $B$ for $B_{m,M}$. We have seen that under (14) $B = \bigcup_{\delta \in A} \{\delta\} \times B^{(\delta)}$, which we can obviously write as $\overline{B} = (\{\delta^*_s\} \times B^{(\delta^*_s)}) \cup \bigcup_{\delta \in A_0} \{\delta\} \times B^{(\delta)}$, where $A_0 \text{ def } A \backslash \{\delta^*_s\}$. Hence

$$\Pi_{\gamma} \left( \{\delta^*_s\} \times B^{(\delta^*_s)}|z \right) = \Pi_{\gamma} (B|z) - \Pi_{\gamma} \left( \bigcup_{\delta \in A_0} \{\delta\} \times B^{(\delta)}|z \right).$$

The proof then boils down to controlling the rightmost term in the last equation. We have by definition

\[
\left. \Pi_{\gamma} \left( \bigcup_{\delta \in A_0} \{\delta\} \times B^{(\delta)}|z \right) \right. 
\[
= \frac{\sum_{\delta \in A_0} \omega_\delta \left( \frac{\rho}{2} \right) \|\delta\| \left( \frac{1}{2\pi\gamma} \right) \frac{p-\|\delta\|_1}{2} \int_{\mathbb{R}^p} e^{-h_{\gamma}(\delta; \theta; z)} d\theta}{\sum_{\delta \in \Delta} \omega_\delta \left( \frac{\rho}{2} \right) \|\delta\| \left( \frac{1}{2\pi\gamma} \right) \frac{p-\|\delta\|_1}{2} \int_{\mathbb{R}^p} e^{-h_{\gamma}(\delta; \theta; z)} d\theta} 
\leq \frac{\sum_{\delta \in A_0} \omega_\delta \left( \frac{\rho}{2} \right) \|\delta\| \left( \frac{1}{2\pi\gamma} \right) \frac{p-\|\delta\|_1}{2} \int_{\mathbb{R}^p; \|\theta_\delta - \theta^*_s\|_2 \leq M\varepsilon} e^{-h_{\gamma}(\delta; \theta; z)} d\theta}{\sum_{\delta \in \Delta} \omega_\delta \left( \frac{\rho}{2} \right) \|\delta\| \left( \frac{1}{2\pi\gamma} \right) \frac{p-\|\delta\|_1}{2} \int_{\mathbb{R}^p; \|\theta_\delta - \theta^*_s\|_2 \leq M\varepsilon} e^{-h_{\gamma}(\delta; \theta; z)} d\theta}. \tag{86}
\]

The first inequality of Lemma 19 says that for all $\theta \in \mathbb{R}^p$,

$$-h_{\gamma}(\delta; \theta; z) \geq \ell(\theta_\delta; z) - \rho \|\theta_\delta\|_1 - \frac{1}{2\gamma} (\theta - \theta^*_s)'(I_p - \gamma S)(\theta - \theta^*_s).$$
Since $\ell(\theta; z) = \ell^{[\delta]}(\theta; z)$, we combine this with the definition of $\varpi_{2,M}$ in (13) to obtain that for all $\theta \in \mathbb{R}^p$, such that $\|\theta - \theta_*\|_2 \leq M\epsilon$,

$$-h_\gamma(\delta, \theta; z) \geq \ell^{[\delta]}(\hat{\theta}_\delta; z) + \left\langle \nabla \ell^{[\delta]}(\hat{\theta}_\delta; z), [\theta]_\delta - \hat{\theta}_\delta \right\rangle - \frac{1}{2}([\theta]_\delta - \hat{\theta}_\delta)' [-\nabla^{(2)} \ell^{[\delta]}(\hat{\theta}_\delta; z)] ([\theta]_\delta - \hat{\theta}_\delta) - \frac{\varpi_{2,M}}{6} \|\theta - \hat{\theta}_\delta\|_2^3 - \rho \|\theta\|_1 - \frac{1}{2\gamma}(\theta - \theta_\delta)' (I_p - \gamma S)(\theta - \theta_\delta).$$

By the first order optimality condition of $\hat{\theta}_\delta$, $\nabla \ell^{[\delta]}(\hat{\theta}_\delta; z) = 0$. Furthermore for any $\theta \in \mathbb{R}^p$ such that $\|\theta - \theta_*\|_2 \leq M\epsilon$, and using the assumption that $\|\theta - [\theta_\delta]_\delta\|_2 \leq \epsilon$, we have

$$-\rho \|\theta\|_1 - \frac{\varpi_{2,M}}{6} \|\theta - \hat{\theta}_\delta\|_2^3 \geq -\rho \|\hat{\theta}_\delta\|_1 - (M + 1)^2 \rho \epsilon^2 - \frac{(M + 1)^3}{6} \varpi_{2,M} \epsilon^3.$$

We then deduce that

$$-h_\gamma(\delta, \theta; z) \geq -(M + 1)^2 \rho \epsilon^2 - \frac{(M + 1)^3}{6} \varpi_{2,M} \epsilon^3 + \ell^{[\delta]}(\hat{\theta}_\delta; z) - \rho \|\hat{\theta}_\delta\|_1 - \frac{1}{2}([\theta]_\delta - \hat{\theta}_\delta)' [-\nabla^{(2)} \ell^{[\delta]}(\hat{\theta}_\delta; z)] ([\theta]_\delta - \hat{\theta}_\delta) - \frac{1}{2\gamma}(\theta - \theta_\delta)' (I_p - \gamma S)(\theta - \theta_\delta). \quad (87)$$

It follows from this last inequality that the denominator of the right-hand side of (86) is lower bounded by

$$e^{-(M + 1)^2 \rho \epsilon^2 - \frac{(M + 1)^3}{6} \varpi_{2,M} \epsilon^3 - \frac{1}{2}([\theta]_\delta - \hat{\theta}_\delta)' [-\nabla^{(2)} \ell^{[\delta]}(\hat{\theta}_\delta; z)] ([\theta]_\delta - \hat{\theta}_\delta) - \frac{1}{2\gamma}(\theta - \theta_\delta)' (I_p - \gamma S)(\theta - \theta_\delta) + R_\gamma(\delta, \theta; z),$$

where $B_\delta \overset{\text{def}}{=} \{ u \in \mathbb{R}^{\|\delta\|_0} : \|u - [\theta_\delta]_\delta\|_2 \leq M\epsilon \}$. Starting with the second inequality of Lemma 19, and with the same calculations as above, we get for any $\theta \in B^{(\delta)}$,

$$-h_\gamma(\delta, \theta; z) \leq -(M + 1)^2 \rho \epsilon^2 - \frac{(M + 1)^3}{6} \varpi_{2,M} \epsilon^3 + \ell^{[\delta]}(\hat{\theta}_\delta; z) - \rho \|\hat{\theta}_\delta\|_1 - \frac{1}{2}([\theta]_\delta - \hat{\theta}_\delta)' [-\nabla^{(2)} \ell^{[\delta]}(\hat{\theta}_\delta; z)] ([\theta]_\delta - \hat{\theta}_\delta) - \frac{1}{2\gamma}(\theta - \theta_\delta)' (I_p - \gamma S)(\theta - \theta_\delta) + R_\gamma(\delta, \theta; z),$$

and for $z \in E_{\rho}$, $\delta \in A$, and $\theta \in B^{(\delta)}$, using (64), we have

$$-\frac{1}{2\gamma}(\theta - \theta_\delta)' (I_p - \gamma S)(\theta - \theta_\delta) \leq -\frac{1}{2\gamma}(\theta - \theta_\delta)' [A]_{\delta^c}(\theta - \theta_\delta) + 2\gamma \kappa(\delta)^2 (M\epsilon)^2 + 3\gamma \rho^2 \|\delta\|_0.$$
The last two inequalities imply that for $z \in \mathcal{E}_\rho$, $\delta \in \mathcal{A}$, and $\theta \in \mathcal{B}^{(\delta)}$
\[ -h_\gamma(\delta; z) \leq (M + 1)\rho s^{3/2} + \frac{(M + 1)^3}{6} \omega_{2M} \epsilon^3 + 2\gamma \tilde{k}(s)^2(M\epsilon)^2 + 3\gamma \rho^2 \|\delta\|_0 \\
+ \ell^{[\delta]}(\hat{\theta}_\delta; z) - \rho \|\hat{\theta}_\delta\|_1 - \frac{1}{2} ([\theta]_\delta - \hat{\theta}_\delta)'[-\nabla(2) \ell(\hat{\theta}_\delta; z)]([\theta]_\delta - \hat{\theta}_\delta) \\
- \frac{1}{2\gamma}(\theta - \hat{\theta}_\delta)'[A]_{\delta^c}(\theta - \hat{\theta}_\delta). \quad (88) \]

Therefore the numerator on the right-hand side of (86) is upper-bounded by
\[ e^{(M + 1)\rho s^{3/2} + \frac{(M + 1)^3}{6} \omega_{2M} \epsilon^3 + 2\gamma \tilde{k}(s)^2(M\epsilon)^2 + 3\gamma \rho^2 \|\delta\|_0} \sum_{\delta \in \mathcal{A}} \left( \frac{1}{2\pi \gamma} \right)^{\rho - \|\delta\|_0} \int_{\mathbb{R}^{p-\|\delta\|_0}} e^{-\frac{1}{2\gamma} u'(0\|\delta^c\|_0)} du \times e^{3\gamma \rho^2 \|\delta\|_0} \omega_{\delta} \left( \frac{\rho}{2} \right)^{\|\delta\|_0} e^{\ell(\hat{\theta}_\delta; z) - \rho \|\hat{\theta}_\delta\|_1} \sqrt{\det(2\pi I_{\gamma,\delta})N(\hat{\theta}_\delta, I_{\gamma,\delta}^{-1})(B_\delta)}, \]

Furthermore, we have seen in (71) that
\[ \sqrt{\det(I_{p-\|\delta\|_0} - \gamma |S|_{\delta^c})} \leq e^{\frac{s}{2}}. \]

Therefore it follows from the above that (86) gives us
\[ \Pi_\gamma \left( \bigcup_{\delta \in \mathcal{A}_0} \{\delta\} \times \mathcal{B}^{(\delta)} \mid z \right) \leq e^{c_0} \sum_{\delta \in \mathcal{A}_0} \omega_{\delta} \left( \frac{\rho}{2} \right)^{\|\delta\|_0} e^{\ell(\hat{\theta}_\delta; z) - \rho \|\hat{\theta}_\delta\|_1} \sqrt{\det(2\pi I_{\gamma,\delta}^{-1})N(\hat{\theta}_\delta, I_{\gamma,\delta}^{-1})(B_\delta)} \times \sum_{\delta \in \mathcal{A}} \omega_{\delta} \left( \frac{\rho}{2} \right)^{s + k} e^{\ell(\hat{\theta}_\delta; z) - \rho \|\hat{\theta}_\delta\|_1} \sqrt{\det(I_{p-s-k} - \gamma |S|_{\delta^c})} \sqrt{\det(2\pi I_{\gamma,\delta}^{-1})N(\hat{\theta}_\delta, I_{\gamma,\delta}^{-1})(B_\delta)}}, \]

where $c_0 \overset{\text{def}}{=} 2(M + 1)\rho s^{3/2} + \frac{(M + 1)^3}{6} \omega_{2M} \epsilon^3 + 2\gamma \tilde{k}(s)^2(M\epsilon)^2 + 3\gamma \rho^2 s + \frac{a}{2}$. We rewrite the last inequality as
\[ \Pi_\gamma \left( \bigcup_{\delta \in \mathcal{A}_0} \{\delta\} \times \mathcal{B}^{(\delta)} \mid z \right) \leq e^{c_0} \sum_{k=0}^{s \cdot s} G_k \sum_{k=1}^{s \cdot s} G_k, \quad (89) \]

where
\[ G_k = \sum_{\delta \in \mathcal{A}, \|\delta\|_0 = s, s + k} \omega_{\delta} \left( \frac{\rho}{2} \right)^{s + k} e^{\ell(\hat{\theta}_\delta; z) - \rho \|\hat{\theta}_\delta\|_1} \sqrt{\det(I_{p-s-k} - \gamma |S|_{\delta^c})} \sqrt{\det(2\pi I_{\gamma,\delta}^{-1})N(\hat{\theta}_\delta, I_{\gamma,\delta}^{-1})(B_\delta)}. \]

We note that
\[ \sqrt{\det(2\pi I_{\gamma,\delta}^{-1})} = \frac{(2\pi)^{s+k}}{\sqrt{\det \left( [-\nabla(2) \ell(\hat{\theta}_\delta; z)] \right)}}. \]
Hence

$$G_k = \sum_{\delta \geq \delta^*, \|\delta\|_0 = s_* + k} \frac{\omega_\delta \left(\frac{\rho}{2}\right)^{s_* + k}}{\sqrt{\det (I_{p-s_*}-\gamma|S|_{\delta^*})}} \frac{e^{\ell(\hat{\theta}_\delta; z) - \rho \|\hat{\theta}_\delta\|_1}}{\sqrt{\det \left(\left[-\nabla(2)\ell(\delta) (\hat{\theta}_\delta; z)\right]\right)}} \frac{(2\pi)^{s_* + k} N(\hat{\theta}_\delta, I_{\gamma, \delta}^{-1}) (B_\delta)}{\sqrt{\det \left(\left[-\nabla(2)\ell(\delta) (\hat{\theta}_\delta; z)\right]\right)}}.$$  

Fix $\delta$ such $\delta \geq \delta^*$, $\|\delta\|_0 = s_* + k$. Firstly, since $|S|_\delta$ is a sub-matrix of $|S|_{\delta^*}$ and the eigenvalues of $I_{p-s_*} - \gamma|S|_{\delta^*}$ are all between $1/2$ and $1$, it is not hard to see that

$$\frac{\sqrt{\det (I_{p-s_*} - \gamma|S|_{\delta^*})}}{\sqrt{\det (I_{p-s_*} - \gamma|S|_{\delta^*})}} \leq 1.$$  

Secondly,

$$-\rho\|\hat{\theta}_\delta\|_1 + \rho\|\hat{\theta}_*\|_1 \leq \rho\|\hat{\theta}_\delta - \hat{\theta}_*\|_1 \leq 2\rho^{3/2} e,$$

and for $z \in \mathcal{E}_{\rho, \Lambda}$, we have

$$\ell(\hat{\theta}_\delta; z) \leq \ell(\hat{\theta}_*; z) + \Lambda k.$$  

It follows that

$$G_k \leq G_0 e^{2\rho^{3/2} e}$$

$$\times \sum_{\delta \geq \delta^*, \|\delta\|_0 = s_* + k} \frac{\omega_\delta e^{\Lambda k} \left(\frac{\rho}{2}\right)^k (2\pi)^{\frac{1}{2}}}{\omega_\delta} \frac{\sqrt{\det \left(\left[-\nabla(2)\ell(\delta) (\hat{\theta}_\delta; z)\right]\right)}}{\sqrt{\det \left(\left[-\nabla(2)\ell(\delta) (\hat{\theta}_\delta; z)\right]\right)}} \frac{N(\hat{\theta}_\delta, I_{\gamma, \delta}^{-1}) (B_\delta)}{N(\hat{\theta}_\delta, I_{\gamma, \delta}) (B_\delta)}.$$  

We have

$$N(\hat{\theta}_\delta, I_{\gamma, \delta}^{-1}) (B_\delta) \geq 1 - \mathbb{P} \left(V' \left([-\nabla(2)\ell(\delta) (\hat{\theta}_\delta; z)]\right)^{-1} V > (M - 1)^2 e^2\right),$$

$$\geq 1 - \mathbb{P} \left(\|V\|_2 \geq (M - 1)e\kappa(s)^{1/2}\right),$$

where $V = (V_1, \ldots, V_s)$ i.i.d. $\mathcal{N}(0, 1)$, and where the second inequality uses H1 and the definition of $\kappa(s)$. By standard exponential bound for Gaussian random variables, we have

$$\mathbb{P} \left(\|V\|_2 \geq (M - 1)e\kappa(s)^{1/2}\right) \leq \exp \left(-\frac{1}{2} \left((M - 1)e\kappa(s)^{1/2} - s_1^{1/2}\right)^2\right).$$  

Hence

$$\frac{N(\hat{\theta}_\delta, I_{\gamma, \delta}^{-1}) (B_\delta)}{N(\hat{\theta}_\delta, I_{\gamma, \delta}) (B_\delta)} \leq \frac{1}{N(\hat{\theta}_\delta, I_{\gamma, \delta}) (B_\delta)} \leq \frac{1}{1 - e^{-\frac{1}{2} \left((M - 1)e\kappa(s)^{1/2} - s_1^{1/2}\right)^2}} \leq 2,$$
using the assumption that $(M - 1)e_K(\bar{s})^{1/2} - s_*^{1/2} \geq 2$. We split

$$\frac{\sqrt{\det \left( [-\nabla^{(2)} \ell^{[\delta]}(\hat{\theta}_\delta ; z)] \right)}}{\sqrt{\det \left( [-\nabla^{(2)} \ell^{[\delta]}(\hat{\theta}_\delta ; z)] \right)}}$$

$$= \frac{\sqrt{\det \left( [-\nabla^{(2)} \ell^{[\delta]}((\hat{\theta}_\delta, 0)\delta_\delta ; z)]_{\delta_\delta, \delta_\delta} \right)}}{\sqrt{\det \left( [-\nabla^{(2)} \ell^{[\delta]}((\hat{\theta}_\delta, 0)\delta_\delta ; z)]_{\delta_\delta, \delta_\delta} \right)}} \frac{\sqrt{\det \left( [-\nabla^{(2)} \ell^{[\delta]}((\hat{\theta}_\delta, 0)\delta_\delta ; z)]_{\delta_\delta, \delta_\delta} \right)}}{\sqrt{\det \left( [-\nabla^{(2)} \ell^{[\delta]}((\hat{\theta}_\delta, 0)\delta_\delta ; z)]_{\delta_\delta, \delta_\delta} \right)}}.$$

The convexity of the function $-\log \det$ can be used to show that for any pair of symmetric positive definite matrices $A, B$ of same size, $|\log \det(A) - \log \det(B)| \leq \max(||A^{-1}||_F, ||B^{-1}||_F) ||B - A||_F$. We use this to conclude that

$$\frac{\sqrt{\det \left( [-\nabla^{(2)} \ell^{[\delta]}((\hat{\theta}_\delta, 0)\delta_\delta ; z)]_{\delta_\delta, \delta_\delta} \right)}}{\sqrt{\det \left( [-\nabla^{(2)} \ell^{[\delta]}((\hat{\theta}_\delta, 0)\delta_\delta ; z)]_{\delta_\delta, \delta_\delta} \right)}} \leq \exp \left( \frac{1}{2} s_*^{1/2} \left\| \left[ \nabla^{(2)} \ell^{[\delta]}((\hat{\theta}_\delta)_{\delta_\delta} ; z) - \nabla^{(2)} \ell^{[\delta]}((\hat{\theta}_\delta)_{\delta_\delta} ; z) \right]_{\delta_\delta, \delta_\delta} \right\|_F \right) \leq e^{\frac{s_*^{1/2} \|e_{\rho M}^2\|_F}{2} \|\theta_\delta - \theta_*\|_2} \leq e^{-\frac{s_*^{1/2} \|e_{\rho M}^2\|_F}{2} \|\theta_\delta - \theta_*\|_2}.$$

Then we use Lemma 26 to get

$$\frac{\sqrt{\det \left( [-\nabla^{(2)} \ell^{[\delta]}((\hat{\theta}_\delta, 0)\delta_\delta ; z)]_{\delta_\delta, \delta_\delta} \right)}}{\sqrt{\det \left( [-\nabla^{(2)} \ell^{[\delta]}((\hat{\theta}_\delta, 0)\delta_\delta ; z)]_{\delta_\delta, \delta_\delta} \right)}} \leq \left( \frac{1}{\sqrt{s}_\delta} \right)^k.$$

It follows that for $z \in \mathcal{E}_{\hat{\theta}, \Lambda}$,

$$G_k \leq G_0 2 e^{2 s_\rho^{1/2} + \frac{1}{2} s_*^{1/2} \|e_{\rho M}^2\|_F} e^{\Lambda k} \left( \frac{p}{2} \right)^k e^{1/2 \left( \frac{1}{\sqrt{s}_\delta} \right)^k} \sum_{\delta \geq \delta_*, \|\delta\|_0 = s_\delta + k} \frac{\omega_\delta}{\omega_{\delta_*}}.$$

and under $H_3$ and for $p$ large enough so that $q \leq 1/2$,

$$\sum_{\delta \geq \delta_*, \|\delta\|_0 = s_\delta + k} \frac{\omega_\delta}{\omega_{\delta_*}} = \left( \frac{q}{1 - q} \right)^k \left( \frac{p - s_*}{k} \right) \leq (2q)^k e^{k \log(p)} \leq \left( \frac{2}{p^{\rho u}} \right)^k,$$
using the fact that \( \binom{p-s}{k} \leq e^{(p-s) \log(k)} \leq e^{p \log(k)} \). Therefore
\[
\sum_{k=1}^{\tilde{s}} G_k \leq G_0 2 \epsilon^{2p^{1/2} \epsilon + s^{1/2} \epsilon \cdot M \pi(s)} \sum_{k=1}^{\tilde{s}} \left( \frac{pe^\Lambda}{p^A} \sqrt{\frac{2\pi}{K(s)}} \right)^k.
\]
(90)
It follows that for \( \frac{2pe^\Lambda}{p^A} \sqrt{\frac{2\pi}{K(s)}} \leq 1 \), we get
\[
\sum_{k=1}^{\tilde{s}} G_k \leq 4 \frac{pe^\Lambda}{p^A} \sqrt{\frac{2\pi}{K(s)}} e^{2p^{1/2} \epsilon + s^{1/2} \epsilon \cdot M \pi(s)},
\]
which, together with [89], implies the stated bound in (18).

**Part two: Bernstein-von Mises approximation.** We introduce the following probability distributions on \( \Delta \times \mathbb{R}^p \):
\[
\hat{\Pi}_{\gamma,B}(\delta, d\theta | z) \propto \omega_\delta(2\pi \gamma)^{\frac{|\delta|_0}{2}} \left( \frac{\rho}{2} \right) \| \delta \|_0 \cdot e^{-h_\gamma(\delta, \theta; z)} \mathbf{1}_B(\delta, \theta) d\theta,
\]
and
\[
\hat{\Pi}_{\gamma,B}(\delta, d\theta | z) \propto \omega_\delta(2\pi \gamma)^{\frac{|\delta|_0}{2}} \left( \frac{\rho}{2} \right) \| \delta \|_0 \cdot e^{\ell(\theta; z) - \rho \| \theta \|_1 - \frac{1}{2\gamma}(\theta - \theta')^t(I_p - \gamma S)(\theta - \theta') \mathbf{1}_B(\delta, \theta) d\theta,
\]
and
\[
\hat{\Pi}_{\gamma,B}(\delta, d\theta | z) \propto \omega_\delta(2\pi \gamma)^{\frac{|\delta|_0}{2}} \left( \frac{\rho}{2} \right) \| \delta \|_0 \cdot e^{\ell(\theta; z) - \rho \| \theta \|_1}
\]
\[
\times e^{-\frac{1}{2}(\theta - \theta')^t(I_p - \gamma S)(\theta - \theta') - \frac{1}{2\gamma}(\theta - \theta')^t(I_p - \gamma S)(\theta - \theta') \mathbf{1}_B(\delta, \theta) d\theta.
\]
For all measurable subset \( C \) of \( \Delta \times \mathbb{R}^p \), we obviously can write
\[
|\hat{\Pi}_\gamma(C | z) - \hat{\Pi}^\infty_{\gamma}(C | z)| \leq |\hat{\Pi}_\gamma(C | z) - \hat{\Pi}^\infty_{\gamma,B}(C | z)| + |\hat{\Pi}^\infty_{\gamma,B}(C | z) - \hat{\Pi}^\infty_{\gamma}(C | z)|.
\]
Since \( \hat{\Pi}^\infty_{\gamma,B}(\cdot | z) \) is one of the component probability measures of \( \hat{\Pi}^\infty_{\gamma,B}(\cdot | z) \), by coupling inequality we have
\[
\| \hat{\Pi}^\infty_{\gamma,B}(\cdot | z) - \hat{\Pi}^\infty_{\gamma}(\cdot | z) \|_{tv} \leq 1 - \hat{\Pi}^\infty_{\gamma}(\{ \delta_\star \} \times B(\delta_\star) | z)
\]
\[
\leq 1 - \hat{\Pi}_\gamma(\{ \delta_\star \} \times B(\delta_\star) | z) + \| \hat{\Pi}_\gamma(\cdot | z) - \hat{\Pi}^\infty_{\gamma,B}(\cdot | z) \|_{tv}.
\]
We conclude that
\[
\| \hat{\Pi}_\gamma(\cdot | z) - \hat{\Pi}^\infty_{\gamma}(\cdot | z) \|_{tv} \leq 1 - \hat{\Pi}_\gamma(\{ \delta_\star \} \times B(\delta_\star) | z) + 2\| \hat{\Pi}_\gamma(\cdot | z) - \hat{\Pi}^\infty_{\gamma,B}(\cdot | z) \|_{tv}.
\]
(91)
Hence it suffices to bound the rightmost term of (91). For a measurable subset \( C \) of \( \Delta \times \mathbb{R}^p \), we have
\[
|\hat{\Pi}_\gamma(C | z) - \hat{\Pi}^\infty_{\gamma,B}(C | z)| \leq |\hat{\Pi}_\gamma(C | z) - \hat{\Pi}^\infty_{\gamma,B}(C | z)| + |\hat{\Pi}^\infty_{\gamma,B}(C | z) - \hat{\Pi}^\infty_{\gamma,B}(C | z)|
\]
\[
+ |\hat{\Pi}^\infty_{\gamma,B}(C | z) - \hat{\Pi}^\infty_{\gamma,B}(C | z)|.
\]
(92)
To deal with the first term on the right hand side of the inequality in (92), we first note that $\tilde{\Pi}_{\gamma,B}(\cdot | z)$ is no other than the restriction of $\tilde{\Pi}_{\gamma}$ to the set $B$. With this in mind, we make the following general observation. For any probability measure $\mu$, and a measurable set $A$ such that $\mu(A) > 0$, if $\mu_A$ denotes the restriction of $\mu$ to $A$ ($\mu_A(B) \equiv \mu(A \cap B)/\mu(A)$), we can decompose $\mu$ as $\mu = \mu_A + \mu(A^c)(\mu_{A^c} - \mu_A)$, where $A^c$ denotes the complement of $A$. This decomposition implies that for all measurable set $B$,

$$|\mu(B) - \mu_A(B)| \leq \max \left( \mu(A^c \cap B), \frac{\mu(A \cap B)}{\mu(A)} \mu(A^c) \right) \leq \mu(A^c).$$

(93)

In the particular case of $\tilde{\Pi}_{\gamma}$ and $\tilde{\Pi}_{\gamma,B}$, this bound readily implies that

$$\sup_{C \text{ meas.}} |\tilde{\Pi}_{\gamma}(C|z) - \tilde{\Pi}_{\gamma,B}(C|z)| \leq 1 - \tilde{\Pi}_{\gamma}(\bar{B}|z).$$

(94)

We claim that for all $z \in \mathcal{E}_{\hat{\rho}}$,

$$\sup_{C \text{ meas.}} \left| \tilde{\Pi}_{\gamma,B}(C|z) - \tilde{\Pi}_{\gamma,G}(C|z) \right| \leq 2t_1,$$

(95)

where

$$t_1 \equiv e^{\frac{2}{3} + 3\gamma\hat{\rho}^2 + 2\kappa(\hat{s})^2(Me)^2} - 1 + \frac{1}{p^m}.$$  

To establish (95), we note that if $f \geq g$ are two unnormalized positive densities on some measurable space with normalizing constants $Z_f, Z_g$ respectively, and $A$ is a measurable set, we have

$$\left| \frac{\int_A f(x)dx}{Z_f} - \frac{\int_A g(x)dx}{Z_g} \right| = \left| \frac{(Z_g - Z_f)\int_A f(x)dx}{Z_g Z_f} + \frac{\int_A (f(x) - g(x))d(x)}{Z_g} \right| \leq \left( \frac{Z_f}{Z_g} - 1 \right).$$

(96)

Owning to the first inequality of Lemma [19], we can apply this result with $f/Z_f$ as $\tilde{\Pi}_{\gamma,B}(\cdot | z)$, and $g/Z_g$ as $\tilde{\Pi}_{\gamma,G}(\cdot | z)$. Now, it suffices to control the ratio of normalizing constants of $\tilde{\Pi}_{\gamma,B}(\cdot | z)$ and $\tilde{\Pi}_{\gamma}(\cdot | z)$ given by

$$\frac{\sum_{\delta \in \Delta_\delta} \omega_\delta(2\pi\gamma)^{\frac{\|\delta\|}{2}} \left( \frac{p}{2} \right)^{\|\delta\|} \int_{B(\delta)} e^{-h_\gamma(\delta,\theta; z)}d\theta}{\sum_{\delta \in \Delta_\delta} \omega_\delta(2\pi\gamma)^{\frac{\|\delta\|}{2}} \left( \frac{p}{2} \right)^{\|\delta\|} \int_{B(\delta)} e^{\ell(\theta; z) - \rho\|\delta\|} - \frac{1}{2\gamma}(\theta - \theta_\delta)^2 A(\theta - \theta_\delta) + 3\gamma\hat{\rho}^2 + 2\kappa(\hat{s})^2(Me)^2}.$$  

(97)

By the second inequality of Lemma [19] and (84), for $z \in \mathcal{E}_{\hat{\rho}}$, $\delta \in \Delta_\delta$, and $\theta \in B(\delta)$, we have

$$-h_\gamma(\delta, \theta; z) \leq \ell(\theta_\delta; z) - \rho\|\delta\| - \frac{1}{2\gamma}(\theta - \theta_\delta)^2 A(\theta - \theta_\delta) + 3\gamma\hat{\rho}^2 + 2\kappa(\hat{s})^2(Me)^2,$$
where $A = I_p - \gamma(1 + 4\gamma(s))S$. Hence the numerator of (97) is upper-bounded by
\[
(2\pi\gamma)^{p/2}e^{3\gamma\rho^2 s + 2\gamma(s)^2(Me)^2} \sum_{\delta \in \Delta_x} \omega_{\delta} \left( \frac{\rho}{2} \right) \|\delta\|_0 \left[ \int_{\{\theta \in \mathbb{R}^p : \|\theta - \theta_0\|_2 \leq Me\}} e^{\ell(\theta;z) - \rho\|\theta\|_1} \mu_{\delta}(d\theta) \right] \times \frac{1}{\sqrt{\det([A]_{\delta})}},
\]
whereas the denominator is equal to
\[
(2\pi\gamma)^{p/2} \sum_{\delta \in \Delta_x} \omega_{\delta} \left( \frac{\rho}{2} \right) \left\|\delta\right\|_0 \left[ \int_{\{\theta \in \mathbb{R}^p : \|\theta - \theta_0\|_2 \leq Me\}} e^{\ell(\theta;z) - \rho\|\theta\|_1} \mu_{\delta}(d\theta) \right] \frac{T_\delta}{\sqrt{\det([I - \gamma S]_{\delta'})}},
\]
where $T_\delta$ is the probability of the set $\{u \in \mathbb{R}^{p-\|\delta\|_0} : \|u\|_2 \leq 2\sqrt{(1 + m)\gamma p}, \|u\|_\infty \leq 2\sqrt{(m + 1)\gamma \log(p)}\}$, under the distribution $N(0, \gamma([I - \gamma S]_{\delta'})^{-1})$, which is easily seen to be larger than $1 - \frac{1}{p^m}$ for all $m \geq 4$, by standard Gaussian tail bound. From these results and (71) we conclude that the ratio (97) is upper bounded by $(1 - \frac{1}{p^m})^{-1} e^{3\gamma\rho^2 s + 2\gamma(s)^2(Me)^2 + \frac{2}{3}}$, and this together with (96) imply (95).

We claim that for all $z \in E_{\rho,\epsilon}$
\[
\sup_{C : \text{meas.}} \left| \hat{\Pi}_{\gamma,B}(C|z) - \hat{\Pi}_{\gamma,B}^\infty(C|z) \right| \leq 16t_1 \left( (M + 1)\rho \epsilon^{1/2}e + \frac{(M + 1)^3}{3} e^{2\omega_{2,M}} \right) e^{4(M + 1)\rho \epsilon^{1/2}e + \frac{2}{3}(M + 1)^3 e^{2\omega_{2,M}}}.
\]
We establish this with the following general observation. Let $\{a_j\}, \{b_j\}$ be two discrete probability distributions on some discrete set $\mathcal{J}$, with $a_j > 0$. Let $\mu(j, dx) = a_j \mu_j(dx)$, $\nu(j, dx) = b_j \nu_j(dx)$ be two probability measures on $\mathcal{J} \times \mathbb{R}^p$, where for each $j$, $\mu_j(dx)$ and $\nu_j(dx)$ are equivalent probability measures supported by some measurable subset of $\mathbb{R}^p$. For any measurable set $A \subset \mathcal{J} \times \mathbb{R}^p$, we have
\[
\mu(A) - \nu(A) = \sum_j \left( 1 - \frac{b_j}{a_j} \right) a_j \mu_j(A^{(j)}) + \sum_j \frac{b_j}{a_j} \int_{A^{(j)}} \left( 1 - \frac{d\nu_j}{d\mu_j}(x) \right) a_j \mu_j(dx),
\]
where $A^{(j)} = \{ x \in \mathbb{R}^p : (j, x) \in A \}$. This implies that
\[
\sup_{A : \text{meas.}} |\mu(A) - \nu(A)| \leq \sup_j \left| 1 - \frac{b_j}{a_j} \right| + \sup_j \left( \frac{b_j}{a_j} \right) \sup_x \left| 1 - \frac{d\nu_j}{d\mu_j}(x) \right|.
\]
We apply this result with $\mu$ taken as $\hat{\Pi}_{\gamma,B}$, and $\nu$ taken as $\hat{\Pi}_{\gamma,B}^\infty$. In that case
\[
a_\delta = \frac{\tilde{a}_\delta}{\sum_{\delta \in A} \tilde{a}_\delta}, \quad b_\delta = \frac{\tilde{b}_\delta}{\sum_{\delta \in A} \tilde{b}_\delta},
\]
where
\[ a_\delta = \omega_\delta (2\pi \gamma)^{\frac{\parallel \delta \parallel_0}{2}} \left( \frac{\rho}{2} \right)^{\parallel \delta \parallel_0} \int_{B(\delta)} e^{\ell(\hat{\theta}_\delta; z) - \rho \parallel \theta_\delta \parallel_1} \, d\theta, \]
\[ b_\delta = \omega_\delta (2\pi \gamma)^{\frac{\parallel \delta \parallel_0}{2}} \left( \frac{\rho}{2} \right)^{\parallel \delta \parallel_0} \int_{B(\delta)} e^{\ell(\hat{\theta}_\delta; z) - \rho \parallel \theta_\delta \parallel_1} \times \int_{B(\delta)} e^{-\frac{1}{2}((\theta_\delta - \delta)^\prime \Gamma_{\gamma, \delta}(\theta_\delta - \delta) - \frac{1}{2}\gamma (\theta_\delta - \delta)\Gamma_{\gamma, \delta}(\theta_\delta - \delta) \delta)} \, d\theta. \]

We have
\[ \min_{\delta \in A} \frac{b_\delta}{a_\delta} \leq b_\delta \leq \frac{\max_{\delta \in A} b_\delta}{a_\delta}. \]

We take a Taylor expansion of the function \( u \mapsto \ell^{[\delta]}(u; z) \) to the third order around \( \hat{\theta}_\delta \), and note that \( \nabla \ell^{[\delta]}(\hat{\theta}_\delta; z) = 0 \) to conclude that for all \( z \in E_\rho, \theta \in B(\delta), \)
\[ \left| \ell(\theta_\delta; z) - \rho \parallel \theta_\delta \parallel_1 - \ell(\hat{\theta}_\delta; z) + \rho \parallel \hat{\theta}_\delta \parallel_1 + \frac{1}{2}([\theta_\delta - \hat{\theta}_\delta])\Gamma_{\gamma, \delta}([\theta_\delta - \hat{\theta}_\delta]) \right| \leq \rho \parallel \theta_\delta - \hat{\theta}_\delta \parallel_1 + \frac{1}{2}\omega_{2, M} \parallel \theta_\delta - \hat{\theta}_\delta \parallel_2^2 \leq c_1, \quad (100) \]
where \( c_1 = (M + 1)\rho \gamma^{1/2} + \frac{(M+1)^3}{6} \omega_{2, M} \epsilon^3 \). It follows easily that \( e^{-c_1} \leq \frac{b_\delta}{a_\delta} \leq e^{c_1} \), so that \( e^{-2c_1} \leq \frac{b_\delta}{a_\delta} \leq e^{2c_1} \). Similarly the Radon-Nykodym derivative satisfies
\[ e^{-2c_1} \leq \frac{d\mu_\rho}{d\mu_\delta}(\theta) \leq e^{2c_1}. \]

With these bounds, (98) easily follows from (99), and the fact that for all \( z \in (e^{-a}, e^{a}) \) for some \( a > 0 \), we have \( |1 - x| \leq ae^a \). The theorem hence follows from (98), (95), (94), (92) and (91). \( \Box \)

**Appendix C. Proof of Corollary 10**

We know from Lemma 22 that \( \hat{\Pi} \) is well-defined for all \( z \in \mathbb{R}^n \) and all \( \gamma > 0 \) such that \( 4\gamma \lambda_{\max}(X'X)/\sigma^2 \leq 1 \). \( \hat{\Pi} \) readily implies \( \hat{\Pi} \). Ignoring constants, it is straightforward that
\[ \ell(\theta; z) = -\frac{1}{2\sigma^2} \left\| z - X\theta \right\|_2^2, \quad \nabla \ell(\theta; z) = \frac{1}{\sigma^2} X'(z - X\theta), \quad \nabla^{(2)} \ell(\theta; z) = -\frac{1}{\sigma^2}(X'X). \]
Hence $H_4$ holds with $S = \tilde{S} = (X'X)/\sigma^2$. To apply Lemma 21 we need to check (62). For $\gamma > 0$, and $\delta \in \Delta$, we have

$$
\mathcal{L}_\gamma(\delta; \theta) = \ell(\theta; z) - \ell(\theta_\ast; z) - \langle \nabla \ell(\theta_\ast; z), \theta - \theta_\ast \rangle
+ \frac{2\gamma}{\sigma^4} (\theta - \theta_\ast)'X'X\delta X(\theta - \theta_\ast),
$$

$$
\leq - \frac{n}{2\sigma^2} (\theta - \theta_\ast)'\left( \frac{X'X}{n} \right) (\theta - \theta_\ast)
+ \frac{2n\gamma \lambda_{\max}(X\delta X_\delta')}{\sigma^4} (\theta - \theta_\ast)'\left( \frac{X'X}{n} \right) (\theta - \theta_\ast).
$$

Using this and the moment generating function of the Gaussian distribution yields

$$
\log E_{\ast} \left[ e^{\mathcal{L}_\gamma(\delta; \theta; Z) + (1 - \frac{\gamma}{2})(\nabla \ell(\theta_\ast; Z), \theta - \theta_\ast) + 1 + 2\frac{\rho}{\bar{\rho}}(\theta - \theta_\ast)'\left( \frac{X'X}{n} \right) (\theta - \theta_\ast)} \right]
\leq - \frac{n}{2\sigma^2} \left( 1 - \frac{4\gamma \lambda_{\max}(XX')}{\sigma^2} \right) \left( 1 - \frac{\rho}{\bar{\rho}} \right)^2 (\theta - \theta_\ast)'\left( \frac{X'X}{n} \right) (\theta - \theta_\ast)
\leq - \frac{n}{2\sigma^2} \left( \frac{\rho}{\bar{\rho}} - \frac{4\gamma \lambda_{\max}(XX')}{\sigma^2} \right) (\theta - \theta_\ast)'\left( \frac{X'X}{n} \right) (\theta - \theta_\ast).
$$

Since $\rho/\bar{\rho} = 1/\log(p)$, $(8/\sigma^2)\gamma \log(p)\lambda_{\max}(X'X) \leq 1$, and given $H_7$, we readily deduce that (62) holds with the rate function $r_0(x) = \frac{n\nu}{2\sigma^2 \log(p)} x^2$. In that case $a_0 = \frac{128\epsilon}{2}$. Furthermore, under the stated assumptions we easily check that $\eta$ in (63) satisfies $\eta \leq \frac{2}{\bar{\sigma}}(m_0 + 1 + 2s_\ast)$. This naturally suggests taking $s = s_\ast + \frac{2}{\bar{\sigma}}(m_0 + 1 + 2s_\ast)$ in $H_4$. For $\|\delta\|_0 \leq s$ and $\theta \in \mathbb{R}_{\delta}^p$, the upper-bound on $\mathcal{L}_\gamma(\delta; \theta; z)$ obtained above readily shows that

$$
\log E_{\ast} \left[ e^{\mathcal{L}_\gamma(\delta; \theta; Z)} \right] \leq \mathcal{L}_\gamma(\delta; \theta; Z)
\leq - \frac{n}{2\sigma^2} \left( 1 - \frac{4\gamma \nu \tilde{v}(\bar{\delta})}{\sigma^2} \right) (\theta - \theta_\ast)'\left( \frac{X'X}{n} \right) (\theta - \theta_\ast).
$$

Since $\gamma n = o(1)$, as $p \to \infty$, we see that for all $p$ large enough, $H_4$ holds with the rate function $r_1(x) = \frac{n\nu v(\bar{\delta})}{2\sigma^2} x^2$, and from the definitions, we obtain that

$$
\epsilon = \frac{6\sigma^2 \tilde{\rho}(s_\ast + \bar{s})^{1/2}}{n\nu v(\bar{\delta})} = \frac{24\sigma}{v(\bar{\delta})} \sqrt{\frac{\tilde{v}(\bar{s}) \log(p)}{n}}.
$$

We can then apply Theorem 5. Condition (13) and $a = o(\log(p))$ are implied by the assumptions on $\gamma$ in (23) of the main manuscript. We also easily see that $\gamma \rho^2 \bar{s} + \gamma \tilde{\kappa}(\bar{s})^2 (M \epsilon_2)^2 = O(\bar{s}) = o(\log(p))$. Consequently for all $m \geq 2$, $4p^{-m} \epsilon_2^2 + 3\gamma \rho^2 \bar{s} + 2\gamma \tilde{\kappa}(\bar{s})^2 (M \epsilon_2)^2 \leq p^{-(m-1)}$. Hence by Theorem 5 for any $m > 1$, $p$ large enough, and for any event $\mathcal{E}$
such that $\mathbb{P}_*(Z \in \mathcal{E}) \geq 1/2$,

$$
\mathbb{E}_* [\tilde{\Pi}_\gamma (\tilde{B}_{m,M} | Z) | Z \in \mathcal{E}] \geq 1 - \frac{1}{p^{m_0}} - \frac{1}{p^{m-1}} - 2e^{\delta \log(9p)} \sum_{j \geq 1} e^{-\frac{\delta}{6} r_1 \left( \frac{1}{2} M \epsilon \right)}
- 4e^{\frac{3}{2} p (2+u) s_*} \left( 1 + \frac{K(s_*)}{\rho^2} \right)^{s_*} \sum_{j \geq 1} e^{-\frac{\delta}{6} r_1 \left( \frac{1}{2} M \epsilon + 8p^{1/2} \left( \frac{1}{2} M \epsilon \right) \right)}.
$$

Given the expression of $r_1$ found above, we check that

$$
\sum_{j \geq 1} e^{-\frac{\delta}{6} r_1 \left( \frac{1}{2} M \epsilon \right)} \leq \frac{e^{-\frac{2\nu(s)}{12\sigma^2} (M \epsilon/2)^2}}{1 - e^{-\frac{2\nu(s)}{12\sigma^2} (M \epsilon/2)^2}},
$$

and noting that $\nu(s) \leq 1$ as a consequence of (20), we have

$$
\frac{2\nu(s)}{12\sigma^2} \left( \frac{M \epsilon}{2} \right)^2 = \frac{12M^2}{\nu(s)} (s_* + \bar{s}) \log(p) \geq 2M^2 \bar{s} \log(9p),
$$

for all $p \geq 2$. Hence,

$$
e^{\nu_0 \log(9p)} \sum_{j \geq 1} e^{-\frac{\delta}{6} r_1 \left( \frac{1}{2} M \epsilon \right)} \leq \frac{2}{(9p)^{M^2 \bar{s}}} \leq \frac{1}{p^{M^2 \bar{s}}}.
$$

Similarly, we check that

$$
- \frac{1}{12} r_1 \left( \frac{jM \epsilon}{2} \right) + 8p^{1/2} \left( \frac{jM \epsilon}{2} \right) \leq 0,
$$

for all $j \geq 1$ if $M \geq 64/n$. Hence, given $M > 2$, for all $p$ large enough (such that $n \geq 32$), we have

$$
\sum_{j \geq 1} e^{-\frac{1}{6} r_1 \left( \frac{1}{2} M \epsilon \right) + 8p^{1/2} \left( \frac{1}{2} M \epsilon \right)} \leq \sum_{j \geq 1} e^{-\frac{1}{12} r_1 \left( \frac{1}{2} M \epsilon \right)} \leq 2 \exp \left( -6M^2 (s_* + \bar{s}) \log(p) \right),
$$

and since $a = o(\log(p))$, and

$$
\log \left( 1 + \frac{K(s_*)}{\rho^2} \right)^{s_*} = s_* \log \left( 1 + \frac{\nu(s_*)}{16} \log(p) \right) = o(s_* \log(p)),
$$

as $p \to \infty$. Let us take $M > 2$ such that $3M^2 \geq 2 + u$. We then get

$$
4e^{\frac{3}{2} \rho^2 (2+u) s_*} \left( 1 + \frac{K(s_*)}{\rho^2} \right)^{s_*} \sum_{j \geq 1} e^{-\frac{1}{6} r_1 \left( \frac{1}{2} M \epsilon \right) + 8p^{1/2} \left( \frac{1}{2} M \epsilon \right)} \leq \frac{1}{p^{M^2 \bar{s}}},
$$

for all $p$ large enough. We conclude that there exist absolute constants $A_0$ such that for all $p \geq A_0$,

$$
\mathbb{E}_* [\tilde{\Pi}_\gamma (\tilde{B}_{m,M} | Z) | Z \in \mathcal{E}] \geq 1 - \frac{1}{p^{m_0}} - \frac{1}{p^{m-1}} - \frac{1}{p^{M^2 \bar{s}}}, \tag{101}
$$

for any measurable subset $\mathcal{E} \subseteq \mathcal{E}_\beta$ such that $\mathbb{P}_*(Z \in \mathcal{E}) \geq 1/2$. 

To apply Theorem 8, we need to check H5 and (14) of the main manuscript. For any given $M \geq \max(2, \sqrt{n^{\frac{2}{3}}})$, (24) of the main manuscript implies (14) of the main manuscript, for all $p$ large enough. Since $-\nabla^{(2)} \ell[\hat{\theta}; z] = (X'_\delta X_\delta)/\sigma^2$, it is straightforward that

$$\kappa(s) = \frac{nv(s)}{\sigma^2} > 0, \text{ and } \nu_2 = 0.$$  

By (16) of the main manuscript, for $\|\delta\|_0 \leq s$, we get

$$\|\hat{\theta}_\delta(z) - [\theta_\delta]\|_{2 \ell_p}(z) \leq \frac{\rho s_{\delta}^{1/2}}{\kappa(s)} \leq \frac{1}{6} \epsilon \leq \epsilon.$$  

This shows that H5 holds. We shall apply Theorem 8 with $\Lambda = 3 \log(n \land p)$. We easily check that $(M - 1)e_{\kappa(s)} 1/2 \geq s_{\delta}^{1/2} + 2$, 

$$\frac{p^e \Lambda}{p^u} \sqrt{\frac{2\pi}{\kappa(s)}} \leq \frac{\log(p)}{p^{u-2} v(s)},$$  

and

$$\left(\frac{4p^e \Lambda}{p^u} \sqrt{\frac{2\pi}{\kappa(s)}}\right) e^{2(M+2) \rho s_{\delta}^{3/2} e^{2\gamma s_{\delta}^2(M e)^2 e^{3\gamma \rho s_{\delta}^2 e_s^2}} \leq \frac{\log(p)}{p^{u-2} v(s)} e^{o(\log(p))} \leq \frac{1}{p^{u-3}},$$  

for all $p$ large enough. Hence we can apply Theorem 8 to conclude that

$$\Pi_{\gamma}(\{\delta_\star\} \times B_{m, M}^{(\delta_\star)}|Z) \Pi_\gamma(Z) \geq \Pi_{\gamma}(B_{m, M}|Z) \Pi_\gamma(Z) - \frac{1}{p^{u-3}} 1_{E_\gamma}(Z).$$  

Taking the expectation on both sides and dividing by $\mathbb{P}_\star(Z \in E_\rho)$ together with (101) yields

$$\mathbb{E}_\star \left[\Pi_{\gamma}(\{\delta_\star\} \times B_{m, M}^{(\delta_\star)}|Z) Z \in E \right] \geq 1 - \frac{1}{p^{u_0}} - \frac{1}{p^{u-1}} - \frac{1}{p^{M^2_{\delta}}} - \frac{1}{p^{u-3}}.$$  

It remains to control the term $\mathbb{P}_\star(Z \notin E_\rho, \Lambda(Z))$. By Gaussian tail bounds, we see that H6 and (20) imply that,

$$\mathbb{P}_\star(Z \notin E_\rho) \leq \frac{2}{p}. \quad (102)$$  

Since $Z = X_\delta \theta_\star + \sigma U$, where $U \sim \mathbf{N}(0, I_n)$, for any $\delta \in \mathcal{A}$, with $\delta \neq \delta_\star$, we have

$$\ell(\hat{\theta}; Z) - \ell(\hat{\theta}; Z) = U' [X_\delta (X'_\delta X_\delta)^{-1} X'_\delta - \delta_\star (X'_\delta X_\delta)^{-1} X'_\delta] U = \|Q_{\delta - \delta_\star} U\|_2^2,$$

where $X = QR$ with $Q \in \mathbb{R}^{n \times (p \land n)}$, $R \in \mathbb{R}^{(p \land n) \times p}$ denotes the QR decomposition of $X$. Using this, and by Lemma 5 of Castillo et al. (2015) which provides a deviation
bound on the maximum of chi-square random variables, we can find an absolute constant $c$ such that

$$
P_\star (Z \notin E_A) = \mathbb{P}_\star \left[ \bigcup_{k=1}^{\bar{s} - s_\star} \max_{\delta \in \mathcal{A} : \|\delta\|_0 = s_\star + k} \ell(\hat{\Theta}_\delta; Z) - \ell(\hat{\Theta}_\star; Z) > 3k \log (n \land p) \right]
\leq \sum_{k=1}^{\bar{s} - s_\star} \mathbb{P}_\star \left[ \max_{\delta \in \mathcal{A} : \|\delta\|_0 = s_\star + k} \ell(\hat{\Theta}_\delta; Z) - \ell(\hat{\Theta}_\star; Z) > 3k \log (n \land p) \right]
\leq \sum_{k=1}^{\bar{s} - s_\star} \frac{e^{ck}}{(n \land p)^{\frac{k}{4}}}.
$$

Since $\binom{n \land p}{k} \geq (n \land p - k)^k = e^{k \log (n \land p - k)} \geq e^{k \log ((n \land p) / 2)}$ for $k \leq \bar{s} \leq (n \land p) / 2$, for $n, p$ large enough, we get

$$
\sum_{k=1}^{\bar{s} - s_\star} \frac{e^{ck}}{(n \land p)^{\frac{k}{4}}} \leq \sum_{k=1}^{\bar{s}} e^{-4} e^{-\frac{1}{4}(\log ((n \land p) / 2) - 4c)} \leq 2e^{-\frac{1}{4}(\log ((n \land p) / 2) - 4c)} = \frac{C_1}{(n \land p)^{\frac{1}{4}}},
$$

for some absolute constant $C_1$. Hence

$$
P_\star [Z \notin E_{\rho, A}(Z)] \leq \frac{2}{p} + \frac{C_1}{(n \land p)^{\frac{1}{4}}}.
$$

The Bernstein-von Mises approximation part of the theorem follows easily.

\[\square\]

**Appendix D. Technical lemmas needed in the proof of Lemma 27.**

**Lemma 27.** Assume $H_6$. Then there exist absolute constants $A_0$, such that for all $p \geq A_0$, all $z \in \mathbb{R}^n$, all $\gamma > 0$, and all $\theta \in \mathbb{R}^p$, we have

$$
|h_\gamma(\delta_\star, \theta; z) - \hat{h}_\gamma(\delta_\star, \theta; z)| \leq \frac{\|\delta_\star\|_0 \gamma}{2} \left( p + \frac{C(X) \sqrt{n} \|\theta - \theta_\delta\|_2}{\sigma^2} \right)^2.
$$

**Proof.** Fix $\gamma > 0$, $\delta = \delta_\star$, $\theta$, and $z$ as above. Using (26) an (30) we have

$$
h_\gamma(\delta, \theta; z) - \hat{h}_\gamma(\delta, \theta; z) = \rho \left( \|J_\gamma(\delta, \theta)\|_1 - \|\hat{J}_\gamma(\delta, \theta)\|_1 \right)
\quad - \frac{1}{2\gamma} \left( \hat{J}_\gamma(\delta, \theta) - J_\gamma(\delta, \theta), J_\gamma(\delta, \theta) - \theta_\delta - \gamma \nabla \ell(\theta; z) + \hat{J}_\gamma(\delta, \theta) - \theta_\delta - \gamma \nabla \ell(\theta; z) \right).
$$

From the definition of the proximal operator there exist $E_{1,j} \in [-1, 1]$ and $E_{2,j} \in [-1, 1]$ such that

$$
J_{\gamma,j}(\delta, \theta) = (\theta_j + \gamma \nabla_j \ell(\theta; z) + \gamma \rho E_{1,j}) \delta_j,
$$

and

$$
\hat{J}_{\gamma,j}(\delta, \theta) = (\theta_j + \gamma \nabla_j \ell(\theta_\delta; z) + \gamma \rho E_{2,j}) \delta_j.
$$

Hence for $j$ such that $\delta_j = 1$, we have

$$
|J_{\gamma,j}(\delta, \theta) - \hat{J}_{\gamma,j}(\delta, \theta)| \leq 2\gamma \rho + \frac{\gamma}{\sigma^2} \left| \sum_{k : \delta_k = 0} \theta_k X_k \right| \leq 2\gamma \rho + \frac{C(X) \sqrt{n}}{\sigma^2} \|\theta - \theta_\delta\|_2,
$$
Lemma 28. Assume $H_0$, $H_7$. Then there exist some constants $A_0, C_0$ such that for all $p \geq A_0$, all $m \geq 1, M > 2$, and all $\theta_1, \theta_2 \in B^{(\delta_*)}_{m,M}$, such that $[\theta_1]_{\delta_*} = [\theta_2]_{\delta_*}$, we have

\[
\sup_{\delta \in \mathcal{E}_\delta} |h_\gamma(\delta, \theta_1; z) - h_\gamma(\delta, \theta_2; z)| \leq C_0 \left( 1 + \frac{\gamma n s_*^{1/2}}{\sigma^2} \right) \left( \rho + n M \epsilon \right) + C_0 \gamma s_* \left( 1 + \frac{\gamma n s_*^{1/2}}{\sigma^2} \right) \left( \rho + n M \epsilon + C(X) \sqrt{(m + 1)\gamma n p} \right).
\]

Proof. For convenience we write $\delta$, and $B^{(\delta)}$ instead of $\delta_*$ and $B^{(\delta_*)}_{m,M}$ respectively. We also set $s = |\delta_0||_0$. Fix $z \in \mathcal{E}_\rho$, and $\theta_1, \theta_2 \in B^{(\delta)}$ such that $[\theta_1]_{\delta_*} = [\theta_2]_{\delta_*}$. We start with some general remarks. For $\theta \in \mathbb{R}^p$, since $\ell$ is quadratic and $\nabla^{(2)} \ell(\theta; z) = -\frac{1}{\sigma^2}(X'X)$, we have

\[
\nabla \ell(\theta; z) = \nabla \ell(\theta_*; z) - \frac{\rho}{\sigma^2} (X'X)(\theta - \theta_*) - \frac{1}{\sigma^2} (X'X)(\theta_\delta - \theta_*).
\]

Hence for $z \in \mathcal{E}_\rho$, and $j$ such that $\delta_j = 1$, if $\nabla_j$ denotes the partial derivative operator with respect to $\theta_j$, we have

\[
|\nabla_j \ell(\theta; z)| \leq \frac{\rho}{2} + \frac{\sqrt{n}C(X)}{\sigma^2} \|\theta - \theta_\delta\|_2 + \frac{n\sqrt{v(s)}}{\sigma^2} \|\theta_\delta - \theta_*\|_2.
\]

Hence for all $\theta \in \mathbb{R}^p$,

\[
\sup_{z \in \mathcal{E}_\rho} \max_{j: \delta_j = 1} |\nabla_j \ell(\theta; z)| \leq \frac{\rho}{2} + \frac{\sqrt{n}C(X)}{\sigma^2} \|\theta - \theta_\delta\|_2 + \frac{n\sqrt{v(s)}}{\sigma^2} \|\theta_\delta - \theta_*\|_2.
\]
From (26), we have
\[
h_\gamma(\delta, \theta; z) = -\ell(\theta; z) - \langle \nabla \ell(\theta; z), J_\gamma(\delta, \theta) - \theta \rangle + \rho \| J_\gamma(\delta, \theta) \|_1 + \frac{1}{2\gamma} \| J_\gamma(\delta, \theta) - \theta \|_2^2.
\]
Since \( \ell \) is quadratic, we have
\[
-\ell(\theta; z) - \langle \nabla \ell(\theta; z), J_\gamma(\delta, \theta) - \theta \rangle = -\ell(J_\gamma(\delta, \theta); z) + \frac{1}{2\sigma^2} [J_\gamma(\delta, \theta) - \theta]' (X'X) [J_\gamma(\delta, \theta) - \theta].
\]
Hence
\[
h_\gamma(\delta, \theta_2; z) - h_\gamma(\delta, \theta_1; z) = U^{(1)} + U^{(2)} + U^{(3)} + U^{(4)},
\]
where
\[
U^{(1)} \overset{\text{def}}{=} \ell \left( J_\gamma(\delta, \theta_1); z \right) - \ell \left( J_\gamma(\delta, \theta_2); z \right),
\]
\[
U^{(2)} \overset{\text{def}}{=} \frac{1}{2\sigma^2} \left( [J_\gamma(\delta, \theta_2) - \theta_2]' (X'X) [J_\gamma(\delta, \theta_2) - \theta_2] \right.
\]
\[
\left. - [J_\gamma(\delta, \theta_1) - \theta_1]' (X'X) [J_\gamma(\delta, \theta_1) - \theta_1] \right),
\]
\[
U^{(3)} \overset{\text{def}}{=} \rho \left( \| J_\gamma(\delta, \theta_2) \|_1 - \| J_\gamma(\delta, \theta_1) \|_1 \right),
\]
\[
U^{(4)} \overset{\text{def}}{=} \frac{1}{2\gamma} \left( \| J_\gamma(\delta, \theta_2) - \theta_2 \|_2^2 - \| J_\gamma(\delta, \theta_1) - \theta_1 \|_2^2 \right).
\]
Note that \( U^{(1)} = \langle \nabla \ell(\bar{\theta}; z), J_\gamma(\delta, \bar{\theta}_1) - J_\gamma(\delta, \bar{\theta}_2) \rangle \), for some \( \bar{\theta} \) on the segment between \( J_\gamma(\delta, \theta_1) \) and \( J_\gamma(\delta, \theta_2) \). Therefore using (103) we get
\[
\left| U^{(1)} + U^{(3)} \right| \leq \left( \frac{3\bar{\rho}}{2} + \sqrt{n} \frac{\mu_0}{\sigma^2} \| \bar{\theta} - \tilde{\theta}_\delta \|_2 + \frac{n\sqrt{v(s)}}{\sigma^2} \| \tilde{\theta}_\delta - \theta_\star \|_2 \right) \| J_\gamma(\delta, \theta_2) - J_\gamma(\delta, \theta_1) \|_1.
\]
Since \( \bar{\theta} \) lies on the segment between \( J_\gamma(\delta, \theta_1) \) and \( J_\gamma(\delta, \theta_2) \), \( \bar{\theta} - \tilde{\theta}_\delta = 0 \), and we have
\[
\| \tilde{\theta}_\delta - \theta_\star \|_2 \leq \max \left( \| J_\gamma(\delta, \theta_1) - \theta_\star \|_2, \| J_\gamma(\delta, \theta_2) - \theta_\star \|_2 \right).
\] (104)
Given \( \theta \in B^{(\delta)} \), we recall that if \( \delta_j = 0 \), \( J_{\gamma,j}(\delta, \theta) = 0 \), and if \( \delta_j = 1 \), we have \( J_{\gamma,j}(\delta, \theta) = \theta_j + \gamma \nabla_j \ell(\theta; Z) + \gamma \rho E_j \), for some \( E_j \in [-1, 1] \) (that depends on \( \theta \)). Hence if \( \delta_j = 1 \), using (103) and the fact that \( \theta \in B^{(\delta)} \), we have \( |J_{\gamma,j}(\delta, \theta) - \theta_\star| \leq \left| \theta_j - \theta_\star \right| + \gamma \left( \frac{3\bar{\rho}}{2} + 2\sqrt{m} C(X) \sqrt{(1 + m) \gamma \log(p) / \sigma^2} + n \sqrt{v(s)} (M\epsilon) / \sigma^2 \right) \). It follows that for any \( \theta \in B^{(\delta)} \),
\[
\| J_\gamma(\delta, \theta) - \theta_\star \|_2 \leq M \epsilon + \gamma s^{1/2} \left( \frac{3\bar{\rho}}{2} + 2\sqrt{m} C(X) \sqrt{(m + 1) \gamma p} + \frac{n \sqrt{v(s)}}{\sigma^2} M \epsilon \right).
\]
Using the last display in (104) yields
\[
\| \tilde{\theta}_\delta - \theta_\star \|_2 \leq M \epsilon + \gamma s^{1/2} \left( \frac{3\bar{\rho}}{2} + 2\sqrt{m} C(X) \sqrt{(m + 1) \gamma p} + \frac{n \sqrt{v(s)}}{\sigma^2} M \epsilon \right).
\]
Similarly, we have
\[
\|J_\gamma(\delta, \theta_2) - J_\gamma(\delta, \theta_1)\|_1 \leq \|\theta_2 - \theta_1\|_1 + \gamma s \left( 2\rho + \frac{n}{\sigma^2} \bar{v}(s)^{1/2} \|\theta_2 - \theta_1\|_2 \right),
\]
\[
\leq \|\theta_2 - \theta_1\|_1 + 2\gamma s \left( \rho + \frac{n}{\sigma^2} \bar{v}(s)^{1/2} M\epsilon \right).
\]

We then get
\[
|U^{(1)} + U^{(3)}| \leq C_1 \left( \gamma s \left[ \rho + \frac{n}{\sigma^2} M\epsilon \right] + \|\theta_2 - \theta_1\|_1 \right)
\times \left[ \left( 1 + \frac{\gamma ns^{1/2}}{\sigma^2} \right) \left( \rho + nM\epsilon \right) + \frac{\gamma ns^{1/2}}{\sigma^2} \frac{C(X) \sqrt{n} \sqrt{(m + 1)\gamma p}}{\sigma^2} \right],
\]
for all \( p \) large enough, and for some absolute constants \( C_1, C_2 \), using the fact that \( \bar{v}(s) = O(1) \). For \( U^{(2)} \), we write
\[
U^{(2)} = \frac{1}{2\sigma^2} \Delta_1'(X'X)\Delta_2,
\]
where \( \Delta_1 = J_\gamma(\delta, \theta_2) - \theta_2 - J_\gamma(\delta, \theta_1) + \theta_1 \), and \( \Delta_2 = J_\gamma(\delta, \theta_2) - \theta_2 + J_\gamma(\delta, \theta_1) - \theta_1 \).

We note that \( \Delta_{1,j} = 0 \) for \( \delta_j = 0 \). Hence
\[
|U_2| \leq \frac{1}{2\sigma^2} \|\Delta_1\|_1 \max_{j: \delta_j = 1} |\langle X_j, X\Delta_2 \rangle| \leq \frac{1}{2\sigma^2} \|\Delta_1\|_1 \left( (n\bar{v}(s)^{1/2} \|\Delta_2,\delta\|_2 + C(X) \sqrt{n} \|\Delta_2 - \Delta_2,\delta\|_2) \right).
\]

With the same calculations as above we have
\[
\|\Delta_1\|_1 \leq 2\gamma s \left( \rho + \frac{n}{\sigma^2} \bar{v}(s)^{1/2} M\epsilon \right),
\]
\[
\|\Delta_2,\delta\|_2 \leq s^{1/2} \gamma \left( 2\rho + \max_{j: \delta_j = 1} |\nabla_j \ell(\theta_1, z)| + \max_{j: \delta_j = 1} |\nabla_j \ell(\theta_2, z)| \right)
\leq \frac{C\gamma s^{1/2}}{\sigma^2} \left( \sigma^2 \rho + nM\epsilon + C(X) \sqrt{n} \sqrt{(m + 1)\gamma p} \right),
\]
and
\[
\|\Delta_2 - \Delta_2,\delta\|_2 = \|(\theta_1 - \theta_{1,\delta}) + (\theta_2 - \theta_{2,\delta})\|_2 \leq 4\sqrt{(m + 1)\gamma p}.
\]

Therefore
\[
|U^{(2)}| \leq \frac{C\gamma s}{\sigma^2} \left( \rho + \frac{n}{\sigma^2} M\epsilon \right) \left[ \frac{n\gamma s^{1/2}}{\sigma^2} (\rho + nM\epsilon) + \left( 1 + \frac{\gamma ns^{1/2}}{\sigma^2} \right) C(X) \sqrt{(m + 1)\gamma p} \right],
\]
for all \( p \) large enough. For \( U^{(4)} \), we proceed similarly as with \( U^{(2)} \) to get
\[
|U^{(4)}| \leq \frac{C\gamma s}{\sigma^2} \left( \rho + \frac{n}{\sigma^2} M\epsilon \right) \left[ \frac{n\gamma s^{1/2}}{\sigma^2} (\rho + nM\epsilon) + \left( 1 + \frac{\gamma ns^{1/2}}{\sigma^2} \right) C(X) \sqrt{(m + 1)\gamma p} \right].
\]
We complete the proof by collecting these bounds together. □

We will also need the following lemma which is adapted from Lemma 4 of [Belloni and Chernozhukov (2009)]. For \( \tau > 0 \), we call \( Q_\tau(\theta, \cdot) \) the density of the normal distribution \( N(\theta, \tau^2 I_d) \).

**Lemma 29.** For some integer \( d \geq 1 \), fix \( \theta_0 \in \mathbb{R}^d \), \( R > 0 \), and define the set \( \Xi \equiv \{ u \in \mathbb{R}^d : \| u - \theta_0 \|_2 \leq R \} \). Let \( 0 < \sigma \leq \frac{R \sqrt{d \pi}}{320d} \), and \( r \equiv 4\sigma \sqrt{d} \). For all \( u, v \in \Xi \) such that \( \| u - v \|_2 \leq \sigma/4 \), we have

\[
\int_{\Xi_{uv}} Q_\sigma(u, z) \wedge Q_\sigma(v, z)dz \geq \frac{1}{4},
\]

where \( \Xi_{uv} \equiv \{ z \in \Xi : \| z - u \|_2 \leq r, \| z - v \|_2 \leq r \} \).

**Proof.** For \( x \in \mathbb{R}^d \), and \( a > 0 \) we shall write \( \mathbb{B}_a(x) \) to denote the ball of radius \( a \) centered at \( x \): \( \mathbb{B}_a(x) \equiv \{ z \in \mathbb{R}^d : \| z - x \|_2 \leq a \} \). Without any loss of generality we can assume that \( \theta_0 = 0_d \), and we can take \( u, v \in \Xi \) such that \( \| u - \theta_0 \|_2 = \| v - \theta_0 \|_2 = R \).

We set \( \Xi \equiv \mathbb{B}_r(u) \cap \mathbb{B}_r(v) \cap \Xi \), and we introduce

\[
\mathcal{H}_1 \equiv \{ \theta \in \mathbb{R}^d : 2 \langle \theta - u, v - u \rangle \geq \| v - u \|_2^2 \},
\]

\[
\mathcal{H}_2 \equiv \{ \theta \in \mathbb{R}^d : 2 \langle \theta - u, v - u \rangle < -\| v - u \|_2^2 \},
\]

\[
\mathcal{H}_3 \equiv \{ \theta \in \mathbb{R}^d : -\| v - u \|_2^2 \leq 2 \langle \theta - u, v - u \rangle < \| v - u \|_2^2 \}.
\]

We have \( \mathbb{R}^d = \mathcal{H}_1 \cup \mathcal{H}_2 \cup \mathcal{H}_3 \), and since \( \| \theta - u \|_2^2 = \| \theta - v \|_2^2 + 2 \langle \theta - v, u - v \rangle - \| v - u \|_2^2 \), we see that

\[
Q_\sigma(u, \theta) \wedge Q_\sigma(v, \theta) = \begin{cases} Q_\sigma(u, \theta) & \text{if } \theta \in \mathcal{H}_1, \\ Q_\sigma(v, \theta) & \text{if } \theta \notin \mathcal{H}_1. \end{cases}
\]

Therefore,

\[
\int_{\Xi_{uv}} Q_\sigma(u, \theta) \wedge Q_\sigma(v, \theta)d\theta \geq \int_{\Xi_{uv} \cap \mathcal{H}_1} Q_\sigma(u, \theta)d\theta + \int_{\Xi_{uv} \cap \mathcal{H}_2} Q_\sigma(v, \theta)d\theta.
\]

Let \( w = (u + v)/2 \), \( \nu \equiv \{ z \in \mathbb{R}^d : \langle z, w \rangle = \| w \|_2^2 \} \), and the half-space \( \nu_- \equiv \{ z \in \mathbb{R}^d : \langle z, w \rangle \leq \| w \|_2^2 \} \). Define also \( \mathbb{B}_r \equiv \mathbb{B}_r(u) \cap \mathbb{B}_r(v) \). It can easily checked that for \( j \in \{ 1, 2 \} \),

\[
\mathbb{B}_r \cap \mathcal{H}_j \cap \left( \nu_- - \frac{r^2 w}{R \| w \|_2^2} \right) \subseteq \Xi_{uv} \cap \mathcal{H}_j.
\]

Indeed any \( \theta \in \mathbb{R}^d \) can be written \( \theta = \frac{\theta w}{\| w \|_2^2} \frac{w}{\| w \|_2} + s \), where \( \langle s, w \rangle = 0 \). And if \( \theta \in \mathbb{B}_r \) then \( \| s \|_2 = \| \theta - \frac{\theta w}{\| w \|_2^2} \frac{w}{\| w \|_2} \|_2 \leq \| \theta - w \|_2 = \| \theta - (u + v)/2 \|_2 \leq r \). Also if \( \theta \in \mathbb{B}_r \cap \mathcal{H}_j \cap \left( \nu_- - \frac{r^2 w}{R \| w \|_2^2} \right) \cap \Xi_{uv} \cap \mathcal{H}_j \).
then $\frac{\theta'w}{\|w\|_2^2} \leq \|w\|_2 - \frac{r^2}{R} \leq R - \frac{r^2}{R}$. Hence $\|\theta\|^2 = \left(\frac{\theta'w}{\|w\|_2^2}\right)^2 + \|s\|^2 \leq \left(R - \frac{r^2}{R}\right)^2 + r^2 \leq R^2$.

We conclude that

$$
\int_{\Xi_{u,v} \cap \mathcal{H}_1} Q_\sigma(u, \theta)d\theta \geq \int_{\Xi_{u,v} \cap \mathcal{H}_1 \cap \mathcal{\varphi}^-} Q_\sigma(u, \theta)d\theta \\
\geq \int_{\mathcal{B}_r \cap \mathcal{H}_1 \cap \mathcal{\varphi}^-} Q_\sigma(u, \theta)d\theta - \int_{\mathcal{\varphi}^- \setminus \left(\varphi - \frac{r^2w}{R\|w\|_2^2}\right)} Q_\sigma(u, \theta)d\theta.
$$

And $\theta \in \mathcal{\varphi}^- \setminus \left(\varphi - \frac{r^2w}{R\|w\|_2^2}\right)$ is equivalent to $-\frac{r^2}{R} \leq \langle \theta - u, \frac{w}{\|w\|_2} \rangle \leq 0$. Hence

$$
\int_{\mathcal{\varphi}^- \setminus \left(\varphi - \frac{r^2w}{R\|w\|_2^2}\right)} Q_\sigma(u, \theta)d\theta = \int_0^{\frac{r^2}{R\sqrt{2\pi}}} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{r^2u^2}{2\sigma}} du \leq \frac{r^2}{\sigma R \sqrt{2\pi}}. A similarly calculations holds for the second integral. We conclude that

$$
\int_{\Xi_{u,v}} Q_\sigma(u, \theta) \wedge Q_\sigma(v, \theta)d\theta \geq \int_{\mathcal{B}_r \cap \mathcal{H}_1 \cap \mathcal{\varphi}^-} Q_\sigma(u, \theta)d\theta + \int_{\mathcal{B}_r \cap \mathcal{H}_2 \cap \mathcal{\varphi}^-} Q_\sigma(v, \theta)d\theta - \frac{2r^2}{\sigma R \sqrt{2\pi}}.
$$

It can be checked that the sets $\mathcal{B}_r(u)$, $\mathcal{B}_r(v)$, $\mathcal{H}_1$ and $\mathcal{H}_2$ are symmetric with respect to $\varphi$. Indeed for $x \in \mathbb{R}^d$, the reflection of $x$ with respect to $\varphi$ is $x - 2\left(\frac{x'w}{\|w\|_2^2} - \|w\|_2\right)w$, and it is easy verification to check that if $x$ belongs to any of these sets, its reflection also belong. Hence,

$$
\int_{\Xi_{u,v}} Q_\sigma(u, \theta) \wedge Q_\sigma(v, \theta)d\theta \geq \frac{1}{2} \int_{\mathcal{B}_r \cap \mathcal{H}_1} Q_\sigma(u, \theta)d\theta + \frac{1}{2} \int_{\mathcal{B}_r \cap \mathcal{H}_2} Q_\sigma(v, \theta)d\theta - \frac{2r^2}{\sigma R \sqrt{2\pi}}
\geq \frac{1}{2} \int_{\mathcal{B}_r \cap \mathcal{H}_1} Q_\sigma(u, \theta)d\theta + \frac{1}{2} \int_{\mathcal{B}_r \cap \mathcal{H}_2} Q_\sigma(v, \theta)d\theta - \frac{1}{2} \int_{\mathcal{H}_3} Q_\sigma(v, \theta)d\theta - \frac{2r^2}{\sigma R \sqrt{2\pi}}.
$$

We have

$$
\int_{\mathcal{B}_r \cap \mathcal{H}_1} Q_\sigma(u, \theta)d\theta + \int_{\mathcal{B}_r \cap \mathcal{H}_2^c} Q_\sigma(v, \theta)d\theta \\
= \int_{\mathbb{R}^d} 1_{\mathcal{B}_r \cap \mathcal{H}_1}(\theta)Q_\sigma(u, \theta)d\theta + \int_{\mathbb{R}^d} 1_{\mathcal{B}_r \cap \mathcal{H}_2}(v - u + \theta)Q_\sigma(u, \theta)d\theta \\
\geq \int_{\mathbb{R}^d} 1_{\mathcal{B}_r \cap \mathcal{H}_1}(\theta)Q_\sigma(u, \theta)d\theta + \int_{\mathbb{R}^d} 1_{\mathcal{B}_r(u) \cap \mathcal{H}_2}(v)Q_\sigma(u, \theta)d\theta,
$$

where the last inequality uses the fact, easy to check that if $\theta \in \mathcal{B}_r(u) \cap \mathcal{H}_2$, then $\theta + v - u \in \mathcal{B}_r \setminus \mathcal{H}_1$. Furthermore,

$$
\mathcal{B}_r(u) = (\mathcal{B}_r(u) \cap \mathcal{H}_1) \cup (\mathcal{B}_r(u) \cap \mathcal{H}_2) \cup (\mathcal{B}_r(u) \cap \mathcal{H}_3),$$
and $B_r(u) \cap H_1 = B_r \cap H_1$. Hence

$$\int_{\Xi_{u,v}} Q_\sigma(u, \theta) \wedge Q_\sigma(v, \theta) d\theta \geq \frac{1}{2} \int_{B_r(u)} Q_\sigma(u, \theta) d\theta - \frac{1}{2} \int_{H_3} Q_\sigma(u, \theta) d\theta - \frac{\sigma^2 r^2}{\sigma R \sqrt{2\pi}}.$$

With $r = 4\sigma \sqrt{d}$, we have $\int_{B_r(u)} Q_\sigma(u, \theta) d\theta \geq 1 - 10^{-4}$. With $\sigma \leq \frac{R \sqrt{2\pi}}{320d}$,

$$\frac{2r^2}{\sigma R \sqrt{2\pi}} \leq \frac{1}{10},$$

and with $\|v - u\|_2 \leq \sigma/4$,

$$\int_{H_3} Q_\sigma(u, \theta) d\theta + \int_{H_3} Q_\sigma(v, \theta) d\theta = \int_{\|v-u\|_2^2 \over 2 \over 2 \sqrt{2\pi} \sigma} e^{-u^2 \over 2 \sigma} du + \int_{\|v-u\|_2^2 \over 2 \over 2 \sqrt{2\pi} \sigma} e^{-u^2 \over 2 \sigma} du \leq \frac{3\|v-u\|_2}{\sqrt{2\pi} \sigma} \leq \frac{3}{4\sqrt{2\pi}}.$$

In conclusion,

$$\int_{\Xi_{u,v}} Q_\sigma(u, \theta) \wedge Q_\sigma(v, \theta) d\theta \geq \frac{1}{2} - \frac{10^{-4}}{2} - \frac{3}{8\sqrt{2\pi}} - \frac{1}{10} \geq \frac{1}{4},$$

as claimed.

\[ \square \]

**Acknowledgements:** The authors are grateful to Galin Jones, Scott Schmidler and Yuekai Sun for very helpful discussions.

**References**

Atchade, Y. A. (2017). On the contraction properties of some high-dimensional quasi-posterior distributions. *Ann. Statist.* 45 2248–2273.

Atchadé, Y. F. (2015). A Moreau-Yosida approximation scheme for high-dimensional posterior and quasi-posterior distributions. *ArXiv e-prints*.

Atchadé, Y. F. and Rosenthal, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* 11 815–828.

Bauschke, H. H. and Combettes, P. L. (2011). *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, Springer, New York. With a foreword by Hédy Attouch. URL [http://dx.doi.org/10.1007/978-1-4419-9467-7](http://dx.doi.org/10.1007/978-1-4419-9467-7)
Belloni, A. and Chernozhukov, V. (2009). On the computational complexity of MCMC-based estimators in large samples. *Ann. Statist.* **37** 2011–2055.

Bhattacharya, A., Chakraborty, A. and Mallick, B. (2016). Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika* **103** 985 – 991.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732.

Bottolo, L. and Richardson, S. (2010). Evolutionary stochastic search for bayesian model exploration. *Bayesian Anal.* **5** 583–618.

Castillo, I., Schmidt-Hieber, J. and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018.

Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101.

Dyer, M., Frieze, A. and Kannan, R. (1991). A random polynomial-time algorithm for approximating the volume of convex bodies. *J. ACM* **38** 1–17.

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

Gottardo, R. and Raftery, A. E. (2008). Markov chain monte carlo with mixtures of mutually singular distributions. *Journal of Computational and Graphical Statistics* **17** 949–975.

Green, P. J. (2003). Trans-dimensional Markov chain Monte Carlo. In *Highly structured stochastic systems*, vol. 27 of *Oxford Statist. Sci. Ser.* 179–206.

Guan, Y. and Krone, S. M. (2007). Small-world mcmc and convergence to multimodal distributions: From slow mixing to fast mixing. *Ann. Appl. Probab.* **17** 284–304.

Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. 2nd ed. Cambridge University Press, New York, NY, USA.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies. *Ann. Statist.* **33** 730–773.

Kleijn, B. J. K. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.* **34** 837–877.

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.

Lovasz, L. and Simonovits, M. (1990). The mixing rate of markov chains, an isoperimetric inequality, and computing the volume. In *Proceedings of the 31st Annual Symposium on Foundations of Computer Science*. SFCS ’90, IEEE Computer
Lovász, L. and Simonovits, M. (1993). Random walks in a convex body and an improved volume algorithm. *Random Structures Algorithms* **4** 359–412.

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

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

Mangoubi, O. and Smith, A. (2017). Rapid Mixing of Hamiltonian Monte Carlo on Strongly Log-Concave Distributions. *ArXiv e-prints* .

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

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *JASA* **83** 1023–1032.

Moreau, J.-J. (1965). Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France* **93** 273–299.

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

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of $m$-estimators with decomposable regularizers. *Statistical Science* **27** 538–557.

Parikh, N. and Boyd, S. (2013). Proximal algorithms. *Foundations and Trends in Optimization* **1** 123–231.

Patrinos, P., Stella, L. and Bemporad, A. (2014). Forward-backward truncated Newton methods for convex composite optimization. *ArXiv e-prints* .

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*. 2nd ed. Springer Texts in Statistics, Springer-Verlag, New York.

Rockova, V. and George, E. I. (2014). Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association* **109** 828–846.

Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Trans. Inf. Theor.* **59** 3434–3447.

Schreck, A., Fort, G., Le Corff, S. and Moulines, E. (2013). A shrinkage-thresholding Metropolis adjusted Langevin algorithm for Bayesian variable selection. *ArXiv e-prints* .

Sinclair, A. and Jerrum, M. (1989). Approximate counting, uniform generation and rapidly mixing Markov chains. *Inform. and Comput.* **82** 93–133.

Sur, P., Chen, Y. and Candès, E. J. (2017). The Likelihood Ratio Test in High-Dimensional Logistic Regression Is Asymptotically a Rescaled Chi-Square. *ArXiv*
ON HIGH-DIMENSIONAL SPIKE-AND-SLAB POSTERIOR DISTRIBUTIONS

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B 58 267–288.

Tierney, L. (1994). Markov chains for exploring posterior distributions. Ann. Statist. 22 1701–1762. With discussion and a rejoinder by the author.

Vempala, S. (2005). Geometric random walks: a survey. In Combinatorial and computational geometry, vol. 52 of Math. Sci. Res. Inst. Publ. Cambridge Univ. Press, Cambridge, 577–616.

Woodard, D. B., Schmidler, S. C. and Huber, M. (2009). Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. Ann. Appl. Probab. 19 617–640.

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.