ON THE CONTRACTION PROPERTIES OF SOME HIGH-DIMENSIONAL QUASI-POSTERIOR DISTRIBUTIONS

YVES F. ATCHADÉ

(April 2016; First version Sept. 2015)

Abstract. We study the contraction properties of a quasi-posterior distribution $\check{\Pi}_{n,d}$ obtained by combining a quasi-likelihood function and a sparsity inducing prior distribution on $\mathbb{R}^d$, as both n (the sample size), and d (the dimension of the parameter) increase. We derive some general results that highlight a set of sufficient conditions under which $\check{\Pi}_{n,d}$ puts increasingly high probability on sparse subsets of $\mathbb{R}^d$, and contracts towards the true value of the parameter. We apply these results to the analysis of logistic regression models, and binary graphical models, in high-dimensional settings. For the logistic regression model, we shows that for well-behaved design matrices, the posterior distribution contracts at the rate $O(\sqrt{s_\star \log(d)/n})$, where $s_\star$ is the number of non-zero components of the parameter. For the binary graphical model, under some regularity conditions, we show that a quasi-posterior analog of the neighborhood selection of Meinshausen and Buhlmann (2006) contracts in the Frobenius norm at the rate $O(\sqrt{(p+S)\log(p)/n})$, where p is the number of nodes, and S the number of edges of the true graph.

1. Introduction

Let $\mathcal{Z}^{(n)}$ denote a measurable space equipped with a reference sigma-finite measure denoted $\mathrm{d}z$. The upper script n represents the sample size. Let Z be a $\mathcal{Z}^{(n)}$-valued random variable that we model as having distribution $\mathbb{P}_\theta^{(n)}$ given a parameter $\theta \in \mathbb{R}^d$. We assume that $\mathbb{P}_\theta^{(n)}$ has a density $f_{n,\theta}$: $\mathbb{P}_\theta^{(n)}(\mathrm{d}z) = f_{n,\theta}(z)\mathrm{d}z$. Let Π be a prior distribution on $\mathbb{R}^d$. The resulting posterior distribution for learning the parameter θ is the random probability measure

$$A \mapsto \frac{\int_A f_{n,\theta}(Z)\Pi(\mathrm{d}\theta)}{\int_{\mathbb{R}^d} f_{n,\theta}(Z)\Pi(\mathrm{d}\theta)}, \quad A \text{ meas.} \subseteq \mathbb{R}^d.$$

2010 *Mathematics Subject Classification.* 62F15, 62Jxx.

Key words and phrases. Quasi-posterior distributions, convergence rate, variable selection, high-dimensional inference, linear regression, logistic regression.

This work is partially supported by the NSF grant DMS 1228164 and DMS 1513040.

Y. F. Atchadé: University of Michigan, 1085 South University, Ann Arbor, 48109, MI, United States. *E-mail address:* yvesa@umich.edu.

In practice, many inference problems are best tackled using quasi-likelihood (or a pseudo-likelihood) functions. In the Bayesian framework, this leads to a quasi-Bayesian inference. Let $(\theta, z) \mapsto q_{n,\theta}(z)$ denote a jointly measurable function such that $0 < \int_{\mathbb{R}^d} q_{n,\theta}(z)\Pi(\mathrm{d}\theta) < \infty$, almost surely $[\mathrm{d}z]$. Substituting $q_{n,\theta}$ in place of $f_{n,\theta}$ yields the quasi-posterior (QP) distribution

$$\check{\Pi}_{n,d}(A|Z) \stackrel{\text{def}}{=} \frac{\int_A q_{n,\theta}(Z)\Pi(\mathrm{d}\theta)}{\int_{\mathbb{R}^d} q_{n,\theta}(Z)\Pi(\mathrm{d}\theta)}, \quad A \subseteq \mathbb{R}^d. \tag{1}$$

Although $\check{\Pi}_{n,d}$ is not a proper posterior distribution, it possesses the key property that it is a probability distribution obtained by tilting a prior distribution using a likelihood-like function. Hence, to the extent that the quasi-likelihood function $\theta \mapsto q_{n,\theta}(Z)$ contains information about the true value of the parameter θ, one can expect the same from the quasi-posterior distribution (1), in which case valid inferential procedures can be derived using $\check{\Pi}_{n,d}$. This idea can be made precise by noting that (1) is a solution of the minimization

$$\min_{\mu \ll \Pi} \left[-\int_{\mathbb{R}^d} \log q_{n,\theta}(Z)\mu(\mathrm{d}\theta) + \mathsf{KL}(\mu|\Pi) \right],$$

where $\mathsf{KL}(\mu|\Pi) \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} \log(\mathrm{d}\mu/\mathrm{d}\Pi)\mathrm{d}\mu$ is the KL-divergence between μ and Π, and where the minimization is over all probability measures that are absolutely continuous with respect to the prior Π. We refer to Zhang (2006) for more details (and in particular to Proposition 5.1 of that paper that contains a proof of the above statement). The implication of this result is that, under appropriate regularity conditions, one can expect the QP distribution to concentrate around the maximizer of the function $\theta \mapsto \log q_{n,\theta}(Z)$, provided that the prior distribution does not prevent it.

As pointed out to us by a referee, QP distributions are commonly used in the PAC-Bayesian framework to aggregate estimators (McAllester (1999); Catoni (2004); Dalalyan and Tsybakov (2007); Alquier and Lounici (2011); Arias-Castro and Lounici (2014)). However in this literature the emphasis is typically on the estimators, not on the QP distributions themselves. An influential work in the quasi-Bayesian literature is Chernozhukov and Hong (2003), which subsequently led to the development of quasi-Bayesian inference in semi-parametric modeling, particularly models arising from moment and conditional moment restrictions (Liao and Jiang (2011); Yang and He (2012); Kato (2013); Li and Jiang (2014)). Approximate Bayesian computation (ABC) methods (see e.g. Marin et al. (2012) and the references therein) that are very popular in Bayesian data analysis can also be viewed, from an inferential standpoint, as a use of QP distributions.

The present paper is motivated by the idea that quasi-Bayesian inference holds a great potential for dealing with high-dimensional statistical models. For some of these models, a likelihood-based inference is often intractable, and this has impeded somewhat the applicability of the Bayesian framework in this area. However, M-estimation procedures that maximizes various quasi/pseudo-likelihood functions are often readily available. Using the quasi-Bayesian framework, these quasi-likelihood functions can be easily employed to derive tractable quasi-Bayesian procedures.

We study the behavior of the QP distribution (1) when the prior distribution Π is given by

$$\Pi(\mathrm{d}\theta) = \sum_{\delta \in \Delta_d} \pi_\delta \Pi(\mathrm{d}\theta|\delta), \tag{2}$$

for a discrete distribution $\{\pi_\delta, \delta \in \Delta_d\}$ on $\Delta_d \stackrel{\text{def}}{=} \{0,1\}^d$, and a sparsity inducing prior $\Pi(\mathrm{d}\theta|\delta)$ on $\mathbb{R}^d$, that we build as follows. Given δ, the components of θ are independent, and for $1 \leq j \leq d$,

$$\theta_j|\delta \sim \begin{cases} \mathsf{Dirac}(0) & \text{if } \delta_j = 0 \\ \mathsf{Laplace}(\rho) & \text{if } \delta_j = 1 \end{cases}, \tag{3}$$

where $\mathsf{Dirac}(0)$ is the Dirac measure on $\mathbb{R}$ with full mass at 0, and $\mathsf{Laplace}(\rho)$ denotes the Laplace distribution with parameter $\rho > 0$. The marginal prior distribution of θ_j implied by (3) belongs to the class of spike-and-slab priors (Mitchell and Beauchamp (1988)).

We work under the assumption that $Z \sim \mathbb{P}_{\theta_\star}^{(n)}$ for some $\theta_\star \in \mathbb{R}^d$. When d is assumed fixed and $n \to \infty$, it is known from the initial work of Chernozhukov and Hong (2003) that $\check{\Pi}_{n,d}$ concentrates around $\theta_\star$, and is asymptotically Gaussian (when properly scaled). Infinite-dimensional extensions of such results have recently been studied (Liao and Jiang (2011); Florens and Simoni (2012); Kato (2013)). The present paper focus on the case where $\check{\Pi}_{n,d}$ arises from a high-dimensional parametric model with the sparsity inducing prior (2-3), and the results that we derive substantially extend previous works by Castillo et al. (2015); Li and Jiang (2014). More precisely, we derive two general results (Theorem 8 and Theorem 9) that highlights the key determinants that control the convergence and convergence rate of $\check{\Pi}_{n,d}$ towards $\theta_\star$. The theorems are obtained by combining ideas from Castillo et al. (2015) together with a general methodology for studying high-dimensional M-estimators synthesized in Negahban et al. (2012), as well as a key technical result by Kleijn and van der Vaart (2006) on the existence of test functions.

We apply these results to the Bayesian analysis of high-dimensional logistic regression models. We derive a non-asymptotic result (Theorem 2) that shows that for large d, and appropriately large sample size n, the resulting posterior distribution

$\check{\Pi}_{n,d}$ puts a high probability on sparse subsets of $\mathbb{R}^d$, and contracts towards the true value of the parameter $\theta_\star$ as $n, d \to \infty$, at the rate

$$O\left(\sqrt{\frac{s_\star \log(d)}{n}}\right),$$

where $s_\star = \|\theta_\star\|_0$. The constant in the big-O notation depends crucially on some smallest restricted eigenvalues of the Fisher information matrix of the model.

We also apply the results to a quasi-Bayesian inference of high-dimensional binary graphical models. Discrete graphical models are known to raise significant difficulties due to the intractable nature of the likelihood function. A very successful frequentist approach to deal with large graphical models is the neighborhood selection method of Meinshausen and Buhlmann (2006) initially proposed for Gaussian graphical models, and extended to the Ising model by Ravikumar et al. (2010). We analyze a quasi-Bayesian version of neighborhood selection applied to binary graphical models. We show that as $n, p \to \infty$ (where p is the number of nodes in the graph), provided that n is sufficiently large, the QP distribution obtained from neighborhood selection contracts towards the true model parameter $\theta_\star$ in the Frobenius norm at the rate

$$O\left(\sqrt{\frac{(p+S)\log(p)}{n}}\right),$$

where S is the number of edges in the graph defined by $\theta_\star$. This convergence compares very well with the best existing frequentist results. For instance Sun and Zhang (2013) shows that the scaled g-Lasso version of neighborhood selection in the Gaussian case converges at the rate $O\left(s_\star\sqrt{\log(d)/n}\right)$ in the spectral norm, where $s_\star$ is the maximum degree of the graph defined by $\theta_\star$. Our theory suggests that good approximation of $\theta_\star$ in the Frobenius norm typically requires much larger sample sizes than typically available in such problems. This motivates us to look at other norms. We analyze the contraction of $\check{\Pi}_{n,d}$ in the norm $|\!|\!|\theta|\!|\!| \stackrel{\text{def}}{=} \max_j \|\theta_{\cdot j}\|_2$, where $\theta_{\cdot j}$ is the j-th column of θ. We show that in this norm, the QP distribution obtained from neighborhood selection contracts towards $\theta_\star$ at the rate

$$O\left(\sqrt{\frac{s_\star \log(p)}{n}}\right),$$

where here $s_\star$ is the maximum degree of the graph defined by the true parameter $\theta_\star$. Furthermore, the sample size n required for this result to hold is comparable to the sample size requirement in high-dimensional logistic regression.

An important issue not addressed in this work is how to obtain Monte Carlo samples from the QP distribution (1). It is well known that posterior and quasi-posterior

distributions built from discrete-continuous mixture priors as in (2)-(3) are computational difficult to handle with standard Markov Chain Monte Carlo algorithms. However there has been some recent progress, including the STMaLa of Schreck et al. (2013), or the Moreau approximation approach of the author developed in Atchadé (2015). We point the reader to these works for more details and some additional references. Further discussion of computational methods can be in Castillo et al. (2015).

The remainder of the paper is organized as follows. First we close the introduction with some notation that will be used throughout the paper. For improved readability of the paper, we first highlight the applications, with a discussion of the logistic regression in Section 2, and the binary graphical model in Section 3. The general analysis of the QP distribution $\check{\Pi}_{n,d}$ is developed in Section 4. All the proofs are gathered together in Section 5.

1.1. **Notation.** For an integer $d \geq 1$, we equip the Euclidean space $\mathbb{R}^d$ with its usual Euclidean inner product $\langle \cdot, \cdot \rangle$, associated norm $\|\cdot\|_2$, and its Borel sigma-algebra. We set $\Delta_d \stackrel{\text{def}}{=} \{0,1\}^d$. We will also use the following norms on $\mathbb{R}^d$: $\|\theta\|_1 \stackrel{\text{def}}{=} \sum_{j=1}^d |\theta_j|$, $\|\theta\|_0 \stackrel{\text{def}}{=} \sum_{j=1}^d \mathbf{1}_{\{|\theta_j|>0\}}$, and $\|\theta\|_\infty \stackrel{\text{def}}{=} \max_{1\leq j\leq d} |\theta_j|$.

For $\delta \in \Delta_d$, $\mu_{d,\delta}$ denotes the product measure on $\mathbb{R}^d$ defined as

$$\mu_{d,\delta}(\mathrm{d}\theta) \stackrel{\text{def}}{=} \prod_{j=1}^d \nu_{\delta_j}(\mathrm{d}\theta_j),$$

where $\nu_0(\mathrm{d}z)$ is the Dirac mass at 0, and $\nu_1(\mathrm{d}z)$ is the Lebesgue measure on $\mathbb{R}$. For $\theta, \theta' \in \mathbb{R}^d$, $\theta \cdot \theta' \in \mathbb{R}^d$ denotes the component-wise product of θ and θ': $(\theta \cdot \theta')_j = \theta_j \theta'_j$, $1 \leq j \leq d$. And for $\delta \in \Delta_d$, we define $\delta^c = 1 - \delta$, that is $\delta^c_j = 1 - \delta_j$, $1 \leq j \leq d$. For $\theta \in \mathbb{R}^d$, the sparsity structure of θ is the element $\delta \in \Delta_d$ defined as $\delta_j = \mathbf{1}_{\{|\theta_j|>0\}}$, $1 \leq j \leq d$.

Throughout the paper e denotes the Euler number. For $x \in \mathbb{R}$, the notation $\lceil x \rceil$ represents the smallest integer larger of equal to x. Finally, for $\theta \in \mathbb{R}^d$, and $A \subset \mathbb{R}^d$, $\theta + A \stackrel{\text{def}}{=} \{\theta + u,\ u \in A\}$.

2. Sparse Bayesian logistic regression

As a first application of the general theory developed in Section 4, we study the contraction behavior of a posterior distribution obtained from a high-dimensional logistic regression model, for large values of the sample size n and the dimension d. Suppose that $Z_1, \ldots, Z_n$ are independent 0-1 binary random variables and we consider

the model

$$\mathbb{P}(Z_i = 1) = \frac{e^{\langle x_i, \theta\rangle}}{1 + e^{\langle x_i, \theta\rangle}},$$

for a parameter $\theta \in \mathbb{R}^d$, where $x_i \in \mathbb{R}^d$ is a known vector of covariates. Writing $z = (z_1, \ldots, z_n)$, the likelihood function is then

$$q_{n,\theta}(z) = \exp\left(\sum_{i=1}^n z_i \langle x_i, \theta\rangle - g\left(\langle x_i, \theta\rangle\right)\right),$$

where

$$g(x) \stackrel{\text{def}}{=} \log(1 + e^x), \quad x \in \mathbb{R}.$$

Using the prior distribution given in (2)-(3), we consider the posterior distribution

$$\check{\Pi}_{n,d}(\mathrm{d}\theta|Z) \propto \exp\left(\sum_{i=1}^n Z_i \langle x_i, \theta\rangle - g\left(\langle x_i, \theta\rangle\right)\right) \sum_{\delta\in\Delta} \pi_\delta \left(\frac{\rho}{2}\right)^{\|\delta\|_1} e^{-\rho\|\theta\|_1} \mu_{d,\delta}(\mathrm{d}\theta). \quad (4)$$

We make the following assumption that implies H1.

B1. *$Z_1, \ldots, Z_n$ are independent* 0-1 *binary random variables, and there exist $\theta_\star \in \mathbb{R}^d$, $x_1, \ldots, x_n \in \mathbb{R}^d$, such that*

$$\mathbb{P}(Z_i = 1) = \frac{e^{\langle x_i, \theta_\star\rangle}}{1 + e^{\langle x_i, \theta_\star\rangle}},\ i = 1, \ldots, n.$$

Following Castillo et al. (2015), we specify the prior $\{\pi_\delta,\ \delta \in \Delta_d\}$ as follows.

B2. *For $\delta \in \Delta_d$ with $\|\delta\|_0 = s$, $\pi_\delta = g_s \binom{d}{s}^{-1}$, for a discrete distribution $\{g_s,\ 0 \le s \le d\}$, for which there exist positive universal constant c_1, c_2, $c_3 \ge c_4$ such that*

$$\frac{c_1}{d^{c_3}} g_{s-1} \le g_s \le \frac{c_2}{d^{c_4}} g_{s-1},\ s = 1, \ldots, d.$$

Remark 1. Castillo and van der Vaart (2012) has several examples of prior distributions that satisfy B2. For instance if, for some hyper-parameter $u > 1$, $\mathsf{q} \sim \mathbf{Beta}(1, d^u)$, and given q, we draw independently $\delta_j \sim \mathsf{Ber}(q)$, then the marginal distribution of δ in this case satisfies B2, with $c_1 = 1/2$, $c_2 = 1$, $c_3 = u$ and $c_4 = u - 1$.

Let $X \in \mathbb{R}^{n\times d}$ denote the design matrix, where the i-th row of X is given by the transpose of x_i. We shall write g', and $g^{(2)}$ to denote the first and second derivatives of g. Let $W \in \mathbb{R}^{n\times n}$ be the diagonal matrix with i-th diagonal entry given by

$$W_i = g^{(2)}\left(\langle x_i, \theta_\star\rangle\right), \quad i = 1, \ldots, n.$$

We define

$$\underline{\kappa}_1 \stackrel{\text{def}}{=} \inf\left\{\frac{\theta'(X'WX)\theta}{n\|\theta\|_2^2} :\ \theta \in \mathbb{R}^d \setminus \{0\},\ \|\theta \cdot \delta_\star^c\|_1 \le 7\|\theta \cdot \delta_\star\|_1\right\}.$$

For $s \in \{1, \dots, d\}$, we define

$$\bar{\kappa}(s) \stackrel{\text{def}}{=} \sup \left\{ \frac{\theta'(X'X)\theta}{n\|\theta\|_2^2} : \ 1 \le \|\theta\|_0 \le s \right\},$$
$$\text{and} \quad \underline{\kappa}_1(s) \stackrel{\text{def}}{=} \inf \left\{ \frac{\theta'(X'WX)\theta}{n\|\theta\|_2^2} : \ 1 \le \|\theta\|_0 \le s \right\}.$$

We choose the regularization parameter ρ in the prior distribution (3) as

$$\rho \stackrel{\text{def}}{=} 4\|X\|_\infty \sqrt{n \log(d)}, \tag{5}$$

where $\|X\|_\infty \stackrel{\text{def}}{=} \max_{i,j} |X_{ij}|$. We notice that $\kappa(1) \le \|X\|_\infty^2$, and $\underline{\kappa}_1(s) \le \|X\|_\infty^2/4$, for all $s \ge 1$.

Theorem 2. *Assume B1-B2, and $\underline{\kappa}_1 > 0$. Choose ρ as in (5). Set $s_\star \stackrel{\text{def}}{=} \|\theta_\star\|_0$,*

$$\zeta = s_\star + \frac{2}{c_4} + \frac{2}{c_4}\left(1 + \frac{64\|X\|_\infty^2}{\underline{\kappa}_1} + \frac{\bar{\kappa}(s_\star)}{64\|X\|_\infty^2 (\log(d))^2} + \frac{\log(4e)}{\log(d)}\right) s_\star. \tag{6}$$

(1) If d, n satisfy

$$d^{c_4} \ge 8c_2 \max(1, 2c_2), \quad n \ge \left(\frac{2^8}{3}\right)^2 \|X\|_\infty^4 \left(\frac{s_\star}{\underline{\kappa}_1}\right)^2 \log(d), \tag{7}$$

then

$$\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}\left(\left\{\theta \in \mathbb{R}^d : \ \|\theta\|_0 \ge \zeta\right\} | Z\right)\right] \le \frac{4}{d}.$$

(2) Set $\bar{s} \stackrel{\text{def}}{=} \lceil s_\star + \zeta \rceil$, and

$$r_{n,d} \stackrel{\text{def}}{=} \frac{16\|X\|_\infty}{\underline{\kappa}_1(\bar{s})} \sqrt{\frac{\bar{s}\log(d)}{n}}.$$

Choose $M_0 \ge \max(500, 1 + (c_3 + c_4/2)/8)$. If $\underline{\kappa}(\bar{s}) > 0$, and d, n satisfy (7) and

$$d \ge e(1 + c_1)/c_1, \quad n \ge (125)^2 \|X\|_\infty^4 \left(\frac{\bar{s}}{\underline{\kappa}_1(\bar{s})}\right)^2 \log(d),$$

then

$$\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}\left(\left\{\theta \in \mathbb{R}^d : \ \|\theta - \theta_\star\|_2 > M_0 r_{n,d}\right\} | Z\right)\right] \le \frac{12}{d}.$$

Proof. See Section 5.3. □

If the dimension d is large, then

$$\zeta \approx s_\star + \frac{2}{c_4} + \frac{2}{c_4}\left(1 + \frac{64\|X\|_\infty^2}{\underline{\kappa}_1}\right) s_\star.$$

Therefore, for design matrices X for which the restricted eigenvalue $\underline{\kappa}_1$ of the matrix $n^{-1}X'WX$ is not too small, Theorem 2 Part(1) implies that most of the probability mass of the posterior distribution is on sparse subsets of $\mathbb{R}^d$. If in addition the $(s_\star + \zeta)$-sparse smallest eigenvalue of $n^{-1}X'WX$ is well-behaved, then the rate of convergence of the posterior distribution (4) is $O\left(\sqrt{\frac{s_\star \log(d)}{n}}\right)$. The frequentist ℓ^1-penalized M-estimator for logistic regression has been analyzed by Negahban et al. (2012) (assuming a random design matrix X), and Li et al. (2014) (assuming a deterministic design matrix X), and is known to converge at the same rate, and under assumptions that are similar to those imposed above. Technically, our approach is closer to Li et al. (2014). The approach of Negahban et al. (2012) leads to slightly better conditions on the sample size n (they require n to increase linearly in $s_\star$, not quadratically, as in (7)), at the expense of more structure on the design matrix (X is assumed to have i.i.d. rows from a sub-Gaussian distribution and positive definite covariance).

Remark 3. It is interesting to observe that the contraction result given in Theorem 2 Part(2) holds, not in spite of the large dimension d, but because d is large. In other words, the result should be viewed as a form of concentration of measure phenomenon for $\check{\Pi}_{n,d}$ as $d \to \infty$. In particular, Theorem 2 Part(2) should not be applied to a fixed-dimension case in an attempt to recover standard (fixed d) Bayesian contraction results. Indeed, note that for d fixed, the prior distribution Π in (2-3) with ρ as in (5) converges weakly to a point-mass at 0 as $n \to \infty$, which is not a standard behavior of a prior in fixed-dimensional settings. However, with more appropriate prior assumptions Theorem 8 can be used to derive convergence rate results that would be applicable to the fixed-dimensional setting. We refer to Ghosh and Ramamoorthi (2003) (and the references therein) for a good presentation of finite-dimensional Bayesian asymptotics.

3. Quasi-Bayesian inference for high-dimensional binary graphical models

As another example, we consider the Bayesian analysis of high-dimensional binary graphical models (sometimes called Ising models). Let $\mathcal{M}_p$ be the space of real-valued $p \times p$ symmetric matrices. For $\theta \in \mathcal{M}_p$, let f_θ be the probability mass function defined on $\{0,1\}^p$ by

$$f_\theta(x_1, \ldots, x_p) = \frac{1}{Z_\theta} \exp\left(\sum_{j=1}^{p} \theta_{jj} x_j + \sum_{i<j} \theta_{ij} x_i x_j\right), \quad x_j \in \{0,1\},\ 1 \le j \le p, \qquad (8)$$

where Z_θ is the normalizing constant. We consider the problem of estimating θ under a sparsity assumption, from a matrix $Z \in \mathbb{R}^{n\times p}$ where each row of Z is an independent realization from $f_{\theta_\star}$ for some sparse $\theta_\star \in \mathcal{M}_p$. This problem has generated some literature in recent years (Banerjee et al. (2008); Höfling and Tibshirani (2009); Ravikumar et al. (2010); Atchadé (2014) and the references therein), all in the frequentist framework.

The Bayesian estimation of θ is significantly more challenging because the normalizing constant Z_θ are typically intractable, and this leads to posterior distributions that are doubly intractable. We note that there has been some recent progress on MCMC methods for doubly-intractable posterior distributions (see e.g. Lyne et al. (2015) and the references therein). However at the moment, these methods cannot handle high-dimensional parameters. In the frequentist literature cited above, the preferred approach for estimating θ is via penalized pseudo-likelihood maximization, which nicely side-steps the intractable normalizing constants issue. The quasi-Bayesian framework developed in this work can be used to combine these pseudo-likelihood functions with a prior distribution to produce quasi-Bayesian posterior distributions.

The most commonly used pseudo-likelihood function is obtained by taking the product of all the conditional densities in (8). This is an idea that goes back at least to Besag (1974). The resulting quasi-likelihood function is

$$\bar{q}_{n,\theta}(Z) = \prod_{j=1}^{p}\prod_{i=1}^{n} \frac{\exp\left(Z_{ij}\left(\theta_{jj} + \sum_{k\neq j}\theta_{kj}Z_{ik}\right)\right)}{\exp\left(\theta_{jj} + \sum_{k\neq j}\theta_{kj}Z_{ik}\right)}, \quad \theta \in \mathcal{M}_p.$$

Combined with a prior distribution Π on $\mathcal{M}_p$, this approach readily yields a quasi-posterior distribution on $\mathcal{M}_p$ that falls in the framework presented above. Note however that when p is large, say $p \geq 500$, the space $\mathcal{M}_p$ has dimension bigger than 10^5, and MCMC sampling from this quasi-posterior distribution becomes a daunting and time consuming task. One interesting idea is to break the symmetry and to consider the quasi-likelihood

$$q_{n,\theta}(Z) = \prod_{j=1}^{p}\prod_{i=1}^{n} \frac{\exp\left(Z_{ij}\left(\theta_{jj} + \sum_{k\neq j}\theta_{kj}Z_{ik}\right)\right)}{\exp\left(\theta_{jj} + \sum_{k\neq j}\theta_{kj}Z_{ik}\right)}, \quad \theta \in \mathbb{R}^{p\times p}. \tag{9}$$

Notice that the only difference between $\bar{q}_{n,\theta}$ and $q_{n,\theta}$ is that the symmetry constraint in θ is relaxed, that is the parameter space of $\theta \mapsto q_{n,\theta}(Z)$ is $\mathbb{R}^{p\times p}$, not $\mathcal{M}_p$. However this difference has a huge impact since now $q_{n,\theta}(Z)$ factorizes along the columns of θ. As a result, maximizing a penalized version of (9) is equivalent to solving p independent logistic regression (assuming a separable penalty), and this can be done efficiently in a parallel computing environment. This pseudo-likelihood approach

was popularized by the influential paper Meinshausen and Buhlmann (2006) in the Gaussian case, and extended to the Ising model by Ravikumar et al. (2010). In a recent work (Atchadé (2015)), the author extended this idea to the Bayesian analysis of large Gaussian graphical models, and analyzed the contraction of the resulting quasi-posterior distribution using Theorem 8. Here we extend the method to the Ising model.

Throughout this section, if $\theta \in \mathbb{R}^{p\times p}$, $\theta_{\cdot j} \in \mathbb{R}^p$ denotes the j-th column of θ. In view of the discussion above, and for a discrete probability distribution $\{\pi_\delta,\ \delta \in \Delta_p\}$ on Δ_p, and $\rho > 0$, we consider the quasi-posterior $\check{\Pi}_{n,d}$ on $\mathbb{R}^{p\times p}$ given by

$$\begin{aligned}\check{\Pi}_{n,d}(\mathrm{d}\theta|Z) &\propto q_{n,\theta}(Z)\prod_{j=1}^{p}\sum_{\delta\in\Delta_p}\pi_\delta\left(\frac{\rho}{2}\right)^{\|\delta\|_0} e^{-\rho\|\theta_{\cdot j}\|_1}\mu_{p,\delta}(\mathrm{d}\theta_{\cdot j}) \qquad (10)\\ &= \prod_{j=1}^{p}\check{\Pi}_{n,d,j}(\mathrm{d}\theta_{\cdot j}|Z)\quad .\end{aligned}$$

where $\check{\Pi}_{n,d,j}(\cdot|Z)$ is the probability measure on $\mathbb{R}^p$ given by

$$\check{\Pi}_{n,d,j}(\mathrm{d}u|Z) \propto \prod_{i=1}^{n}\frac{\exp\left(Z_{ij}\left(u_j+\sum_{k\neq j}u_kZ_{ik}\right)\right)}{\exp\left(u_j+\sum_{k\neq j}u_kZ_{ik}\right)}\sum_{\delta\in\Delta_p}\pi_\delta\left(\frac{\rho}{2}\right)^{\|\delta\|_0} e^{-\rho\|u\|_1}\mu_{p,\delta}(\mathrm{d}u).$$

Remark 4. One of the limitation of the approach is that the distribution $\check{\Pi}_{n,d}$ does not necessarily produce symmetric matrices. However, because of the contraction properties discussed below, typical realizations of $\check{\Pi}_{n,d}$ will be close to be symmetric. Furthermore, from a practical viewpoint, one can easily remedy a broken symmetry using various symmetrization rules as suggested for instance in Meinshausen and Buhlmann (2006).

It is clear from (10) that the quasi-posterior distribution $\check{\Pi}_{n,d}$ falls in the framework developed in Section 4, and we will use Theorem 8 to derive a bound on its contraction rate. We make the following assumptions.

C 1. *The rows of $Z \in \mathbb{R}^{n\times p}$ are independent $\{0,1\}^p$-valued random variables with common probability mass function $f_{\theta_\star}$, for some $\theta_\star \in \mathcal{M}_p$.*

We define

$$s_{\star j} \stackrel{\text{def}}{=} \|\theta_{\star\cdot j}\|_0, \quad \text{and} \quad s_\star \stackrel{\text{def}}{=} \max_{1\le j\le p} s_{\star j}.$$

Hence $s_\star$ is the maximum degree of the undirected graph encoded by $\theta_\star$. The sparsity structure of $\theta_\star$ is the matrix $\delta_\star \in \mathcal{M}_p$ defined as $\delta_{\star,jk} = \mathbf{1}_{\{|\theta_{\star jk}|>0\}}$. For $X \sim f_{\theta_\star}$, and

$1 \le j \le p$, define $X_{(j)} \stackrel{\text{def}}{=} (X_1, \ldots, X_{j-1}, 1, X_{j+1}, \ldots, X_p) \in \mathbb{R}^p$ (viewed as a column vector), and

$$\mathcal{H}^{(j)} \stackrel{\text{def}}{=} \mathbb{E}\left[g^{(2)}\left(\langle \theta_{\star \cdot j}, X_{(j)} \rangle\right) X_{(j)} X_{(j)}'\right].$$

The matrix $\mathcal{H}^{(j)}$ is the Fisher information matrix in the node j regression. We set

$$\underline{\kappa}_2(s) \stackrel{\text{def}}{=} \inf_{1 \le j \le p} \inf \left\{ \frac{u' \mathcal{H}^{(j)} u}{\|u\|_2^2},\ u \in \mathbb{R}^p \setminus \{0\},\ \ \|u\|_0 \le s \right\}, \quad \text{and}$$

$$\underline{\kappa}_2 \stackrel{\text{def}}{=} \inf_{1 \le j \le p} \inf \left\{ \frac{u' \mathcal{H}^{(j)} u}{\|u\|_2^2},\ \ u \in \mathbb{R}^p \setminus \{0\},\ \ \sum_{k:\ \delta_{\star kj} \ne 0} |u_k| \le \sum_{k:\ \delta_{\star kj} = 0} |u_k| \right\}. \quad (11)$$

The quantities $\underline{\kappa}_2(s)$ and $\underline{\kappa}_2$ are (the minimum over j of) restricted smallest eigenvalues of the matrices $\mathcal{H}^{(j)}$. We will work under the assumption that $\underline{\kappa}_2(s) > 0$ and $\underline{\kappa}_2 > 0$, for some well-chosen s. Admittedly, these assumptions are not easy to check in practice, particularly for discrete graphical models. But to our defense, we note that statistical inference for models with information singularity is still largely an under-developed topic, and it would be a daunting task to tackle this issue in the present context.

As above, we will also assume that

C 2. *The regularization parameter ρ in (10) is taken as*

$$\rho = 24\sqrt{n \log(p)}. \quad (12)$$

And the distribution $\{\pi_\delta,\ \delta \in \Delta_p\}$ used in (10) satisfies the following. For $\delta \in \Delta_p$ with $\|\delta\|_0 = s$, $\pi_\delta = g_s \binom{p}{s}^{-1}$, for a discrete distribution $\{g_s,\ 0 \le s \le p\}$, for which there exist positive universal constant c_1, c_2, $c_3 \ge c_4$ such that

$$\frac{c_1}{p^{c_3}} g_{s-1} \le g_s \le \frac{c_2}{p^{c_4}} g_{s-1},\ s = 1, \ldots, p.$$

To apply Theorem 8 and 9, we view $\mathbb{R}^{p \times p}$ as $\mathbb{R}^d$, with $d = p^2$, equipped with the Frobenius norm $\|\theta\|_{\mathsf{F}} \stackrel{\text{def}}{=} \sqrt{\mathsf{Tr}(\theta'\theta)}$, and inner product $\langle \theta, \vartheta \rangle_{\mathsf{F}} \stackrel{\text{def}}{=} \mathsf{Tr}(\theta'\vartheta)$, where $\mathsf{Tr}(\theta)$ denotes the trace of the matrix θ. Throughout this section, the norm $\|\cdot\|_2$ always denotes the Euclidean norm on $\mathbb{R}^p$. We will work with split cones (see Section 4 below for definition) of the form $\{\theta \in \mathbb{R}^{p \times p} :\ \|\theta_{\cdot j}\|_0 \le s_j,\ 1 \le j \le p\}$. First we show that if $\underline{\kappa}_2 > 0$, then the quasi-posterior distribution concentrates most of its mass on such split cones.

Theorem 5. *Assume C1-C2, and $\underline{\kappa}_2 > 0$. Set $\Theta \stackrel{\text{def}}{=} \{\theta \in \mathbb{R}^{p \times p} :\ \|\theta_{\cdot j}\|_0 \le \zeta_j,\ 1 \le j \le p\}$, where*

$$\zeta_j = s_{\star j} + \frac{4}{c_4} + \frac{2}{c_4}\left(1 + \frac{128}{\underline{\kappa}_2} + \frac{s_{\star j}}{64(\log(p))^2} + \frac{\log(4e)}{\log(p)}\right) s_{\star j}. \quad (13)$$

Then there exists universal finite positive constant A_1, A_2 such that if

$$p^{c_4} \geq 8c_2 \max(1, 2c_2), \quad \textit{and} \quad n \geq A_1 \left(\frac{s_\star}{\underline{\kappa}_2}\right)^2 \log(p), \tag{14}$$

then

$$\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}\left(\mathbb{R}^{d\times d} \setminus \Theta | Z\right)\right] \leq e^{-A_2 n} + \frac{4}{p}.$$

Proof. See Section 5.4 □

The following gives a bound on the contraction rate of $\check{\Pi}_{n,d}$.

Theorem 6. *Assume C1-C2, and $\underline{\kappa}_2 > 0$. For $1 \leq j \leq p$, set $\bar{s}_j \stackrel{\text{def}}{=} \lceil s_{\star j} + \zeta_j \rceil$, where ζ_j is as defined in (13), and $\bar{s} \stackrel{\text{def}}{=} \max_{1\leq j\leq p} \bar{s}_j$. Also define*

$$r_{n,p} \stackrel{\text{def}}{=} \frac{96}{\underline{\kappa}_2(\bar{s})} \sqrt{\left(\sum_{j=1}^{p} \bar{s}_j\right) \frac{\log(p)}{n}}.$$

Fix $M_0 \geq \max(500, 1 + (c_3 + c_4/2)/8)$. Then the constants A_1 and A_2 in Theorem 5 can be chosen such that if $\underline{\kappa}_2(\bar{s}) > 0$ and n, p satisfy (14) and

$$p \geq e(1 + c_1)/c_1, \quad \textit{and} \quad n \geq A_1 \left(\frac{1}{\underline{\kappa}(\bar{s})} \sum_{j=1}^{p} \bar{s}_j\right)^2 \log(p), \tag{15}$$

then

$$\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}\left(\left\{\theta \in \mathbb{R}^{d\times d} : \ \|\theta - \theta_\star\|_{\mathsf{F}} > M_0 r_{n,d}\right\} | Z\right)\right] \leq 2e^{-A_2 n} + \frac{12}{d}.$$

Proof. See Section 5.4 □

If p and n are large while $\underline{\kappa}_2$ remains bounded away from zero, Theorem 5 implies that the quasi-posterior distribution $\check{\Pi}_{n,d}$ puts high probability on matrices of $\mathbb{R}^{d\times d}$ with similar induced-graphs as $\theta_\star$, and Theorem 6 implies that in this case, the rate of convergence in the Frobenius norm is of order

$$O\left(\sqrt{\frac{(p + S)\log(p)}{n}}\right),$$

where $S \stackrel{\text{def}}{=} \sum_{j=1}^{p} s_{\star j}$ is twice the number of non-zero components of $\theta_\star$. This convergence rate matches up well against known results. For instance Sun and Zhang (2013) shows that in the Gaussian case, the convergence rate of a scaled g-Lasso estimator of θ based on the quasi-likelihood $q_{n,\theta}$ has convergence rate $O(s_\star\sqrt{\log(p)/n})$, in the spectral norm. We note however that the convergence in the Frobenius norm requires

a very high sample size, and the rate of convergence is slow. To get a faster rate we consider the norm

$$|||\theta||| \stackrel{\text{def}}{=} \max_{1\le j\le p} \|\theta_{\cdot j}\|_2, \quad \theta \in \mathbb{R}^{p\times p}.$$

The next result shows that the contraction rate of $\check{\Pi}_{n,d}$ in the norm $|||\cdot|||$ is

$$O\left(\sqrt{\frac{s_\star \log(p)}{n}}\right),$$

where $s_\star$ is the maximum degree of the graph defined by $\theta_\star$.

Theorem 7. *Assume C1-C2, and $\underline{\kappa}_2 > 0$. For $1 \le j \le p$, set $\bar{s}_j \stackrel{\text{def}}{=} \lceil s_{\star j} + \zeta_j \rceil$, where ζ_j is as defined in (13), and $\bar{s} \stackrel{\text{def}}{=} \max_{1\le j\le p} \bar{s}_j$. Also define*

$$r_{n,p} \stackrel{\text{def}}{=} \frac{32}{\underline{\kappa}_2(\bar{s})}\sqrt{\frac{\bar{s}\log(p)}{n}}.$$

Fix $M_0 \ge \max(500, 1 + (c_3 + c_4/2)/8)$. Then the constants A_1 and A_2 in Theorem 5 can be chosen such that if $\underline{\kappa}_2(\bar{s}) > 0$ and n, p satisfy (14) and

$$p \ge e(1+c_1)/c_1, \quad \textit{and} \quad n \ge A_1 \left(\frac{\bar{s}}{\underline{\kappa}(\bar{s})}\right)^2 \log(p),$$

then

$$\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}\left(\left\{\theta \in \mathbb{R}^{d\times d} : \ |||\theta - \theta_\star||| > M_0 r_{n,d}\right\} | Z\right)\right] \le 2e^{-A_2 n} + \frac{12}{d}.$$

Proof. See Section 5.4 □

4. Contraction rate of $\check{\Pi}_{n,d}$: general results

In this section we resume the general notation of the introduction, and we consider the QP distribution (1) on $\mathbb{R}^d$, with the prior distribution (2-3). Using the notation of Section 1.1, $\check{\Pi}_{n,d}$ can be written as

$$\check{\Pi}_{n,d}(\mathrm{d}\theta|Z) \propto q_{n,\theta}(Z) \sum_{\delta\in\Delta_d} \pi_\delta \left(\frac{\rho}{2}\right)^{\|\delta\|_0} e^{-\rho\|\theta\|_1} \mu_{d,\delta}(\mathrm{d}\theta). \tag{16}$$

We are interesting in the contraction behavior of $\check{\Pi}_{n,d}$ for large n, d. We take the usual frequentist view of Bayesian procedures by assuming the following.

H 1. *There exists $\theta_\star \in \mathbb{R}^d$ such that $Z \sim \mathbb{P}^{(n)}_{\theta_\star}(\mathrm{d}z) = f_{n,\theta_\star}(z)\mathrm{d}z$.*

We write $\mathbb{E}^{(n)}$ for the expectation operator with respect to $\mathbb{P}^{(n)}_{\theta_\star}(\mathrm{d}z)$. We also make the basic assumption that the quasi-likelihood function is smooth, and we use the notation $\nabla \log q_{n,u}(z)$ to denote the derivative of the map $\theta \mapsto \log q_{n,\theta}(z)$ at u. The j-th component of $\nabla \log q_{n,u}(z)$ is written as $(\nabla \log q_{n,u}(z))_j$.

H 2. *For all $z \in \mathcal{Z}^{(n)}$, the map $\theta \mapsto \log q_{n,\theta}(z)$ is differentiable.*

We consider the contraction properties of $\check{\Pi}_{n,d}$ towards $\theta_\star$. The results that we derive are non-asymptotic and can be useful to understand the behavior of a large class of quasi-posterior distributions as both the sample size n and the parameter dimension d grow. It seems natural to view $\check{\Pi}_{n,d}$ as a special type of mis-specified posterior distribution. This is the approach taken here, and we borrow ideas from the analysis of general Bayesian nonparametric missspecified models as developed by Kleijn and van der Vaart (2006). Although the results obtained by Kleijn and van der Vaart (2006) cannot be applied in our setting, these authors solve a general technical problem that plays a key role in our analysis: they prove the existence of test functions to test a given probability measure against a set of finite measure alternatives.

In the proofs, we also borrow a strategy developed mostly for the analysis of high-dimensional M-estimators, that consists in identifying a "good" subset $\mathcal{E}_n$ of the sample space $\mathcal{Z}^{(n)}$ on which the map $\theta \mapsto q_{n,\theta}(Z)$ has good curvature properties (see e.g. Negahban et al. (2012) for an excellent presentation of these ideas). Using this idea, the task at hand then boils down to controlling the probability of the set $\mathcal{E}_n$ and showing that $\check{\Pi}_{n,d}$ has good contraction properties when $Z \in \mathcal{E}_n$. To that end, and to shorten notation, we introduce the function

$$\mathcal{L}_{n,\theta}(z) \stackrel{\text{def}}{=} \log q_{n,\theta}(z) - \log q_{n,\theta_\star}(z) - \langle \nabla \log q_{n,\theta_\star}(z), \theta - \theta_\star \rangle \,, \quad \theta \in \mathbb{R}^d, \; z \in \mathcal{Z}^{(n)}.$$

This function plays a key role in statistical inference as it informs on the curvature of the objective function $\theta \mapsto \log q_{n,\theta}(Z)$ around $\theta_\star$. However, in high-dimensional settings, it is typically not realistic to assume that $\theta \mapsto \log q_{n,\theta}(Z)$ has good curvature on the entire parameter space $\mathbb{R}^d$. As well explained in Negahban et al. (2012), one should look at restrictions of $\mathcal{L}_{n,\theta}(z)$ to interesting subsets of $\mathbb{R}^d$.

We will use a rate function to express the curvature of $\theta \mapsto \log q_{n,\theta}(Z)$. Throughout the paper, a continuous function $\mathsf{r} : \; [0,\infty) \to [0,\infty)$ is a rate function if r is strictly increasing, $\mathsf{r}(0) = 0$, and $\lim_{x \downarrow 0} \mathsf{r}(x)/x = 0$. Given a rate function r, and $a \geq 0$, we define

$$\phi_{\mathsf{r}}(a) \stackrel{\text{def}}{=} \inf\{x > 0 : \; \mathsf{r}(z) \geq az, \text{ for all } z \geq x\}, \tag{17}$$

with the convention that $\inf \emptyset = +\infty$. The main example of a rate function is $\mathsf{r}(x) = \tau x^2$, for some $\tau > 0$. However, both examples below use $\mathsf{r}(x) = \tau x^2/(1+bx)$.

A non-empty subset Θ of $\mathbb{R}^d$ is a cone if for all $\lambda \geq 0$, and all $x \in \Theta$, $\lambda x \in \Theta$. We will say that a cone Θ is a split cone if $u \cdot x \in \Theta$ for all $x \in \Theta$, and all $u \in \{-1,1\}^d$ (we recall that the notation $u \cdot x$ denotes the component-by-component product). Split cones will serve here as generalizations of sparse subsets of $\mathbb{R}^d$. The archetype example of a split cone is the set of s-sparse elements: $\{\theta \in \mathbb{R}^d : \; \|\theta\|_0 \leq s\}$. However in some

problems, one might have to work with sparse elements that have some additional structure. This motivates the introduction of the split cones. A particularly important subset of $\mathbb{R}^d$ is the set of elements of $\mathbb{R}^d$ with the same sparsity structure as $\theta_\star$:

$$\Theta_\star \stackrel{\text{def}}{=} \Big\{\theta \in \mathbb{R}^d : \ \theta_j = 0 \text{ for all } j \text{ s.t. } \theta_{\star j} = 0\Big\}. \tag{18}$$

Given a rate function r, and a split cone $\Theta \subseteq \mathbb{R}^d$, we set

$$\check{\mathcal{E}}_{n,1}(\Theta, \mathsf{r}) \stackrel{\text{def}}{=} \left\{z \in \mathcal{Z}^{(n)} : \ \text{ for all } \theta \in \theta_\star + \Theta, \ \mathcal{L}_{n,\theta}(z) \le -\frac{1}{2}\mathsf{r}(\|\theta - \theta_\star\|_2)\right\}. \tag{19}$$

Here as in classical Bayesian asymptotics, in order to control the normalizing constant of the quasi-posterior distribution, we need a lower bound on the function $\theta \mapsto \mathcal{L}_{n,\theta}(z)$. Again, a restricted version will suffice. For $L \ge 0$, we set

$$\hat{\mathcal{E}}_{n,1}(\Theta, L) \stackrel{\text{def}}{=} \left\{z \in \mathcal{Z}^{(n)} : \ \text{ for all } \theta \in \theta_\star + \Theta, \ \mathcal{L}_{n,\theta}(z) \ge -\frac{L}{2}\|\theta - \theta_\star\|_2^2\right\}. \tag{20}$$

Finally, for $\lambda > 0$ we set

$$\mathcal{E}_{n,0}(\Theta, \lambda) \stackrel{\text{def}}{=} \left\{z \in \mathcal{Z}^{(n)} : \sup_{u \in \Theta, \, \|u\|_2 = 1} |\langle \nabla \log q_{n,\theta_\star}(z), u\rangle| \le \frac{\lambda}{2}\right\}. \tag{21}$$

The main idea behind these definitions is that on the event $\{Z \in \hat{\mathcal{E}}_{n,1}(\Theta, L) \cap \check{\mathcal{E}}_{n,1}(\Theta, \mathsf{r})\}$ the quasi-log-likelihod function $\theta \mapsto \log q_{n,\theta}(Z)$ has very nice curvature properties when restricted to the set $\theta_\star + \Theta$. The definition of $\mathcal{E}_{n,0}(\Theta, \lambda)$ implies that on the event $\{Z \in \mathcal{E}_{n,0}(\Theta, \lambda)\}$, $\theta_\star$ is close to the maximizer of the map $\theta \mapsto \log q_{n,\theta}(Z)$. Hence the set $\mathcal{E}_{n,0}(\Theta, \lambda) \cap \hat{\mathcal{E}}_{n,1}(\Theta, L) \cap \check{\mathcal{E}}_{n,1}(\Theta, \mathsf{r})$ is our example of a "good set", and on that set, we expect $\check{\Pi}_{n,d}(\cdot|Z)$ to have good concentration properties around $\theta_\star$, provided that the prior Π does not prevent it. This is the substance of the next result. Before stating the main theorem, we introduce few more notation.

For $M > 0$, let $\mathsf{B}_d(\Theta, M) \stackrel{\text{def}}{=} \{\theta \in \theta_\star + \Theta, \ \text{ s.t. } \ \|\theta - \theta_\star\|_2 \le M\}$. For $\epsilon > 0$, let $\mathsf{D}(\epsilon, \mathsf{B}_d(\Theta, M))$ denote the ϵ-packing number of the ball $\mathsf{B}_d(\Theta, M)$, defined as the maximal number of points in $\mathsf{B}_d(\Theta, M)$ such that the $\|\cdot\|_2$-distance between any pair of such points is at least ϵ.

Theorem 8. *Assume H1-H2 and let prior Π be as in (2)-(3). Fix $\bar{\lambda} \ge 0$, $\bar{L}$, a split cone $\bar{\Theta} \supseteq \Theta_\star$ and a rate function r such that $\bar{\epsilon} \stackrel{\text{def}}{=} \phi_{\mathsf{r}}(2\bar{\lambda})$ is finite. Let $s_\star \stackrel{\text{def}}{=} \|\theta_\star\|_0$, and $\mathcal{E}_n \stackrel{\text{def}}{=} \mathcal{E}_{n,0}(\bar{\Theta}, \bar{\lambda}) \cap \hat{\mathcal{E}}_{n,1}(\Theta_\star, \bar{L}) \cap \check{\mathcal{E}}_{n,1}(\bar{\Theta}, \mathsf{r})$. Choose $M_0 > 2$, and define the set*

$U(\bar{\epsilon}) \stackrel{\text{def}}{=} \left\{\theta \in \theta_\star + \bar{\Theta} : \ \|\theta - \theta_\star\|_2 > M_0\bar{\epsilon}\right\}$. *Then*

$$\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}\left(U(\bar{\epsilon})|Z\right)\right] \leq \mathbb{P}^{(n)}\left[Z \notin \mathcal{E}_n\right] + \sum_{j\geq 1} \mathsf{D}_j e^{-\frac{1}{8}\mathsf{r}(\frac{jM_0\bar{\epsilon}}{2})} + \frac{1}{\pi_{\delta_\star}}\left(\sum_{\delta\in\Delta_d} 2^{\|\delta\|_0}\pi_\delta\right)\left(1+\frac{\rho^2}{\bar{L}}\right)^{s_\star}\sum_{k\geq 1} e^{-\frac{1}{8}\mathsf{r}(\frac{kM_0\bar{\epsilon}}{2})}e^{3\rho c_0 kM_0\bar{\epsilon}}, \quad (22)$$

where $\mathsf{D}_j \stackrel{\text{def}}{=} \mathsf{D}\left(\frac{jM_0\bar{\epsilon}}{2}, \mathsf{B}_d(\bar{\Theta}, (j+1)M_0\bar{\epsilon})\right)$, *and* $c_0 \stackrel{\text{def}}{=} \sup_{u\in\bar{\Theta}}\sup_{v\in\bar{\Theta},\,\|v\|_2=1} |\,\langle \mathsf{sign}(u), v\rangle\,|$.

Proof. See Section 5.1. □

Theorem 8 decomposes the convergence rate of the quasi-posterior distribution into parts that can then be handled separately. For the term $\sum_{j\geq 1}\mathsf{D}_j e^{-\frac{1}{8}\mathsf{r}(\frac{jM_0\bar{\epsilon}}{2})}$ to be finite and small, it is sufficient and necessary that the rate term $\mathsf{r}(\frac{jM_0\bar{\epsilon}}{2})$ grows faster than the entropy term $\log \mathsf{D}_j$, as $j\to\infty$. In many cases, it is possible to derive bounds on the ϵ-packing numbers D_j, following for instance the arguments in Example 7.1 of Ghosal et al. (2000) (this is the approach taken in the examples below). Such bounds then allows us to work out simple conditions under which the term $\sum_{j\geq 1}\mathsf{D}_j e^{-\frac{1}{8}\mathsf{r}(\frac{jM_0\bar{\epsilon}}{2})}$ converges to zero with d. A similar argument applies to the last term of (22), the control of which boils down to comparing $\mathsf{r}(\frac{jM_0\bar{\epsilon}}{2})$ and $3\rho c_0 jM_0\bar{\epsilon}$, as $j\to\infty$.

To use the theorem, we will also need to find appropriate values for $\bar{\lambda}, \bar{L}$, split cone $\bar{\Theta}$ and a rate function r, such that the probability of the event $\mathcal{E}_n$ is high. Notice that the same type of events $\mathcal{E}_n$ appear in the analysis of high-dimensional M-estimators. Hence, it is then typically the case, as we will see in the examples, that one can take advantage of several existing results in the literature to deal with $\mathcal{E}_n$.

Finally, we note that

$$\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}\left(\left\{\theta\in\mathbb{R}^d : \ \|\theta-\theta_\star\|_2 > M_0\bar{\epsilon}\right\}|Z\right)\right] \leq \mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}\left(U(\bar{\epsilon})|Z\right)\right] + \mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}\left(\mathbb{R}^d \setminus (\theta_\star + \bar{\Theta})|Z\right)\right]. \quad (23)$$

The term on the left-side of (23) is typically the quantity of interest, whereas Theorem 8 gives a control only on the first term on the right-side of (23). Hence to use the theorem, the values of $\bar{\lambda}, \bar{L}$, the split cone $\bar{\Theta}$, and the rate function r, should also be such that the expectation $\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}\left(\theta_\star + \bar{\Theta}|Z\right)\right]$ is high. In other words, we need to establish by other means that the quasi-posterior distribution put high probability on $\theta_\star + \bar{\Theta}$. In connection with this, we have the following result which generalizes Theorem 1 of Castillo et al. (2015).

Theorem 9. *Assume H1-H2, and let the prior* Π *be as in (2)-(3). Fix* $\bar{L} > 0$*, and set* $\mathcal{E}_n \subseteq \hat{\mathcal{E}}_{n,1}(\Theta_\star, \bar{L})$*. Suppose that there exist* $A \geq 0, \beta > 0$ *such that for all* $\theta \in \mathbb{R}^d$*,*

$$\mathbb{E}^{(n)}\left[\boldsymbol{1}_{\mathcal{E}_n}(Z)\frac{q_{n,\theta}(Z)}{q_{n,\theta_\star}(Z)}\frac{e^{-\rho\|\theta\|_1}}{e^{-\rho\|\theta_\star\|_1}}\right] \leq e^A e^{-\beta\rho\|\theta-\theta_\star\|_1}. \tag{24}$$

Then for any measurable set $\mathcal{B} \subseteq \mathbb{R}^d$*,*

$$\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}\left(\mathcal{B}|Z\right)\right] \leq \mathbb{P}^{(n)}\left[Z \notin \mathcal{E}_n\right] + \frac{e^A}{\pi_{\delta_\star}}\left(1+\frac{\bar{L}}{\rho^2}\right)^{s_\star}\sum_{\delta\in\Delta(\mathcal{B})}\pi_\delta\left(\frac{\rho}{2}\right)^{\|\delta\|_0}\int_{\mathcal{B}} e^{-\beta\rho\|\theta-\theta_\star\|_1}\mu_{d,\delta}(\mathrm{d}\theta),$$

where $\Delta(\mathcal{B}) \stackrel{\text{def}}{=} \{\delta \in \Delta_d : \ \mu_{d,\delta}\left(\mathcal{B}\right) > 0\}$.

Proof. See Section 5.2. □

The condition (24) in Theorem 9 is in general difficult to check. It is possible to give a weaker, but easier to check version. To that end, let $\delta_\star$ denote the sparsity structure of $\theta_\star$: $\delta_{\star j} = \mathbf{1}_{\{|\theta_{\star j}|\neq 0\}}$, $1 \leq j \leq d$. We also define

$$\mathcal{N} \stackrel{\text{def}}{=} \left\{\theta \in \mathbb{R}^d : \ \theta \neq 0, \ \text{ and } \ \|\theta\cdot\delta_\star^c\|_1 \leq 7\|\theta\cdot\delta_\star\|_1\right\},$$

$$\check{\mathcal{E}}_{n,2}(\mathsf{r}) \stackrel{\text{def}}{=} \left\{z \in \mathcal{Z}^{(n)} : \ \text{ for all } \theta \in \theta_\star + \mathcal{N}, \ \mathcal{L}_{n,\theta}(z) \leq -\frac{1}{2}\mathsf{r}(\|\theta-\theta_\star\|_2)\right\},$$

$$\text{and} \quad \mathcal{E}_{n,0}(\lambda) \stackrel{\text{def}}{=} \left\{z \in \mathcal{Z}^{(n)} : \ \left\|\nabla \log q_{n,\theta_\star}(z)\right\|_\infty \leq \frac{\lambda}{2}\right\}.$$

Corollary 10. *Assume H1-H2, and let prior* Π *be as in (2)-(3). Set* $s_\star \stackrel{\text{def}}{=} \|\theta_\star\|_0$*,* $\mathcal{E}_n = \mathcal{E}_{n,0}(\rho) \cap \check{\mathcal{E}}_{n,2}(\mathsf{r}) \cap \hat{\mathcal{E}}_{n,1}(\Theta_\star, \bar{L})$*, for some constant* $\bar{L} \geq 0$*, and a rate function* r*. If the map* $\theta \mapsto \log q_{n,\theta}(z)$ *is concave for* $[\mathrm{d}z]$*-almost all* z*, then for any measurable set* $\mathcal{B} \subseteq \mathbb{R}^d$*,*

$$\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}\left(\mathcal{B}|Z\right)\right] \leq \mathbb{P}^{(n)}\left[Z \notin \mathcal{E}_n\right] + \frac{e^A}{\pi_{\delta_\star}}\left(1+\frac{\bar{L}}{\rho^2}\right)^{s_\star}\sum_{\delta\in\Delta(\mathcal{B})}\pi_\delta 4^{\|\delta\|_0}, \tag{25}$$

where $\Delta(\mathcal{B}) \stackrel{\text{def}}{=} \{\delta \in \Delta_d : \ \mu_{d,\delta}\left(\mathcal{B}\right) > 0\}$*, and* $A = -\frac{1}{2}\inf_{x>0}\left[\mathsf{r}(x) - 4\rho\sqrt{s_\star}x\right]$*, if* $\mathcal{N} \neq \emptyset$*, and* $A = 0$ *if* $\mathcal{N} = \emptyset$*.*

Proof. See Section 5.2.1. □

5. Proofs

5.1. **Proof of Theorem 8.** To improve readability we split the proof in three parts. The first part deals with the normalizing constant of the quasi-posterior distribution, the second part deals with the existence of test functions, and the proof of the theorem itself is given in the third part.

5.1.1. *On the normalizing constant of the quasi-posterior distribution.* The next lemma provides a lower bound on the normalizing constant of the quasi-posterior distribution (16), following an approach initially developed by Castillo et al. (2015).

Lemma 11. *Assume H1-H2. Fix $L \geq 0$, and a split cone $\Theta \supseteq \Theta_\star$. For all $z \in \hat{\mathcal{E}}_{n,1}(\Theta, L)$,*

$$\int_{\mathbb{R}^d} \frac{q_{n,\theta}(z)}{q_{n,\theta_\star}(z)} \Pi(\mathrm{d}\theta) \geq \pi_{\delta_\star} \left(\frac{\rho^2}{L+\rho^2} \right)^{s_\star} e^{-\rho\|\theta_\star\|_1}. \tag{26}$$

Proof. Using the definition of the prior Π, we have

$$\int_{\mathbb{R}^d} \frac{q_{n,\theta}(z)}{q_{n,\theta_\star}(z)} \Pi(\mathrm{d}\theta) \geq \pi_{\delta_\star} \left(\frac{\rho}{2} \right)^{s_\star} \int_{\theta_\star+\Theta_\star} \frac{q_{n,\theta}(z)}{q_{n,\theta_\star}(z)} e^{-\rho\|\theta\|_1} \mu_{d,\delta_\star}(\mathrm{d}\theta). \tag{27}$$

For $z \in \hat{\mathcal{E}}_{n,1}(\Theta, L)$, and $\theta \in \theta_\star + \Theta_\star \subseteq \theta_\star + \Theta$,

$$\log q_{n,\theta}(z) - \log q_{n,\theta_\star}(z) \geq \langle \nabla \log q_{n,\theta_\star}(z), \theta - \theta_\star \rangle - \frac{L}{2}\|\theta - \theta_\star\|_2^2.$$

Setting $\vartheta = \nabla \log q_{n,\theta_\star}(z)$, (27) then gives

$$\begin{aligned} \int_{\mathbb{R}^d} \frac{q_{n,\theta}(z)}{q_{n,\theta_\star}(z)} \Pi(\mathrm{d}\theta) \quad &\geq \quad \pi_{\delta_\star} \left(\frac{\rho}{2} \right)^{s_\star} e^{-\rho\|\theta_\star\|_1} \\ &\quad \times \int_{\theta_\star+\Theta_\star} e^{\langle \vartheta, \theta-\theta_\star \rangle - \frac{L}{2}\|\theta-\theta_\star\|_2^2} e^{-\rho\|\theta-\theta_\star\|_1} \mu_{d,\delta_\star}(\mathrm{d}\theta). \end{aligned} \tag{28}$$

We note that the support of the measure $\mu_{d,\theta_\star}$ is $\Theta_\star = \theta_\star + \Theta_\star$. Using this and the change of variable $\theta = \theta_\star + z$, we see that the integral on the right-hand size of (28) is

$$\int_{\mathbb{R}^d} e^{\langle \vartheta, z \rangle - \frac{L}{2}\|z\|_2^2 - \rho\|z\|_1} \mu_{d,\delta_\star}(\mathrm{d}z).$$

By Jensen's inequality,

$$\begin{aligned} \int_{\mathbb{R}^d} e^{\langle \vartheta, z \rangle} \frac{e^{-\frac{L}{2}\|z\|_2^2 - \rho\|z\|_1}}{\int_{\mathbb{R}^d} e^{-\frac{L}{2}\|u\|_2^2 - \rho\|u\|_1} \mu_{d,\delta_\star}(\mathrm{d}u)} \mu_{d,\delta_\star}(\mathrm{d}z) \\ \geq \exp\left(\int_{\mathbb{R}} \langle \vartheta, z \rangle \frac{e^{-\frac{L}{2}\|z\|_2^2 - \rho\|z\|_1}}{\int_{\mathbb{R}^d} e^{-\frac{L}{2}\|u\|_2^2 - \rho\|u\|_1} \mu_{d,\delta_\star}(\mathrm{d}u)} \mu_{d,\delta_\star}(\mathrm{d}z) \right) = 1. \end{aligned}$$

Using this, and going back to (27) we conclude that

$$\int_{\mathbb{R}^d} \frac{q_{n,\theta}(z)}{q_{n,\theta_\star}(z)} \Pi(\mathrm{d}\theta) \geq \pi_{\delta_\star} \left(\frac{\rho}{2}\right)^{s_\star} e^{-\rho\|\theta_\star\|_1} \int_{\mathbb{R}^d} e^{-\frac{L}{2}\|u\|_2^2 - \rho\|u\|_1} \mu_{d,\delta_\star}(\mathrm{d}u).$$

Now, note that

$$\int_{\mathbb{R}^d} e^{-\frac{L}{2}\|u\|_2^2 - \rho\|u\|_1} \mu_{d,\delta_\star}(\mathrm{d}u) = \left(\int_{\mathbb{R}} e^{-\rho|z| - \frac{L}{2}z^2} \mathrm{d}z\right)^{s_\star}.$$

It is easy to calculate that for $a \geq 0, b > 0$

$$\int_{\mathbb{R}} e^{-\frac{a}{2}u^2 - b|u|} \mathrm{d}u = \frac{2}{\sqrt{a}} \frac{1 - \Phi\left(\frac{b}{\sqrt{a}}\right)}{\phi\left(\frac{b}{\sqrt{a}}\right)}, \tag{29}$$

where ϕ is the density of the standard normal distribution, and Φ its cdf. The formula continues to hold by continuity at $a = 0$. The ratio $(1 - \Phi(z))/\phi(z)$ (known as Mills' ratio), satisfies

$$\frac{z}{1+z^2} \leq \frac{2}{z + \sqrt{z^2+4}} \leq \frac{1-\Phi(z)}{\phi(z)} \leq \frac{4}{3z + \sqrt{z^2+8}}, \quad z \geq 0, \tag{30}$$

see for instance Baricz (2008) Theorem 2.3 for a proof. We use this inequality and (29) to conclude that

$$\int_{\mathbb{R}} e^{-\rho|z| - \frac{L}{2}z^2} \mathrm{d}z \geq \frac{2\rho}{L + \rho^2},$$

and the lemma follows easily. □

5.1.2. *On the existence of test functions.* In this paragraph we establish the existence of test functions to test the density $f_{n,\theta_\star}$ against some mis-specified alternatives $Q_{n,\theta}$ defined below. The result is based on Lemma 6.1 of Kleijn and van der Vaart (2006), that we shall recall first for completeness. For any two integrable non-negative functions q_1, q_2 on $\mathcal{Z}^{(n)}$, and for $\alpha \in (0,1)$, the Hellinger transform $\mathcal{H}_\alpha(q_1, q_2)$ is defined as

$$\mathcal{H}_\alpha(q_1, q_2) \stackrel{\text{def}}{=} \int_{\mathcal{Z}^{(n)}} q_1^\alpha(z) q_2^{1-\alpha}(z) \mathrm{d}z.$$

Here we work with the case $\alpha = 1/2$, and set $\mathcal{H}(q_1, q_2) \stackrel{\text{def}}{=} \mathcal{H}_{1/2}(q_1, q_2)$.

Lemma 12 (Kleijn and van der Vaart (2006))**.** *Let p be a probability density function on $\mathcal{Z}^{(n)}$ and $\mathcal{Q}$ a class of non-negative integrable functions on $\mathcal{Z}^{(n)}$. Then*

$$\inf_\phi \sup_{q \in \mathcal{Q}} \left[\int_{\mathcal{Z}^{(n)}} \phi(z) p(z) \mathrm{d}z + \int_{\mathcal{Z}^{(n)}} (1 - \phi(z)) q(z) \mathrm{d}z\right] \leq \sup_{q \in \mathsf{conv}(\mathcal{Q})} \mathcal{H}(p, q), \tag{31}$$

where $\mathsf{conv}(\mathcal{Q})$ is the convex hull of $\mathcal{Q}$, and the infimum in (31) is taken over all test functions, that is all measurable functions $\phi : \mathcal{Z}^{(n)} \to [0,1]$. Furthermore, there exists a test function ϕ that attains the infimum.

To derive the test function for our quasi-likelihood setting, we will also need the following easy result.

Lemma 13. *Fix $\lambda \geq 0$, a split cone Θ, and a rate function r such that $\phi_{\mathsf{r}}(2\lambda)$ is finite. For any $\theta \in \theta_\star + \Theta$ such that $\|\theta - \theta_\star\|_2 \geq \phi_{\mathsf{r}}(2\lambda)$, we have*

$$\frac{q_{n,\theta}(z)}{q_{n,\theta_\star}(z)} \leq e^{-\frac{1}{4}\mathsf{r}(\|\theta-\theta_\star\|_2)},\ z \in \mathcal{E}_{n,0}(\Theta,\lambda) \cap \check{\mathcal{E}}_{n,1}(\Theta,\mathsf{r}).$$

Proof. For all $z \in \mathcal{Z}^{(n)}$, and $\theta \in \mathbb{R}^d$, we have

$$\frac{q_{n,\theta}(z)}{q_{n,\theta_\star}(z)} = \exp\left[\langle \nabla \log q_{n,\theta_\star}(z), \theta - \theta_\star\rangle + \mathcal{L}_{n,\theta}(z)\right].$$

By the definition of $\check{\mathcal{E}}_{n,1}(\Theta,\mathsf{r})$, for $\theta \in \theta_\star + \Theta$ and $z \in \check{\mathcal{E}}_{n,1}(\Theta,\mathsf{r})$, we have $\mathcal{L}_{n,\theta}(z) \leq -\frac{1}{2}\mathsf{r}(\|\theta - \theta_\star\|_2)$. And by the definition of $\mathcal{E}_{n,0}(\Theta,\lambda)$, for $z \in \mathcal{E}_{n,0}(\Theta,\lambda)$, and $\theta \in \theta_\star + \Theta$, we have

$$|\langle \nabla \log q_{n,\theta_\star}(z), \theta - \theta_\star\rangle| \leq \frac{\lambda}{2}\|\theta - \theta_\star\|_2.$$

Hence, for $z \in \mathcal{E}_{n,0}(\Theta,\lambda) \cap \check{\mathcal{E}}_{n,1}(\Theta,\mathsf{r})$, and $\theta \in \theta_\star + \Theta$,

$$\frac{q_{n,\theta}(z)}{q_{n,\theta_\star}(z)} \leq \exp\left[\frac{\lambda}{2}\|\theta - \theta_\star\|_2 - \frac{1}{2}\mathsf{r}(\|\theta - \theta_\star\|_2)\right]. \tag{32}$$

If in addition $\|\theta - \theta_\star\|_2 \geq \phi_{\mathsf{r}}(2\lambda)$, then from the properties of the rate function r, we have $2\lambda\|\theta - \theta_\star\|_2 - \mathsf{r}(\|\theta - \theta_\star\|_2) \leq 0$, and the result follows. □

Our main result on the existence of test functions follows. We recall that for $M > 0$, and a split cone Θ, $\mathsf{B}_d(\Theta, M) \stackrel{\text{def}}{=} \{\theta \in \theta_\star + \Theta : \text{ s.t. } \|\theta - \theta_\star\|_2 \leq M\}$, and for $\epsilon > 0$, $\mathsf{D}(\epsilon, \mathsf{B}_d(\Theta, M))$ denotes the ϵ-packing number of $\mathsf{B}_d(\Theta, M)$ in the norm $\|\cdot\|_2$.

Lemma 14. *Fix $\lambda \geq 0$, a split cone Θ, and a rate function r such that $\tilde{\epsilon} \stackrel{\text{def}}{=} \phi_{\mathsf{r}}(2\lambda)$ is finite. Set $\bar{\mathcal{E}}_n \stackrel{\text{def}}{=} \mathcal{E}_{n,0}(\Theta,\lambda) \cap \check{\mathcal{E}}_{n,1}(\Theta,\mathsf{r})$. For $\theta \in \mathbb{R}^d$, define the function*

$$Q_{n,\theta}(z) \stackrel{\text{def}}{=} \mathbf{1}_{\bar{\mathcal{E}}_n}(z)\frac{q_{n,\theta}(z)}{q_{n,\theta_\star}(z)} f_{n,\theta_\star}(z), \quad z \in \mathcal{Z}^{(n)}. \tag{33}$$

For any $M > 2$, there exists a measurable function $\phi : \mathcal{Z}^{(n)} \to [0,1]$ such that,

$$\mathbb{E}^{(n)}(\phi(Z)) \leq \sum_{j \geq 1} \mathsf{D}_j e^{-\frac{1}{8}\mathsf{r}(\frac{jM\tilde{\epsilon}}{2})},$$

where $\mathsf{D}_j \stackrel{\text{def}}{=} \mathsf{D}\left(\frac{jM\tilde{\epsilon}}{2}, \mathsf{B}_d(\Theta, (j+1)M\tilde{\epsilon})\right)$. Furthermore, for all $j \geq 1$, all $\theta \in \theta_\star + \Theta$ such that $\|\theta - \theta_\star\|_2 > jM\tilde{\epsilon}$,

$$\int_{\mathcal{Z}^{(n)}} (1 - \phi(z)) Q_{n,\theta}(z) \mathrm{d}z \leq e^{-\frac{1}{8}\mathsf{r}(\frac{jM\tilde{\epsilon}}{2})}.$$

Proof. First, notice that the function $z \mapsto Q_{n,\theta}(z)$ is integrable for all $\theta \in \theta_\star + \Theta$. Indeed, using (32) for any such θ, and for $z \in \bar{\mathcal{E}}_n$: $\frac{q_{n,\theta}(z)}{q_{n,\theta_\star}(z)} \leq \exp\left(\frac{\lambda}{2}\left\|\theta - \theta_\star\right\|_2\right)$. Hence,

$$\int_{\mathcal{Z}^{(n)}} Q_{n,\theta}(z)\mathrm{d}z = \int_{\bar{\mathcal{E}}_n} \frac{q_{n,\theta}(z)}{q_{n,\theta_\star}(z)} f_{n,\theta_\star}(z)\mathrm{d}z \leq e^{\frac{\lambda}{2}\left\|\theta-\theta_\star\right\|_2}.$$

Now, fix $\epsilon > 2\tilde{\epsilon}$ (where $\tilde{\epsilon} = \phi_{\mathsf{r}}(2\lambda)$), and fix $\theta \in \theta_\star + \Theta$ such that $\left\|\theta - \theta_\star\right\|_2 > \epsilon$. Set $\mathcal{P}_\theta \stackrel{\text{def}}{=} \{Q_{n,u} :\ u \in \theta_\star + \Theta \text{ and } \ \left\|u - \theta\right\|_2 \leq \epsilon/2\}$, and let $\mathsf{conv}(\mathcal{P}_\theta)$ denote the convex hull of the set $\mathcal{P}_\theta$. By Lemma 12 applied with $p = f_{n,\theta_\star}$, and $\mathcal{Q} = \mathcal{P}_\theta$, there exists a measurable function $\phi_\theta :\ \mathcal{Z}^{(n)} \to [0,1]$ such that

$$\mathbb{E}^{(n)}\left[\phi_\theta(Z)\right] \leq \sup_{Q \in \mathsf{conv}(\mathcal{P}_\theta)} \mathcal{H}(f_{n,\theta_\star}, Q)$$
$$\text{and} \quad \sup_{Q\in\mathcal{P}_\theta} \int_{\mathcal{Z}^{(n)}} (1-\phi_\theta(z))Q(z)\mathrm{d}z \leq \sup_{Q\in\mathsf{conv}(\mathcal{P}_\theta)} \mathcal{H}(f_{n,\theta_\star}, Q). \quad (34)$$

Any $Q \in \mathsf{conv}(\mathcal{P}_\theta)$ can be written as a finite convex combination $Q = \sum_j \alpha_j Q_{n,u_j}$ where $\alpha_j \geq 0$, $\sum_j \alpha_j = 1$, $u \in \theta_\star + \Theta$, and $\left\|u_j - \theta\right\|_2 \leq \epsilon/2$. However, since $\left\|\theta - \theta_\star\right\|_2 > \epsilon$, and $\left\|u_j - \theta\right\|_2 \leq \epsilon/2$, we see that $\left\|u_j - \theta_\star\right\|_2 > \epsilon/2 > \tilde{\epsilon}$. Hence, using Lemma 13 and the definition of the Hellinger transform, we have

$$\mathcal{H}(f_{n,\theta_\star}, Q) = \int_{\mathcal{Z}^{(n)}} \sqrt{\sum_j \alpha_j 1_{\bar{\mathcal{E}}_n}(z) \frac{q_{n,u_j}(z)}{q_{n,\theta_\star}(z)}} f_{n,\theta_\star}(z)\mathrm{d}z \leq \sqrt{\sum_j \alpha_j e^{-\frac{1}{4}\mathsf{r}(\left\|u_j-\theta_\star\right\|_2)}}.$$

Hence (34) becomes

$$\mathbb{E}^{(n)}\left[\phi_\theta(Z)\right] \ \leq\ e^{-\frac{1}{8}\mathsf{r}(\frac{\epsilon}{2})} \qquad \text{and} \qquad \sup_{Q\in\mathcal{P}_\theta} \int_{\mathcal{Z}^{(n)}} (1-\phi_\theta(z))Q(z)\mathrm{d}z \ \leq\ e^{-\frac{1}{8}\mathsf{r}(\frac{\epsilon}{2})}. \quad (35)$$

Now, given $M > 2$, we write $\mathsf{B}_d(\Theta, M) = \cup_{j\geq 1}\mathsf{B}(j)$, where

$$\mathsf{B}(j) = \{\theta \in \theta_\star + \Theta, \text{ s.t. } \ jM\tilde{\epsilon} < \left\|\theta - \theta_\star\right\|_2 \leq (j+1)M\tilde{\epsilon}\}.$$

For each $j \geq 1$, let $\mathcal{S}_j$ be a maximal $(jM\tilde{\epsilon}/2)$-separated points in $\mathsf{B}(j)$. For each j for which $\mathsf{B}(j) \neq \emptyset$, and each point $\theta_k \in \mathcal{S}_j$ we can construct a test function ϕ_{θ_k} as above, with $\epsilon = jM\tilde{\epsilon}$. Then we set

$$\phi = \sup_{j\geq 1} \max_{\theta_k \in \mathcal{S}_j} \phi_{\theta_k},$$

where the supremum in j is over the indexes for which $\mathsf{B}(j) \neq \emptyset$. Now, any $\theta \in \theta_\star + \Theta$ such that $\left\|\theta - \theta_\star\right\|_2 > jM\tilde{\epsilon}$ will be within $iM\tilde{\epsilon}/2$ of a point θ_k in $\mathcal{S}_i$ for some $i \geq j$. Hence by (35), for any such θ,

$$\int_{\mathcal{Z}^{(n)}} (1-\phi(z))Q_{n,\theta}(z)\mathrm{d}z \leq \int_{\mathcal{Z}^{(n)}} (1-\phi_{\theta_k}(z))Q_{n,\theta}(z)\mathrm{d}z \leq e^{-\frac{1}{8}\mathsf{r}(\frac{jM\tilde{\epsilon}}{2})}.$$

Notice that the size of $\mathcal{S}_j$ is upper bounded by D_j. Using this and (35), we get

$$\mathbb{E}^{(n)}\left[\phi(Z)\right] \leq \sum_{j\geq 1} \mathsf{D}_j e^{-\frac{1}{8}\mathsf{r}(\frac{jM\bar{\epsilon}}{2})},$$

which proves the lemma. □

5.1.3. *Proof of the theorem.* Define $U(\bar{\epsilon}) \stackrel{\text{def}}{=} \{\theta \in \theta_\star + \bar{\Theta} : \ \|\theta - \theta_\star\|_2 > M_0\bar{\epsilon}\}$. We apply Lemma 14 with $\lambda = \bar{\lambda}$, $\Theta = \bar{\Theta}$, the rate function r and with $M = M_0 > 2$. Notice $\bar{\epsilon} = \phi_{\mathsf{r}}(2\bar{\lambda})$ is called $\tilde{\epsilon}$ in Lemma 14. By Lemma 14 there exists a measurable functions $\phi : \ \mathcal{Z}^{(n)} \to [0,1]$ such that

$$\mathbb{E}^{(n)}\left[\phi(Z)\right] \leq \sum_{j\geq 1} \mathsf{D}_j e^{-\frac{1}{8}\mathsf{r}(\frac{jM_0\bar{\epsilon}}{2})}, \tag{36}$$

where $\mathsf{D}_j \stackrel{\text{def}}{=} \mathsf{D}\left(\frac{jM_0\bar{\epsilon}}{2}, \mathsf{B}_d(\bar{\Theta}, (j+1)M_0\bar{\epsilon})\right)$. Using the test function ϕ, we have

$$\check{\Pi}_{n,d}\left(U(\bar{\epsilon})|Z\right) \leq \phi(Z) + (1-\phi(Z))\check{\Pi}_{n,d}\left(U(\bar{\epsilon})|Z\right).$$

In view of (36), it remains only to control the expectation of $(1-\phi(Z))\check{\Pi}_{n,d}\left(U(\bar{\epsilon})|Z\right)$. To do so, we set $\bar{\mathcal{E}}_n \stackrel{\text{def}}{=} \mathcal{E}_{n,0}(\bar{\Theta}, \bar{\lambda}) \cap \check{\mathcal{E}}_{n,1}(\bar{\Theta}, \mathsf{r})$, so that $\mathcal{E}_n \subseteq \bar{\mathcal{E}}_n \cap \hat{\mathcal{E}}_{n,1}(\bar{\Theta}, \bar{L})$, and use Lemma 11 and Fubini's theorem to write

$$\begin{aligned}
\mathbb{E}^{(n)}\left[(1-\phi(Z))\check{\Pi}_{n,d}\left(U(\bar{\epsilon})|Z\right)\right] &= \mathbb{E}^{(n)}\left[(1-\phi(Z))\frac{\int_{U(\bar{\epsilon})}\frac{q_{n,\theta}(Z)}{q_{n,\theta_\star}(Z)}\Pi(\mathrm{d}\theta)}{\int\frac{q_{n,\theta}(Z)}{q_{n,\theta_\star}(Z)}\Pi(\mathrm{d}\theta)}\right] \\
&\leq \mathbb{P}^{(n)}\left(Z \notin \mathcal{E}_n\right) + \frac{1}{\pi_{\delta_\star}}\left(1+\frac{\rho^2}{\bar{L}}\right)^{s_\star} e^{\rho\|\theta_\star\|_1} \\
&\qquad \times \int_{U(\bar{\epsilon})} \mathbb{E}^{(n)}\left[\mathbf{1}_{\bar{\mathcal{E}}_n}(Z)(1-\phi(Z))\frac{q_{n,\theta}(Z)}{q_{n,\theta_\star}(Z)}\right]\Pi(\mathrm{d}\theta). \qquad (37)
\end{aligned}$$

We split $U(\bar{\epsilon})$ as $U(\bar{\epsilon}) = \cup_{j\geq 1}\mathsf{B}(j)$, where

$$\mathsf{B}(j) = \{\theta \in \theta_\star + \bar{\Theta} \text{ s.t. } \ jM_0\bar{\epsilon} < \|\theta - \theta_\star\|_2 \leq (1+j)M_0\bar{\epsilon}\}.$$

Therefore, and using the notation of Lemma 14, the integral in (37) is

$$\begin{aligned}
&\int_{U_1(\bar{\epsilon})} \mathbb{E}^{(n)}\left[\mathbf{1}_{\bar{\mathcal{E}}_n}(Z)(1-\phi(Z))\frac{q_{n,\theta}(Z)}{q_{n,\theta_\star}(Z)}\right]\Pi(\mathrm{d}\theta) \\
&\qquad = \sum_{j\geq 1}\int_{\mathsf{B}(j)}\left[\int_{\mathcal{Z}^{(n)}}(1-\phi(z))Q_{n,\theta}(z)\mathrm{d}z\right]\Pi(\mathrm{d}\theta) \leq \sum_{j\geq 1} e^{-\frac{1}{8}\mathsf{r}(\frac{jM_0\bar{\epsilon}}{2})}\Pi(\mathsf{B}(j)).
\end{aligned}$$

From the prior Π, we have

$$e^{\rho\|\theta_\star\|_1}\Pi(\mathsf{B}(j)) = \sum_{\delta\in\Delta_d}\pi_\delta\left(\frac{\rho}{2}\right)^{\|\delta\|_0}\int_{\mathsf{B}(j)} e^{\rho(\|\theta_\star\|_1 - \|\theta\|_1)}\mu_{d,\delta}(\mathrm{d}\theta).$$

and for $\theta \in \mathsf{B}(j)$,

$$\rho(\|\theta_\star\|_1 - \|\theta\|_1) \le \rho\|\theta - \theta_\star\|_1 \le -\frac{\rho}{2}\|\theta - \theta_\star\|_1 + \frac{3}{2}\rho\|\theta - \theta_\star\|_1$$
$$\le -\frac{\rho}{2}\|\theta - \theta_\star\|_1 + \frac{3}{2}\rho c_0 \|\theta - \theta_\star\|_2 \le -\frac{\rho}{2}\|\theta - \theta_\star\|_1 + 3\rho c_0 j M_0 \bar{\epsilon}$$

where $c_0 = \sup_{u\in\bar{\Theta}} \sup_{v\in\bar{\Theta},\, \|v\|_2=1} |\, \langle \mathsf{sign}(u), v\rangle \,|$. Hence

$$\begin{aligned} e^{\rho\|\theta_\star\|_1}\Pi(\mathsf{B}(j)) &\le e^{3\rho c_0 j M_0\bar{\epsilon}} \sum_{\delta\in\Delta_d} \pi_\delta \left(\frac{\rho}{2}\right)^{\|\delta\|_0} \int_{\mathsf{B}(j)} e^{-\frac{\rho}{2}\|\theta-\theta_\star\|_1} \mu_{d,\delta}(\mathrm{d}\theta), \\ &\le e^{3\rho c_0 j M_0\bar{\epsilon}} \sum_{\delta\in\Delta_d} \pi_\delta \left(\frac{\rho}{2}\right)^{\|\delta\|_0} \left(\int_{\mathbb{R}} e^{-\frac{\rho}{2}|z|}\mathrm{d}z\right)^{\|\delta\|_0}, \\ &= e^{3\rho c_0 j M_0\bar{\epsilon}} \sum_{\delta\in\Delta_d} \pi_\delta 2^{\|\delta\|_0}. \end{aligned}$$

Therefore, the second term on the right-hand side of (37) is upper bounded by

$$\frac{1}{\pi_{\delta_\star}} \left(\sum_{\delta\in\Delta_d} \pi_\delta 2^{\|\delta\|_0}\right) \left(1 + \frac{\rho^2}{\bar{L}}\right)^{s_\star} \sum_{k\ge 1} e^{-\frac{1}{8}\mathsf{r}(\frac{kM_0\bar{\epsilon}}{2})} e^{3\rho c_0 k M_0\bar{\epsilon}}.$$

This ends the proof. □

5.2. **Proof of Theorem 9 and Corollary 10.** Let $\Delta(\mathcal{B}) \stackrel{\text{def}}{=} \{\delta \in \Delta_d :\ \mu_{d,\delta}(\mathcal{B}) > 0\}$. We have

$$\mathbb{E}^{(n)}\left(\check{\Pi}_{n,d}(\mathcal{B}|Z)\right) \le \mathbb{P}^{(n)}\left(Z \notin \mathcal{E}_n\right) + T,$$

where $T = \mathbb{E}^{(n)}\left[\mathbf{1}_{\mathcal{E}_n}(Z) \frac{\int_{\mathcal{B}} \frac{q_{n,\theta}(Z)}{q_{n,\theta_\star}(Z)}\Pi(\mathrm{d}\theta)}{\int_{\mathbb{R}^d} \frac{q_{n,\theta}(Z)}{q_{n,\theta_\star}(Z)}\Pi(\mathrm{d}\theta)}\right]$. We use Lemma 11, and Fubini's theorem to write

$$\begin{aligned} T &\le \frac{1}{\pi_{\delta_\star}} \left(1 + \frac{\bar{L}}{\rho^2}\right)^{s_\star} e^{\rho\|\theta_\star\|_1} \mathbb{E}^{(n)}\left[\mathbf{1}_{\mathcal{E}_n}(Z) \int_{\mathcal{B}} \frac{q_{n,\theta}(Z)}{q_{n,\theta_\star}(Z)}\Pi(\mathrm{d}\theta)\right] \\ &= \frac{1}{\pi_{\delta_\star}} \left(1 + \frac{\bar{L}}{\rho^2}\right)^{s_\star} \\ &\quad \times \sum_{\delta\in\Delta(\mathcal{B})} \pi_\delta \left(\frac{\rho}{2}\right)^{\|\delta\|_0} \int_{\mathcal{B}} \mathbb{E}^{(n)}\left[\mathbf{1}_{\mathcal{E}_n}(Z) \frac{q_{n,\theta}(Z)}{q_{n,\theta_\star}(Z)} \frac{e^{-\rho\|\theta\|_1}}{e^{-\rho\|\theta_\star\|_1}}\right] \mu_{d,\delta}(\mathrm{d}\theta). \end{aligned} \tag{38}$$

Using (24) which gives a bound on the inner expectation in (38), we conclude that

$$T \le \frac{e^A}{\pi_{\delta_\star}} \left(1 + \frac{\bar{L}}{\rho^2}\right)^{s_\star} \sum_{\delta\in\Delta(\mathcal{B})} \pi_\delta \left(\frac{\rho}{2}\right)^{\|\delta\|_0} \int_{\mathcal{B}} e^{-\beta\rho\|\theta-\theta_\star\|_1} \mu_{d,\delta}(\mathrm{d}\theta),$$

hence, the theorem. □

5.2.1. *Proof of Corollary 10.* Firstly, we need to check that under the conditions of the corollary, (24) holds. Note that to check (24), we only need to check (24) for $\theta \neq \theta_\star$. For $z \in \mathcal{E}_n \subseteq \mathcal{E}_{n,0}(\rho)$, and $\theta \in \mathbb{R}^d$, we have

$$\begin{aligned}\frac{q_{n,\theta}(z)}{q_{n,\theta_\star}(z)} &= \exp\left[\langle \nabla \log q_{n,\theta_\star}(z), \theta - \theta_\star\rangle + \mathcal{L}_{n,\theta}(z)\right], \\ &\leq \exp\left[\frac{\rho}{2}\|\theta - \theta_\star\|_1 + \mathcal{L}_{n,\theta}(z)\right].\end{aligned}$$

It follows that for all $\theta \in \mathbb{R}^d$,

$$\mathbb{E}^{(n)}\left[\mathbf{1}_{\mathcal{E}_n}(Z)\frac{q_{n,\theta}(Z)}{q_{n,\theta_\star}(Z)}\frac{e^{\rho\|\theta\|_1}}{e^{-\rho\|\theta_\star\|_1}}\right] \leq e^{B(\theta)}\mathbb{E}^{(n)}\left[\mathbf{1}_{\mathcal{E}_n}(Z)\exp\left(\mathcal{L}_{n,\theta}(Z)\right)\right]. \tag{39}$$

where

$$B(\theta) \stackrel{\text{def}}{=} \frac{\rho}{2}\|\theta - \theta_\star\|_1 + \rho(\|\theta_\star\|_1 - \|\theta\|_1).$$

We then write

$$\begin{aligned}\|\theta_\star\|_1 + \frac{1}{2}\|\theta - \theta_\star\|_1 &= \|\theta_\star\|_1 + \frac{1}{2}\|\theta\cdot\delta_\star^c\|_1 + \frac{1}{2}\|(\theta-\theta_\star)\cdot\delta_\star\|_1 \\ &\leq \|\theta\|_1 - \frac{1}{2}\|\theta\cdot\delta_\star^c\|_1 + \frac{3}{2}\|(\theta-\theta_\star)\cdot\delta_\star\|_1. \end{aligned}\tag{40}$$

Using this bound in the expression of $B(\theta)$ shows that if $\theta \notin \theta_\star + \mathcal{N}$, then we have

$$\begin{aligned}B(\theta) &\leq -\frac{\rho}{2}\|\theta\cdot\delta_\star^c\|_1 + \frac{3\rho}{2}\|(\theta-\theta_\star)\cdot\delta_\star\|_1 \qquad (41)\\ &\leq -\frac{\rho}{4}\|\theta-\theta_\star\|_1.\end{aligned}$$

This bound together with the fact that the expectation on the right-side of (39) is always smaller or equal to 1 (which follows from the concaveness assumption) show that when $\theta \notin \theta_\star + \mathcal{N}$,

$$\mathbb{E}^{(n)}\left[\mathbf{1}_{\mathcal{E}_n}(Z)\frac{q_{n,\theta}(Z)}{q_{n,\theta_\star}(Z)}\frac{e^{-\rho\|\theta\|_1}}{e^{-\rho\|\theta_\star\|_1}}\right] \leq e^{-\frac{\rho}{4}\|\theta-\theta_\star\|_1}.$$

Now, consider the case where $\mathcal{N} \neq \emptyset$, and $\theta - \theta_\star \in \mathcal{N}$. In that case, the definition of the set $\check{\mathcal{E}}_{n,2}(\mathsf{r})$ and (39) yield

$$\mathbb{E}^{(n)}\left[\mathbf{1}_{\mathcal{E}_n}(Z)\frac{q_{n,\theta}(Z)}{q_{n,\theta_\star}(Z)}\frac{e^{-\rho\|\theta\|_1}}{e^{-\rho\|\theta_\star\|_1}}\right] \leq e^{B(\theta)-\frac{1}{2}\mathsf{r}(\|\theta-\theta_\star\|_2)}.$$

From (41),

$$B(\theta) - \frac{1}{2}\mathsf{r}(\|\theta-\theta_\star\|_2) \leq -\frac{\rho}{2}\|\theta-\theta_\star\|_1 + 2\rho\|(\theta-\theta_\star)\cdot\delta_\star\|_1 - \frac{1}{2}\mathsf{r}(\|\theta-\theta_\star\|_2),$$

and

$$\begin{aligned} 2\rho\|(\theta-\theta_\star)\cdot\delta_\star\|_1 - \frac{1}{2}\mathsf{r}(\|\theta-\theta_\star\|_2) &\leq 2\rho\sqrt{s_\star}\|\theta-\theta_\star\|_2 - \frac{1}{2}\mathsf{r}(\|\theta-\theta_\star\|_2), \\ &\leq -\frac{1}{2}\left[\mathsf{r}(\|\theta-\theta_\star\|_2) - 4\rho\sqrt{s_\star}\|\theta-\theta_\star\|_2\right] \\ &\leq -\frac{1}{2}\inf_{x>0}\left[\mathsf{r}(x) - 4\rho s_\star^{1/2}x\right]. \end{aligned}$$

Therefore, when $\theta \neq \theta_\star$, and $\theta \in \theta_\star + \mathcal{N}$, we have

$$\mathbb{E}^{(n)}\left[\mathbf{1}_{\mathcal{E}_n}(Z)\frac{q_{n,\theta}(Z)}{q_{n,\theta_\star}(Z)}\frac{e^{-\rho\|\theta\|_1}}{e^{-\rho\|\theta_\star\|_1}}\right] \leq e^A e^{-\frac{\rho}{2}\|\theta-\theta_\star\|_1},$$

where $A = -\frac{1}{2}\inf_{x>0}\left[\mathsf{r}(x) - 4\rho s_\star^{1/2}x\right]$. Note that $A > 0$, since $\lim_{x\downarrow 0}\mathsf{r}(x)/x = 0$. This proves that (24) holds with A and $\beta = 1/4$. Secondly, for $\delta \in \Delta(\mathcal{B})$, we will make use of the bound

$$\left(\frac{\rho}{2}\right)^{\|\delta\|_0}\int_{\mathcal{B}} e^{-\beta\rho\|\theta-\theta_\star\|_1}\mu_{d,\delta}(\mathrm{d}\theta) \leq \left(\frac{\rho}{2}\right)^{\|\delta\|_0}\left(\int_{\mathbb{R}} e^{-\beta\rho|z|}\mathrm{d}z\right)^{\|\delta\|_0} = \left(\frac{1}{\beta}\right)^{\|\delta\|_0}.$$

This proves the result. □

5.3. **Proof of Theorem 2.** We prove Theorem 2 by applying Corollary 10 and Theorem 9. Clearly, B1 implies H1, and H2 trivially holds true. Furthermore the function $\mathcal{L}_{n,\theta}$ is given by

$$\mathcal{L}_{n,\theta}(z) = -\sum_{i=1}^n g(\langle x_i,\theta\rangle) - g(\langle x_i,\theta_\star\rangle) - g'(\langle x_i,\theta_\star\rangle)\langle x_i,\theta-\theta_\star\rangle,$$

which does not depend on z. To control this term, we will rely on a nice self-concordant properties of the logistic function $g(x) = \log(1+e^x)$ developed by Bach (2010) Lemma 1, which states that for all $x_0, u \in \mathbb{R}$,

$$\begin{aligned} g^{(2)}(x_0)\left(e^{-|u|} + |u| - 1\right) \leq g(x_0+u) - g(x_0) - g'(x_0)u \\ \leq g^{(2)}(x_0)\left(e^{|u|} - |u| - 1\right). \end{aligned} \quad (42)$$

Proof of Part(1). We shall apply Corollary 10. Clearly, $\theta \mapsto \log q_{n,\theta}(z)$ is concave for all $z \in \{0,1\}^n$. We define $H(x) \stackrel{\text{def}}{=} e^{-x} + x - 1$. It can be checked that H satisfies

$$H(x) \geq \frac{x^2}{2+x}, \quad x \geq 0. \quad (43)$$

This holds because $(2+x)H(x) - x^2 = (2+x)e^{-x} + x - 2$, the derivative of which is $1 - \frac{x+1}{e^x} \geq 0$, for all $x \geq 0$. Using (42), we get

$$\mathcal{L}_{n,\theta}(z) \leq -\sum_{i=1}^{n} g^{(2)}\left(\langle x_i, \theta_\star\rangle\right) H\left(|\,\langle x_i, \theta - \theta_\star\rangle\,|\right).$$

Furthermore, for $\theta - \theta_\star \in \mathcal{N}$, we have

$$|\langle x_i, \theta - \theta_\star\rangle| \leq \|X\|_\infty \|\theta - \theta_\star\|_1 \leq 8\|X\|_\infty s_\star^{1/2}\|\theta - \theta_\star\|_2.$$

Using this, (43), and the definition of $\underline{\kappa}_1$, we get for all $z \in \{0,1\}^n$,

$$\begin{aligned}
\mathcal{L}_{n,\theta}(z) &\leq -\frac{n}{2 + \max_i |\,\langle x_i, \theta - \theta_\star\rangle\,|}(\theta - \theta_\star)'\frac{X'WX}{n}(\theta - \theta_\star) \\
&\leq -\frac{n\underline{\kappa}_1\|\theta - \theta_\star\|_2^2}{2 + 8\sqrt{s_\star}\|X\|_\infty\|\theta - \theta_\star\|_2}, \\
&= -\frac{1}{2}\mathsf{r}(\|\theta - \theta_\star\|_2), \qquad (44)
\end{aligned}$$

where $\mathsf{r}(x) = n\underline{\kappa}_1 x^2/(1 + 4\sqrt{s_\star}\|X\|_\infty x)$. Hence, with this particular choice of rate function $\mathbb{P}^{(n)}(Z \notin \check{\mathcal{E}}_{n,2}(\mathsf{r})) = 0$. Since $g^{(2)}(x) \leq 1/4$, it follows that

$$\mathcal{L}_{n,\theta}(z) \geq -\frac{n}{8}(\theta - \theta_\star)'\frac{X'X}{n}(\theta - \theta_\star).$$

As a result, if $\theta - \theta_\star \in \Theta_\star$, $\mathcal{L}_{n,\theta}(z) \geq -(n/8)\bar{\kappa}(s_\star)\|\theta - \theta_\star\|_2^2$. Hence $\mathbb{P}^{(n)}(Z \notin \hat{\mathcal{E}}_{n,1}(\Theta_\star, \bar{L})) = 0$, for $\bar{L} = n\bar{\kappa}(s_\star)/4$. Finally, we have $\nabla \log q_{n,\theta_\star}(Z) = \sum_{i=1}^n \left(Z_i - g'(\langle x_i, \theta - \theta_\star\rangle)\right) x_i$, and by Hoeffding's inequality, and a standard union bound argument,

$$\begin{aligned}
\mathbb{P}^{(n)}\left(Z \notin \mathcal{E}_{n,0}(\rho)\right) &= \mathbb{P}^{(n)}\left(\max_{1\leq j\leq d}\left|\sum_{i=1}^n \left(Z_i - g'(\langle x_i, \theta - \theta_\star\rangle)\right) X_{ij}\right| > \frac{\rho}{2}\right) \\
&\leq 2\exp\left(\log(d) - \frac{\rho^2}{8\|X\|_\infty^2 n}\right) = \frac{2}{d},
\end{aligned}$$

given the choice of ρ in (5). Hence we can apply Corollary 10. This says that for any $k \geq 0$,

$$\begin{aligned}
\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}\left(\{\theta \in \mathbb{R}^d :\ \|\theta\|_0 > k\}|Z\right)\right] &\leq \frac{2}{d} \\
&\quad + e^{-A}\left(1 + \frac{\bar{\kappa}(s_\star)}{64\|X\|_\infty^2 \log(d)}\right)^{s_\star}\frac{1}{\pi_{\delta_\star}}\sum_{\delta:\ \|\delta\|_0 > k}\pi_\delta 4^{\|\delta\|_0},
\end{aligned}$$

where $A = (1/2)\inf_{x>0}\left[\mathsf{r}(x) - 4\rho s_\star^{1/2}x\right]$. It is not hard to verify that for $\tau, b, c > 0$, $\inf_{x>0}\left[\frac{\tau x^2}{1+bx} - cx\right] \geq -\frac{c^2}{4\sqrt{\tau}\sqrt{\tau - cb}} \geq -\frac{c^2}{2\tau}$, if $\tau \geq (4/3)bc$. In the case of A, the

condition $\tau \geq (4/3)bc$ is satisfies if $\sqrt{n} \geq 64 \times (4/3)\|X\|_\infty^2 s_\star \sqrt{\log(d)}/\underline{\kappa}_1$, and we have

$$A \geq -\frac{64\|X\|_\infty^2 s_\star \log(d)}{\underline{\kappa}_1}.$$

Using B2, and for $k \geq s_\star$,

$$\frac{1}{\pi_{\delta_\star}} \sum_{\delta:\ \|\delta\|_0 > k} \pi_\delta 4^{\|\delta\|_0} = \frac{\binom{d}{s_\star}}{g_{s_\star}} \sum_{j=k+1}^{d} 4^j g_j \leq \frac{\binom{d}{s_\star}}{g_{s_\star}} \sum_{j=k+1}^{d} 4^j \left(\frac{c_2}{d^{c_4}}\right)^{j-s_\star} g_{s_\star}$$
$$= \binom{d}{s_\star} 4^{s_\star} \sum_{j=k+1}^{d} \left(\frac{4c_2}{d^{c_4}}\right)^{j-s_\star}.$$

Therefore, if d is large enough so that $d^{c_4} \geq 8c_2$, then for $k \geq s_\star$,

$$\frac{1}{\pi_{\delta_\star}} \sum_{\delta:\ \|\delta\|_0 > k} \pi_\delta 4^{\|\delta\|_0} \leq 2\binom{d}{s_\star} 4^{s_\star} \left(\frac{4c_2}{d^{c_4}}\right)^{k-s_\star+1}$$
$$\leq 2\exp\left(s_\star \log(4) + s_\star \log(de) + (k+1-s_\star)\log\left(\frac{4c_2}{d^{c_4}}\right)\right),$$

using the combinatorial inequality $\binom{d}{s} \leq e^{s\log(de)}$. It follows that for d large enough such that $d^{c_4} \geq 8c_2$,

$$\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}\left(\{\theta \in \mathbb{R}^d :\ \|\theta\|_0 > k\}|Z\right)\right] \leq \frac{2}{d}$$
$$+2\exp\left[s_\star \log(d)\left(1 + \frac{64\|X\|_\infty^2}{\underline{\kappa}_1} + \frac{\bar{\kappa}(s_\star)}{64\|X\|_\infty^2 \log(d)^2} + \frac{\log(4e)}{\log(d)}\right) + (k+1-s_\star)\log\left(\frac{4c_2}{d^{c_4}}\right)\right].$$

Then for $\alpha > 0$, choose

$$k + 1 = s_\star + \frac{2\alpha}{c_4} + \frac{2}{c_4}\left(1 + \frac{64\|X\|_\infty^2}{\underline{\kappa}_1} + \frac{\bar{\kappa}(s_\star)}{64\|X\|_\infty^2 (\log(d))^2} + \frac{\log(4e)}{\log(d)}\right) s_\star, \qquad (45)$$

to conclude that the second term on the right-hand side of the above inequality is upper-bounded by $\frac{2}{d^\alpha}$, provided that $d^{c_4/2} \geq 4c_2$. Setting $\alpha = 1$ proves the theorem. □

Proof of Part(2). We apply Theorem 8 with $\bar{\lambda} = \rho\sqrt{\bar{s}}$ with ρ as in (5), and $\bar{s} = \zeta + s_\star$, with ζ as in Part (1). We also choose $\bar{L} = n\bar{\kappa}(s_\star)/4$, $\bar{\Theta} = \{\theta \in \mathbb{R}^d :\ \|\theta - \theta_\star\|_0 \leq \bar{s}\}$, the rate function $\mathsf{r}(x) = n\underline{\kappa}_1(\bar{s})x^2/(1 + \sqrt{\bar{s}}\|X\|_\infty x/2)$, and $\mathcal{E}_n = \mathcal{E}_{n,0}(\bar{\Theta}, \bar{\lambda}) \cap \hat{\mathcal{E}}_{n,1}(\Theta_\star, \bar{L}) \cap \check{\mathcal{E}}_{n,1}(\bar{\Theta}, \mathsf{r})$. With similar calculations as in Part (1), we conclude that

$$\mathbb{P}^{(n)}(Z \notin \mathcal{E}_n) \leq \frac{2}{d}.$$

If $\theta \notin \bar{\Theta}$, then $\|\theta\|_0 > \bar{s} - s_\star = \zeta$, and by Part (1), we conclude that

$$\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}(\mathbb{R}^d \setminus \bar{\Theta}|Z)\right] \leq \frac{4}{d}.$$

Recall that $\phi_{\mathsf{r}}(a) = \inf\{x > 0 : \ \mathsf{r}(z) - az \geq 0, \text{ for all } z \geq x\}$. Since $\mathsf{r}(x) = n\underline{\kappa}_1(\bar{s})x^2/(1+\sqrt{\bar{s}}\|X\|_\infty x/2)$, if $n\underline{\kappa}_1(\bar{s}) - \bar{s}^{1/2}\bar{\lambda}\|X\|_\infty > 0$, then

$$\bar{\epsilon} = \phi_{\mathsf{r}}(2\bar{\lambda}) = \frac{\bar{\lambda}}{n\underline{\kappa}_1(\bar{s}) - \bar{s}^{1/2}\bar{\lambda}\|X\|_\infty}.$$

Then we take n large enough so that $(3/4)n\underline{\kappa}_1(\bar{s}) \geq \bar{s}^{1/2}\bar{\lambda}\|X\|_\infty$, to conclude that

$$\bar{\epsilon} = \frac{\bar{\lambda}}{n\underline{\kappa}_1(\bar{s}) - \bar{s}^{1/2}\bar{\lambda}\|X\|_\infty} \leq \frac{4\bar{\lambda}}{n\underline{\kappa}_1(\bar{s})} = \frac{16\|X\|_\infty}{\underline{\kappa}_1(\bar{s})}\sqrt{\frac{\bar{s}\log(d)}{n}} < \infty.$$

The condition $(3/4)n\underline{\kappa}_1(\bar{s}) \geq \bar{s}^{1/2}\bar{\lambda}\|X\|_\infty$ translates into the sample size condition $\sqrt{n} \geq (16/3)\|X\|_\infty^2(\bar{s}/\underline{\kappa}_1(\bar{s}))\sqrt{\log(d)}$, which holds by assumption. We fix $M_0 \geq \max(500, 1 + (c_3 + c_4/2)/8)$, and apply Theorem 8 to get:

$$\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}\left(\left\{\theta \in \mathbb{R}^d : \ \|\theta - \theta_\star\| > M_0\bar{\epsilon}\right\}|Z\right)\right] \leq \frac{6}{d} + \sum_{j\geq 1}\mathsf{D}_j e^{-\frac{1}{8}\mathsf{r}(\frac{jM_0\bar{\epsilon}}{2})} + \frac{1}{\pi_{\delta_\star}}\left(\sum_{\delta\in\Delta_d} 2^{\|\delta\|_0}\pi_\delta\right)\left(1+\frac{\bar{L}}{\rho^2}\right)^{s_\star}\sum_{j\geq 1} e^{3\rho\bar{s}^{1/2}jM_0\bar{\epsilon}}e^{-\frac{1}{8}\mathsf{r}(\frac{jM_0\bar{\epsilon}}{2})}. \quad (46)$$

Since $\phi_{\mathsf{r}}(a)$ is defined as $\inf\{x > 0 : \ \mathsf{r}(z) \geq az, \ \text{ for all } z \geq x\}$, and $jM_0\bar{\epsilon}/2 \geq \bar{\epsilon} = \phi_{\mathsf{r}}(2\bar{\lambda})$, we have $\mathsf{r}(jM_0\bar{\epsilon}/2) \geq 2\bar{\lambda}(jM_0\bar{\epsilon}/2) = \rho\sqrt{\bar{s}}jM_0\bar{\epsilon}$. Hence

$$\sum_{j\geq 1} e^{-\frac{1}{8}\mathsf{r}\left(\frac{jM_0\bar{\epsilon}}{2}\right)} \leq \sum_{j\geq 1} e^{-\frac{1}{8}jM_0\sqrt{\bar{s}}\rho\bar{\epsilon}} = \frac{e^{-\frac{1}{8}M_0\sqrt{\bar{s}}\rho\bar{\epsilon}}}{1-e^{-\frac{1}{8}M_0\sqrt{\bar{s}}\rho\bar{\epsilon}}} \leq 2e^{-8M_0\bar{s}\log(d)}, \quad (47)$$

where the last inequality follows from the bounds

$$\frac{1}{8}M_0\sqrt{\bar{s}}\rho\bar{\epsilon} \geq \frac{1}{8}M_0\sqrt{\bar{s}}\rho\left(\frac{\bar{\lambda}}{n\underline{\kappa}_1(\bar{s})}\right) = 2M_0\frac{\bar{s}\|X\|_\infty^2}{\underline{\kappa}_1(\bar{s})}\log(d) \geq 8M_0\bar{s}\log(d) \geq 1$$

since $8M_0\bar{s} \geq 16M_0/c_4 \geq 1$, and $\log(d) \geq 1$, by assumption. Using the arguments in Example 7.1 of Ghosal et al. (2000) shows that the packing numbers D_j satisfies $\sup_{j\geq 1}\mathsf{D}_j \leq \binom{d}{\bar{s}}(24)^{\bar{s}} \leq (24)^{\bar{s}}e^{\bar{s}\log(de)}$. It follows that

$$\begin{aligned}\sum_{j\geq 1}\mathsf{D}_j e^{-\frac{1}{8}\mathsf{r}(\frac{jM_0\bar{\epsilon}}{2})} &\leq 2\exp\left[\bar{s}\log(d)\left(1+\frac{\log(24e)}{\log(d)} - 8M_0\right)\right] \\ &\leq \frac{2}{d},\end{aligned}$$

provided that $\log(d) \geq 1$, and using the condition $8M_0 \geq c_4/2 + 1 + \log(24e)$. Setting $x = jM_0\bar{\epsilon}/2$, we have

$$\begin{aligned} 3\rho\sqrt{\bar{s}}jM_0\bar{\epsilon} - \frac{1}{8}\mathsf{r}(\frac{jM_0\bar{\epsilon}}{2}) &\leq -\frac{x}{8}\left(\frac{n\underline{\kappa}_1(\bar{s})x}{1+\frac{1}{2}\sqrt{\bar{s}}\|X\|_\infty x} - 48\rho\sqrt{\bar{s}}\right), \\ &\leq -\frac{x}{8}\left(\frac{n\underline{\kappa}_1(\bar{s})\frac{M_0\bar{\epsilon}}{2}}{1+\frac{1}{2}\sqrt{\bar{s}}\|X\|_\infty\frac{M_0\bar{\epsilon}}{2}} - 48\rho\sqrt{\bar{s}}\right) \\ &\leq -\frac{2\rho\sqrt{\bar{s}}x}{8}, \end{aligned} \tag{48}$$

provided that

$$\frac{n\underline{\kappa}_1(\bar{s})\frac{M_0\bar{\epsilon}}{2}}{1+\frac{1}{2}\sqrt{\bar{s}}\|X\|_\infty\frac{M_0\bar{\epsilon}}{2}} - 48\rho\sqrt{\bar{s}} \geq 2\rho\sqrt{\bar{s}}.$$

This latter condition holds for all $M_0 \geq 500$, if $\sqrt{n} > 125\bar{s}\|X\|_\infty^2\sqrt{\log(d)}/\underline{\kappa}_1(\bar{s})$. In which case, from (48) we have

$$\sum_{j\geq 1} e^{3\rho\bar{s}^{1/2}jM_0\bar{\epsilon}}e^{-\frac{1}{8}\mathsf{r}(\frac{jM_0\bar{\epsilon}}{2})} \leq \sum_{j\geq 1} e^{-\frac{1}{8}jM_0\sqrt{\bar{s}}\rho\bar{\epsilon}} \overset{(a)}{\leq} 2e^{-8M_0\bar{s}\log(d)},$$

where the inequality (a) uses (47). Furthermore, using B2, and for $d^{c_4} \geq 4c_2$,

$$\pi_{\delta_\star}^{-1}\sum_{\delta\in\Delta_d}\pi_\delta 2^{\|\delta\|_0} = \frac{\binom{d}{s_\star}}{g_{s_\star}}\sum_{j=0}^{d}2^j g_j \leq \frac{\binom{d}{s_\star}}{g_{s_\star}}g_0\sum_{j=0}^{d}\left(\frac{2c_2}{d^{c_4}}\right)^j \leq 2\binom{d}{s_\star}\frac{g_0}{g_{s_\star}} \leq 2\binom{d}{s_\star}\left(\frac{d^{c_3}}{c_1}\right)^{s_\star}.$$

In conclusion, the last term on the right-hand side of (46) is upper-bounded by

$$4\binom{d}{s_\star}\left(\frac{d^{c_3}}{c_1}\right)^{s_\star}\left(1+\frac{\bar{L}}{\rho^2}\right)^{s_\star}e^{-8M_0\bar{s}\log(d)}.$$

And

$$\begin{aligned} \binom{d}{s_\star}\left(\frac{d^{c_3}}{c_1}\right)^{s_\star}\left(1+\frac{\bar{L}}{\rho^2}\right)^{s_\star}e^{-8M_0\bar{s}\log(d)} \leq \exp\Bigg[s_\star\log(d)\Bigg(1+c_3+\frac{\log(e/c_1)}{\log(d)} \\ +\frac{\bar{\kappa}(s_\star)}{64\|X\|_\infty^2\log(d)^2}\Bigg) - 8M_0\bar{s}\log(d)\Bigg]. \end{aligned} \tag{49}$$

Given that $\bar{s} = s_\star + \zeta$ with ζ as in Part (1), since $\log(d) \geq \log(e/c_1)$, and $8M_0 \geq 2 + c_3$, we see that the right-side of (49) is upper-bounded by $(1/d)^{16M_0/c_4} \leq (1/d)$. The theorem follows. □

5.4. **Proof of Theorem 5, 6, and 7.** It is obvious that H1 and H2 hold for this example. For convenience in the notation, for $z \in \mathbb{R}^{n\times p}$, $1 \le j \le p$, we let $z^{(j)} \in \mathbb{R}^{n\times p}$ be the matrix obtained by replacing all the components of the j-th column of z by 1. We introduce

$$q^{(j)}_{n,u}(z) \stackrel{\text{def}}{=} \prod_{i=1}^{n} \frac{\exp\left(z_{ij}\left(u_j + \sum_{k\neq j} u_k z_{ik}\right)\right)}{\exp\left(u_j + \sum_{k\neq j} u_k z_{ik}\right)},$$

$$\text{and} \quad \mathcal{L}^{(j)}_{n,u}(z) \stackrel{\text{def}}{=} \log q^{(j)}_{n,u}(z) - \log q^{(j)}_{n,\theta_{\star\cdot j}}(z) - \left\langle \nabla \log q^{(j)}_{n,\theta_{\star\cdot j}}(z), u - \theta_{\star\cdot j}\right\rangle, \quad u \in \mathbb{R}^p.$$

The function $u \mapsto q^{(j)}_{n,u}(z)$ is the likelihood function of the logistic regression model of the j-column of z on $z^{(j)}$. Let $\mathcal{H}^{(j)}_n(z) \stackrel{\text{def}}{=} \nabla^{(2)} \log q_{n,\theta_{\star\cdot j}}(z)$. Specifically, we have

$$(\mathcal{H}^{(j)}_n(z))_{st} \stackrel{\text{def}}{=} \sum_{i=1}^{n} g^{(2)}\left(\theta_{\star jj} + \sum_{k\neq j} z_{ik}\theta_{\star kj}\right) z^{(j)}_{is} z^{(j)}_{it}, \quad 1 \le s,t \le p.$$

We will need the following restricted smallest eigenvalues of $\mathcal{H}^{(j)}_n(z)$.

$$\underline{\kappa}^{(j)}_2(z) \stackrel{\text{def}}{=} \inf\left\{ \frac{u'(\mathcal{H}^{(j)}_n(z))u}{n\|u\|^2},\ u \in \mathbb{R}^p \setminus \{0\},\ \sum_{k:\ \delta_{\star,kj}=0} |u_k| \le 7 \sum_{k:\ \delta_{\star,kj}=1} |u_k|. \right\}.$$

$$\underline{\kappa}^{(j)}_2(s,z) \stackrel{\text{def}}{=} \inf\left\{ \frac{u'(\mathcal{H}^{(j)}_n(z))u}{n\|u\|^2},\ u \in \mathbb{R}^p \setminus \{0\},\ \|u\|_0 \le s. \right\}.$$

$$\underline{\kappa}_2(z) = \inf_{1\le j\le p} \underline{\kappa}^{(j)}_2(z), \quad \text{and} \quad \underline{\kappa}_2(s,z) = \inf_{1\le j\le p} \underline{\kappa}^{(j)}_2(s,z).$$

The next result shows that if $\underline{\kappa}_2(s) > 0$ and $\underline{\kappa}_2 > 0$ (with $\underline{\kappa}_2(s)$ and $\underline{\kappa}_2$ as defined in (11), then with high probability $\underline{\kappa}_2(Z) > 0$ and $\underline{\kappa}_2(s,Z) > 0$. The proof is an easy modification of the argument of Atchadé (2014) Lemma 2.5. We omit the details.

Lemma 15. *Assume C1. Let $Z \in \{0,1\}^{n\times p}$ be such that the row of Z are i.i.d. random variables with distribution $f_{\theta_\star}$. There exist finite universal constants a_1, a_2 such that the following two statements holds true.*

(1) For $1 \le s \le p$, if $\underline{\kappa}_2(s) > 0$, and $n \ge a_1 \left(\frac{s}{\underline{\kappa}_2(s)}\right)^2 \log(p)$, then

$$\mathbb{P}^{(n)}\left(\underline{\kappa}_2(s,Z) \le \frac{\underline{\kappa}_2(s)}{2}\right) \le e^{-a_2 n}.$$

(2) If $\underline{\kappa}_2 > 0$, and $n \ge a_1 \left(\frac{s_\star}{\underline{\kappa}}\right)^2 \log(p)$, then

$$\mathbb{P}^{(n)}\left(\underline{\kappa}_2(Z) \le \frac{\underline{\kappa}_2}{2}\right) \le e^{-a_2 n}.$$

Proof of Theorem 5. We will reduce the result to Theorem 2 Part(1). We set

$$\mathcal{G}^{(j)} \stackrel{\text{def}}{=} \{z \in \mathbb{R}^{n\times p} : \underline{\kappa}_2^{(j)}(z) > \underline{\kappa}_2/2\}, \quad \text{and} \quad \mathcal{G} \stackrel{\text{def}}{=} \{z \in \mathbb{R}^{n\times p} : \underline{\kappa}_2(z) > \underline{\kappa}_2/2\}.$$

We also define $\mathcal{A}^{(j)} \stackrel{\text{def}}{=} \{u \in \mathbb{R}^p : \|u\|_0 \le \zeta_j\}$. Recall that $\Theta = \{\theta \in \mathbb{R}^{p\times p} : \|\theta_{\cdot j}\|_0 \le \bar{s}_j,\ 1 \le j \le p\}$. Hence if $\theta \notin \Theta$, then $\theta_{\cdot j} \notin \mathcal{A}^{(j)}$, for some j. Therefore,

$$\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}\left(\mathbb{R}^{d\times d} \setminus \Theta | Z\right)\right] \le \mathbb{P}^{(n)}\left(Z \notin \mathcal{G}\right) + \sum_{j=1}^{p} \mathbb{E}^{(n)}\left[\mathbf{1}_{\mathcal{G}}(Z)\check{\Pi}_{n,d,j}\left(\mathbb{R}^d \setminus \mathcal{A}^{(j)} | Z\right)\right].$$

Note that $\mathcal{G} \subseteq \mathcal{G}^{(j)}$, and $\{Z \in \mathcal{G}^{(j)}\}$ is $Z^{(j)}-$measurable. Hence by conditioning on $Z^{(j)}$, we get

$$\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}\left(\mathbb{R}^{d\times d} \setminus \Theta | Z\right)\right] \le \mathbb{P}^{(n)}\left(Z \notin \mathcal{G}\right) + \sum_{j=1}^{p} \mathbb{E}^{(n)}\left[\mathbf{1}_{\mathcal{G}^{(j)}}(Z)\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d,j}\left(\mathbb{R}^d \setminus \mathcal{A}^{(j)} | Z\right) | Z^{(j)}\right]\right]$$

By conditioning on $Z^{(j)}$, and for $Z \in \mathcal{G}^{(j)}$, we are taken back to the setting of the standard logistic regression with a well-behaved design matrix. With the choice of ζ_j, and since ρ in (12) is taken larger than $4\sqrt{n\log(p)}$, by Theorem 2 (1), there exists an absolute constant A_1 such that for $p^{c_4} \ge 8c_2\max(1, 2c_2)$, and $n \ge A_1(s_\star/\underline{\kappa}_2)^2\log(p)$, we have

$$\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d,j}\left(\mathbb{R}^d \setminus \mathcal{A}^{(j)} | Z\right) | Z^{(j)}\right] \le \frac{4}{p^2}.$$

The term p^2 in $4/p^2$ comes from using $\alpha = 2$ in (45). Without any loss of generality we can take A_1 as large as the constant a_1 in Lemma 15 to conclude that $\mathbb{P}^{(n)}\left(Z \notin \mathcal{G}\right) \le e^{-a_2 n}$. Hence

$$\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}\left(\mathbb{R}^{d\times d} \setminus \Theta | Z\right)\right] \le e^{-a_2 n} + \frac{4}{p},$$

as claimed. □

Proof of Theorem 6. We shall apply Theorem 8. We will apply the theorem with the split cone

$$\bar{\Theta} \stackrel{\text{def}}{=} \{\theta \in \mathbb{R}^{d\times d} : \|\theta_{\cdot j}\|_0 \le \bar{s}_j,\ 1 \le j \le p\}.$$

Here the norm $\|\cdot\|_2$ in Theorem 8 is the Frobenius norm $\|\cdot\|_{\mathsf{F}}$, whereas the notation $\|\cdot\|_2$ in what follows will denote the Euclidean norm on $\mathbb{R}^p$. Notice that if $\theta \notin \theta_\star + \bar{\Theta}$, then $\theta \notin \Theta$ (where Θ is as defined in Theorem 5). Hence we will use Theorem 5 to control the term $\mathbb{E}^{(n)}\left(\check{\Pi}_{n,p}(\mathbb{R}^{d\times d} \setminus (\theta_\star + \bar{\Theta})|Z)\right)$. More precisely, there exist universal positive constants A_1, A_2 such that for $p^{c_4} \ge 8c_2\max(1, 2c_2)$, and $n \ge A_1(s_\star/\underline{\kappa}_2)^2\log(p)$,

$$\mathbb{E}^{(n)}\left(\check{\Pi}_{n,p}(\mathbb{R}^{d\times d} \setminus (\theta_\star + \bar{\Theta})|Z)\right) \le e^{-A_2 n} + \frac{4}{p}. \tag{50}$$

Set $\bar{S} \stackrel{\text{def}}{=} \sum_{j=1}^{p} \bar{s}_j$, $\bar{\lambda} = \rho\sqrt{\bar{S}}$, $\bar{L} = ns_\star/4$, $\mathsf{r}(x) = n\underline{\kappa}(\bar{s})x^2/(2 + \bar{S}^{1/2}x)$, and consider $\mathcal{E}_n = \mathcal{E}_{n,0}(\bar{\Theta}, \bar{\lambda}) \cap \hat{\mathcal{E}}_{n,1}(\Theta_\star, \bar{L}) \cap \check{\mathcal{E}}_{n,1}(\bar{\Theta}, \mathsf{r})$. We have

$$\sup_{u\in\bar{\Theta},\, \|u\|_{\mathsf{F}}=1} \left|\langle \nabla \log q_{n,\theta_\star}(Z), u\rangle_{\mathsf{F}}\right| \le \sqrt{\sum_{j=1}^{p} \bar{s}_j} \|\nabla \log q_{n,\theta_\star}(Z)\|_\infty.$$

Using this and a standard Hoeffding inequality, we obtain that

$$\mathbb{P}^{(n)}\left(Z \notin \mathcal{E}_{n,0}(\bar{\Theta}, \bar{\lambda})\right) \le 2\exp\left(2\log(p) - \frac{1}{2n}\left(\frac{\bar{\lambda}}{2\sqrt{\sum_{j=1}^{p} \bar{s}_j}}\right)^2\right) \le \frac{2}{p}, \tag{51}$$

given the choice of $\bar{\lambda}$, and ρ in (12).

We use a second order Taylor expansion of $u \mapsto q^{(j)}_{n,u}(z)$ around $\theta_{\star\cdot j}$ and the fact that $g^{(2)}(x) \le 1/4$ to deduce that for all $\theta \in \theta_\star + \Theta_\star$

$$\mathcal{L}_{n,\theta}(z) = -\frac{n}{8}\sum_{j=1}^{p}(\theta_{\cdot j} - \theta_{\star\cdot j})' \left(\frac{[Z^{(j)}]'[Z^{(j)}]}{n}\right)(\theta_{\cdot j} - \theta_{\star\cdot j}) \ge -\frac{ns_\star}{8}\left\|\theta - \theta_\star\right\|^2_{\mathsf{F}}.$$

Hence with $\bar{L} = ns_\star/4$,

$$\mathbb{P}^{(n)}(Z \notin \hat{\mathcal{E}}_{n,1}(\Theta_\star, \bar{L})) = 0. \tag{52}$$

Consider the set

$$\mathcal{G}^{(j)} \stackrel{\text{def}}{=} \{z \in \mathbb{R}^{n\times p} :\ \underline{\kappa}^{(j)}(\bar{s}, z) > \underline{\kappa}(\bar{s})/2\}, \quad \text{and} \quad \mathcal{G} \stackrel{\text{def}}{=} \{z \in \mathbb{R}^{n\times p} :\ \underline{\kappa}(\bar{s}, z) > \underline{\kappa}(\bar{s})/2\}.$$

Take $Z \in \mathcal{G}$. Then for all j, $\underline{\kappa}(\bar{s}, Z^{(j)}) > \underline{\kappa}(\bar{s})/2$ and we can use the same argument in (44) to conclude that for $\theta - \theta_\star \in \bar{\Theta}$,

$$\mathcal{L}^{(j)}_{n,\theta_{\cdot j}}(Z) \le -\frac{n\underline{\kappa}(\bar{s})\|\theta_{\cdot j} - \theta_{\star\cdot j}\|_2^2/2}{2 + \sqrt{\bar{s}_j}\|\theta_{\cdot j} - \theta_{\star\cdot j}\|_2}.$$

It follows that for $\theta - \theta_\star \in \bar{\Theta}$,

$$\mathcal{L}_{n,\theta}(Z) \le -\sum_{j=1}^{p}\frac{n\underline{\kappa}(\bar{s})\|\theta_{\cdot j} - \theta_{\star\cdot j}\|_2^2/2}{2 + \sqrt{\bar{s}_j}\|\theta_{\cdot j} - \theta_{\star\cdot j}\|_2} \le -\frac{1}{2}\frac{n\underline{\kappa}(\bar{s})\left\|\theta - \theta_\star\right\|^2_{\mathsf{F}}}{2 + \bar{S}^{1/2}\left\|\theta - \theta_\star\right\|_{\mathsf{F}}} = -\frac{1}{2}\mathsf{r}(\|\theta - \theta_\star\|_{\mathsf{F}}).$$

Hence, with the rate function $\mathsf{r}(x) = n\underline{\kappa}(\bar{s})x^2/(2 + \bar{S}^{1/2}x)$, we have

$$\mathbb{P}^{(n)}\left(Z \notin \check{\mathcal{E}}_{n,1}(\bar{\Theta}, \mathsf{r})\right) \le \mathbb{P}^{(n)}\left(Z \notin \mathcal{G}\right) \le e^{-a_2 n}, \tag{53}$$

as seen in Lemma 15, provided that $n \ge A_1\left(\frac{\bar{s}}{\underline{\kappa}(\bar{s})}\right)^2\log(p)$ (without any loss of generality, we take A_1 greater than the constant a_1 in Lemma 15). Hence, with $\mathcal{E}_n = \mathcal{E}_{n,0}(\bar{\Theta}, \bar{\lambda}) \cap \hat{\mathcal{E}}_{n,1}(\Theta_\star, \bar{L}) \cap \check{\mathcal{E}}_{n,1}(\bar{\Theta}, \mathsf{r})$, it follows from (51)-(53) that for $n \ge A_1\left(\frac{\bar{s}}{\underline{\kappa}(\bar{s})}\right)^2\log(p)$

$$\mathbb{P}^{(n)}\left(Z \notin \mathcal{E}_n\right) \le e^{-a_2 n} + \frac{2}{p}. \tag{54}$$

Finally, we note that with the same calculations as in the proof of Theorem 2 Part(2), we can choose the constant a_1 such that for $n \geq A_1 \left(\frac{\bar{S}}{\underline{\kappa}(\bar{s})}\right)^2 \log(p)$,

$$\frac{2\bar{\lambda}}{n\underline{\kappa}(\bar{s})} \leq \bar{\epsilon} = \phi_{\mathsf{r}}(2\bar{\lambda}) \leq \frac{4\bar{\lambda}}{n\underline{\kappa}(\bar{s})} \leq \frac{96}{\underline{\kappa}(\bar{s})}\sqrt{\frac{\bar{S}\log(p)}{n}} < \infty.$$

We are then ready to apply Theorem 9. Fix $M_0 \geq \max(500, 1 + (c_3 + c_4/2)/8)$, set $V \stackrel{\text{def}}{=} \{\theta \in \mathbb{R}^{d\times d} : \ \|\theta - \theta_\star\|_{\mathsf{F}} > M_0\bar{\epsilon}\}$, then for $n \geq A_1 \left(\frac{\bar{S}}{\underline{\kappa}(\bar{s})}\right)^2 \log(p)$, (22), (23), (50), and (54) give

$$\begin{aligned}\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}(V|Z)\right] \leq \left(e^{-a_2 n} + \frac{4}{p}\right) + \left(e^{-a_2 n} + \frac{2}{p}\right) + \sum_{j\geq 1} \mathsf{D}_j e^{-\frac{1}{8}\mathsf{r}(\frac{jM_0\bar{\epsilon}}{2})} \\ + \frac{1}{\pi_{\delta_\star}}\left(\sum_{\delta\in\Delta_p} 2^{\|\delta\|_0}\pi_\delta\right)^p \left(1 + \frac{\rho^2}{\bar{L}}\right)^{\sum_{j=1}^p s_{\star j}} \sum_{k\geq 1} e^{-\frac{1}{8}\mathsf{r}(\frac{kM_0\bar{\epsilon}}{2})} e^{3\rho c_0 k M_0 \bar{\epsilon}}. \quad (55)\end{aligned}$$

Similar calculations as in the proof of Theorem 2 Part(2) shows that

$$\sum_{j\geq 1} \mathsf{D}_j e^{-\frac{1}{8}\mathsf{r}(\frac{jM_0\bar{\epsilon}}{2})} \leq \frac{2}{p}, \ \text{ and } \ \sum_{j\geq 1} e^{-\frac{1}{8}\mathsf{r}(\frac{jM_0\bar{\epsilon}}{2})} e^{3\rho c_0 j M_0\bar{\epsilon}} \leq 2e^{-16M_0\bar{S}\log(p)},$$

and

$$\begin{aligned}\frac{1}{\pi_{\delta_\star}}\left(\sum_{\delta\in\Delta_p} 2^{\|\delta\|_0}\pi_\delta\right)^p \left(1 + \frac{\rho^2}{\bar{L}}\right)^{\sum_{j=1}^p s_{\star j}} &\leq \prod_{j=1}^p 2\binom{p}{s_{\star j}} \left(\frac{p^{c_3}}{c_1}\right)^{s_{\star j}} \left(1 + \frac{\bar{L}}{\rho^2}\right)^{s_{\star j}} \\ &\leq 2^p \exp\left(\left(\sum_{j=1}^p s_{\star j}\right)\log(p)\left(1 + c_3 + \frac{\log(e/c_1)}{\log(p)} + \frac{s_\star}{4(24^2)\log(p)^2}\right)\right).\end{aligned}$$

Hence, and by the same argument as in the proof of Theorem 2 Part(2), the last term on the right-side of (55) is bounded by $4/p$. □

Proof of Theorem 7. We will reduce this result to Theorem 2 Part(2). We set

$$\mathcal{V} \stackrel{\text{def}}{=} \{\theta \in \mathbb{R}^{p\times p} : \ \|\theta_{\cdot j}\|_2 > \epsilon_j, \ \text{ for some } j\},$$

and $\bar{\mathcal{V}} \stackrel{\text{def}}{=} \bar{\Theta} \cap \mathcal{V}$, where $\bar{\Theta} = \{\theta \in \mathbb{R}^{d\times d} : \ \|\theta_{\cdot j}\|_0 \leq \bar{s}_j, \ 1 \leq j \leq p\}$. Using Theorem 5 as we did in (50), there exist universal positive constants A_1, A_2 such that for $p^{c-4} \geq 8c_2\max(1, 2c_2)$, and $n \geq A_1(s_\star/\underline{\kappa}_2)^2\log(p)$,

$$\begin{aligned}\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}(\mathcal{V}|Z)\right] &\leq \mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}(\mathbb{R}^{d\times d}\setminus\bar{\Theta}|Z)\right] + \mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}(\bar{\mathcal{V}}|Z)\right], \\ &\leq e^{-A_2 n} + \frac{4}{p} + \mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}(\bar{\mathcal{V}}|Z)\right].\end{aligned}$$

We define

$$\mathcal{G}^{(j)} \stackrel{\text{def}}{=} \{z \in \mathbb{R}^{n\times p} : \underline{\kappa}_2^{(j)}(z) > \underline{\kappa}_2/2\}, \quad \text{and} \quad \mathcal{G} \stackrel{\text{def}}{=} \{z \in \mathbb{R}^{n\times p} : \underline{\kappa}_2(z) > \underline{\kappa}_2/2\}.$$

We also define $\mathcal{A}^{(j)} \stackrel{\text{def}}{=} \{u \in \mathbb{R}^p : \|u\|_0 \leq \bar{s}_j, \text{ and } \|u\|_2 > \epsilon_j\}$. Hence, if $\theta \in \bar{\mathcal{V}}$, then $\theta_{\cdot j} \in \mathcal{A}^{(j)}$, for some j. Therefore,

$$\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}(\bar{\mathcal{V}}|Z)\right] \leq \mathbb{P}^{(n)}\left(Z \notin \mathcal{G}\right) + \sum_{j=1}^{p} \mathbb{E}^{(n)}\left[\mathbf{1}_{\mathcal{G}^{(j)}}(Z)\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d,j}\left(\mathcal{A}^{(j)}|Z\right)|Z^{(j)}\right]\right]$$

Fix $M_0 \geq \max(125, c_4(1+c_3)/64)$. $\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d,j}\left(\mathcal{A}^{(j)}|Z\right)\right]$ is the same as the posterior distribution of the logistic regression of the j-th column of Z on $Z^{(j)}$, and for $Z^{(j)} \in \mathcal{G}^{(j)}$, we can apply Theorem 2 Part(2). Hence, we can take A_1 large enough so that for $p \geq e(1+c_1)/c_1$, and $n \geq A_1(\bar{s}/\underline{\kappa}_2)^2\log(p)$,

$$\mathbf{1}_{\mathcal{G}^{(j)}}(Z)\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d,j}\left(\mathcal{A}^{(j)}|Z\right)|Z^{(j)}\right] \leq \frac{8}{p^2}.$$

Hence

$$\mathbb{E}^{(n)}\left[\check{\Pi}_{n,d}(\mathcal{V}|Z)\right] \leq 2e^{-A_2 n} + \frac{12}{p}.$$

□

Acknowledgements

The author would like to thank Shuheng Zhou for very helpful conversations.

References

Alquier, P. and Lounici, K. (2011). PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electron. J. Stat.* **5** 127–145.

Arias-Castro, E. and Lounici, K. (2014). Estimation and variable selection with exponential weights. *Electron. J. Stat.* **8** 328–354.

Atchadé, Y. F. (2014). Estimation of high-dimensional partially-observed discrete Markov random fields. *Electronic Journal of Statistics* **8** 2242–2263.

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

Atchadé, Y. F. (2015). A scalable quasi-Bayesian framework for Gaussian graphical models. *ArXiv e-prints* .

Bach, F. (2010). Self-concordant analysis for logistic regression. *Electron. J. Statist.* **4** 384–414.

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516.

Baricz, A. (2008). Mills' ratio: Monotonicity patterns and functional inequalities. *Journal of Mathematical Analysis and Applications* **340** 1362 – 1370.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236.

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.

Catoni, O. (2004). *Statistical learning theory and stochastic optimization*, vol. 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *J. Econometrics* **115** 293–346.

Dalalyan, A. S. and Tsybakov, A. B. (2007). Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory*, vol. 4539 of *Lecture Notes in Comput. Sci.* Springer, Berlin, 97–111.

Florens, J.-P. and Simoni, A. (2012). Nonparametric estimation of an instrumental regression: A quasi-bayesian approach based on regularized posterior. *Journal of Econometrics* **170** 458 – 475.

Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian nonparametrics*. Springer Series in Statistics, Springer-Verlag, New York.

Höfling, H. and Tibshirani, R. (2009). Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *J. Mach. Learn. Res.* **10** 883–906.

Kato, K. (2013). Quasi-Bayesian analysis of nonparametric instrumental variables models. *Ann. Statist.* **41** 2359–2390.

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

Li, C. and Jiang, W. (2014). Model Selection for Likelihood-free Bayesian Methods Based on Moment Conditions: Theory and Numerical Examples. *ArXiv e-prints* .

Li, Y.-H., Scarlett, J., Ravikumar, P. and Cevher, V. (2014). Sparsistency of ℓ_1-regularized M-estimators. *arXiv preprint arXiv:1410.7605* .

Liao, Y. and Jiang, W. (2011). Posterior consistency of nonparametric conditional moment restricted models. *Ann. Statist.* **39** 3003–3031.

Lyne, A.-M., Girolami, M., Atchadé, Y., Strathmann, H. and Simpson, D. (2015). On russian roulette estimates for bayesian inference with doubly-intractable likelihoods. *Statist. Sci.* **30** 443–467.

Marin, J.-M., Pudlo, P., Robert, C. and Ryder, R. (2012). Approximate bayesian computational methods. *Statistics and Computing* **22** 1167–1180.

McAllester, D. A. (1999). Some pac-bayesian theorems. *Machine Learning* **37** 355–363.

Meinshausen, N. and Buhlmann, P. (2006). High-dimensional graphs with the lasso. *Annals of Stat.* **34** 1436–1462.

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

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.

Ravikumar, P., Wainwright, M. J. and Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319.

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* .

Sun, T. and Zhang, C.-H. (2013). Sparse Matrix Inversion with Scaled Lasso. *Journal of Machine Learning Research* **14** 3385–3418.

Yang, W. and He, X. (2012). Bayesian empirical likelihood for quantile regression. *Ann. Statist.* **40** 1102–1131.

Zhang, T. (2006). From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics* **34** pp. 2180–2210.