Posterior contraction in sparse generalized linear models

BY SEONGHYUN JEONG
Booth School of Business, University of Chicago, Chicago, Illinois 60637, U.S.A.
seonghyun.jeong@chicagobooth.edu

AND SUBHASHIS GHOSAL
Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.
sghosal@ncsu.edu

SUMMARY

We study posterior contraction rates in sparse high-dimensional generalized linear models using priors incorporating sparsity. A mixture of a point mass at zero and a continuous distribution is used as the prior distribution on regression coefficients. In addition to the usual posterior, the fractional posterior, which is obtained by applying the Bayes theorem on a fractional power of the likelihood, is also considered. The latter allows uniformity in posterior contraction over a larger subset of the parameter space. In our setup, the link function of the generalized linear model need not be canonical. We show that Bayesian methods achieve convergence properties analogous to lasso-type procedures. Our results can be used to derive posterior contraction rates in many generalized linear models including logistic, Poisson regression, and others.

Some key words: Fractional posterior; Generalized linear model; High-dimensional regression; Posterior contraction rates; Sparsity-inducing priors.

1. INTRODUCTION

Generalized linear models (McCullagh and Nelder, 1989) provide a convenient unified framework for linear and nonlinear regression, and are arguably the most widely used models in applications. But despite their practical appeal, there has been relatively little theoretical work on generalized linear models in high-dimensional settings, especially compared to their linear model counterpart. In the non-Bayesian literature, pioneering work includes van de Geer (2008), who studies the oracle rate for lasso estimators in generalized linear models under a Lipschitz loss function, and Abramovich and Grinshtein (2016), who derived convergence rates for penalized estimators relative to the Kullback-Leibler risk for a wide range of penalty functions. For further results on the optimality of estimators in high-dimensional generalized linear models, see Negahban et al. (2012) and Rigollet (2012).

In the Bayesian literature, while linear models in high dimensions received considerable attention (e.g., Castillo et al., 2015; Martin et al., 2017; Bai et al., 2020; Ning et al., 2020; Belitser and Ghosal, 2020; Gao et al., 2020), the work on generalized linear models is sparse. For generalized linear models with increasing dimensional parameters, Ghosal (1997) studied asymptotic normality of the posterior distribution where the parameter dimension grows at a rate slower than the sample size. To the best of our knowledge, only Jiang (2007) studied the asymptotic proper-
ties of Bayesian procedures in generalized linear models with dimension potentially exceeding
the sample size. In the setting where the predictor variables are stochastic, so that the observa-
tions can be treated as independent and identically distributed, he obtained posterior contraction
rates relative to the Hellinger metric on the joint distribution of the predictor and response vari-
able. Liang et al. (2013) also obtained comparable contraction rates for maximum a posteriori
models relative to the Hellinger metric. Although the Hellinger metric is useful for comparison
of densities, it lacks the strength and interpretability of a metric on the regression coefficient
space. A much clearer interpretation of convergence is obtained from contraction with respect to
$\ell_q$-type metrics directly on the parameter space, but this has not been addressed in the literature.
As the regression coefficients are intertwined with a random design matrix, such a rate may not
be obtained from the contraction rate for the joint density. Our study aims to fill this gap in the
deterministic predictor setting. To do so, we impose appropriate conditions on the design matrix.

We consider high-dimensional generalized linear models with fixed design and investigate the
desired posterior contraction rates relative to the squared loss of the linear predictor and the $\ell_q$
-norms, $q = 1, 2$. The distribution belongs to the overdispersed exponential family in its natural
form; that is, the density of the $i$th observation with respect to a dominating measure $\nu$ is

$$f_i(y_i) = \exp\left(\{y_i\theta_i - b(\theta_i)\}/\tau_i + k(y_i, \tau_i)\right),$$

where $\theta_i$ is a natural parameter lying in $\Theta \subset \mathbb{R}$, $\tau_i$ is a known dispersion factor, and $b$ and $k$
are known functions. Following the usual convention, the function $b$ is assumed to be twice-
differentiable and strictly convex on $\Theta$ so that the second derivative $b''$ satisfies $b''(x) > 0$ for
every $x \in \Theta$. Using the derivatives of $b$, it is easy to see that the expected value and the variance
of the $i$th response variable $Y_i$ are given by $b'(\theta_i)$ and $\tau_i b''(\theta_i)$, respectively. For a fixed covariate
$X_i \in \mathbb{R}^p$ and a vector of high-dimensional regression coefficient vector $\beta \in \mathbb{R}^p$ with $p > n$, we
consider an increasing link function $h$ such that $(h \circ b') : \Theta \mapsto \mathbb{R}$ is strictly increasing and

$$(h \circ b')(\theta_i) = X_i^T \beta,$$

where $(h \circ b')(\cdot) = h\{b'(\cdot)\}$. As $p > n$, we assume that the coefficient $\beta$ is sparse, which means
that many components of $\beta$ are zero. We note that our formulation does not require $h$ to be $b'^{-1}$,
the inverse function of $b$, so high-dimensional generalized linear models with non-canonical link
functions are also included. We believe that this generalization is also a substantial contribution
of this paper because our results apply to many models with non-canonical links including probit
regression, negative binomial regression, and gamma regression with the logarithmic link.

Ghosal and van der Vaart (2007) developed the general posterior contraction theory for observ-
ations not necessarily independent or identically distributed. Using a result about the existence
of exponentially powerful tests in Birgé (1983), they characterized the posterior contraction rate
relative to the root-average-squared Hellinger metric for models with independent but possibly non-identically distributed observations. With an appropriate class of prior distributions, we fol-
low their approach to derive the posterior contraction rate relative to the root-average-squared
Hellinger metric in high-dimensional generalized linear models with deterministic predictors
and the response variables independent but not identically distributed. In this sense, our study is
fundamentally different from Jiang (2007), where the predictors are random and hence the ob-
servations are independent and identically distributed. The posterior contraction theory requires
control on the metric entropy, the decay of the prior distribution in its tail area, and the concen-
tration of the prior distribution near the true value (Ghosal and van der Vaart, 2017). As shown
in the Poisson regression example in this paper, the first two can be more restrictive than having
a good prior concentration. The contraction rate for the fractional posterior distribution can be
obtained only depending on the prior concentration rate (Zhang, 2006; Ghosal and van der Vaart,
Posterior contraction in sparse generalized linear models

2017; Bhattacharya et al., 2019). We obtain the \(\ell_q\)-contraction rates \((q = 1, 2)\) of the usual and fractional posterior distributions for the regression coefficients under suitable boundedness conditions. Our results imply that the contraction rates are adaptive to the unknown sparsity level. The resulting \(\ell_1\) and \(\ell_2\)-contraction rates are given by \(s_0\{\log p\}/n^{1/2}\) and \(\{s_0\log p\}/n^{1/2}\) (up to some compatibility coefficients defined later), where \(s_0\) is the number of the true nonzero coefficients. The rates are comparable to those given in the frequentist literature, up to logarithmic factors (e.g., van de Geer, 2008; Abramovich and Grinshtein, 2016).

2. Setup and Prior Specification

2.1. Notation

We have \(n\) independent observations denoted by \(Y^{(n)} = (Y_1, \ldots, Y_n)^T\), generated from the model in (1) and (2) with \(\beta_0\) the true value of \(\beta\). Let \(X \in \mathbb{R}^{n \times p}\) be the design matrix whose rows consist of \(X_i, i = 1, \ldots, n\). For a vector \(\beta \in \mathbb{R}^p\) and a set \(S \subset \{1, \ldots, p\}\), we use the notations \(\beta_S = \{\beta_j, j \in S\}\) and \(\beta_{\bar{S}} = \{\beta_j, j \notin S\}\) to separate \(\beta\) into zero and nonzero coefficients using \(S\). We also denote by \(S_\beta = \{j : \beta_j \neq 0\} \subset \{1, \ldots, p\}\) the effective support determined by \(\beta\). The cardinalities of \(S\) and \(S_\beta\) are denoted by \(s = |S|\) and \(s_\beta = |S_\beta|\). In particular, we write the true support and its cardinality as \(S_0\) and \(s_0\), respectively. We write \(X_{S,i} \in \mathbb{R}^s\) for the matrix whose columns are chosen from \(X\) by a given \(S\). The vector \(X_{S,i} \in \mathbb{R}^s\) is the \(i\)th row of \(X_{S}\). The notation \(I_s\) denotes the \(s\)-dimensional identity matrix.

For two sequences \(a_n\) and \(b_n\), \(a_n \lesssim b_n\) implies \(a_n \leq C b_n\) for some constant \(C > 0\) independent of \(n\). Let \(\| \cdot \|_q\) denote the \(\ell_q\)-norm and \(\| \cdot \|_\infty\) denote the max-norm of vectors. The max-norm notation \(\| \cdot \|_\infty\) is also used for a matrix to denote the maximum absolute value of entries. We write \(\rho_{\min}(\cdot)\) and \(\rho_{\max}(\cdot)\) for the minimum and maximum eigenvalues of a square matrix, respectively.

For \(i = 1, \ldots, n\), let \(f_{\beta,i}\) stand for the density of \(Y_i\) with the associated value of the predictor variable being \(X_i\) and an arbitrary parameter \(\beta\). The joint density of \(Y^{(n)}\) is written as \(f_{\beta} = \prod_{i=1}^n f_{\beta,i}\). Thus the likelihood ratio is given by \(\Lambda_n(\beta) = (f_{\beta}/f_{\beta_0})(Y^{(n)})\). Let \(\Pi\) be a given prior distribution for \(\beta\). For any given \(\alpha \in (0, 1]\), we define the fractional posterior \(\Pi_{\alpha}(\cdot | Y^{(n)})\) by

\[
\Pi_{\alpha}(\beta \in B | Y^{(n)}) = \frac{\int_{\mathbb{R}^p} \Lambda_n^\alpha(\beta) d\Pi(\beta)}{\int_{\mathbb{R}^p} \Lambda_n^\alpha(\beta) d\Pi(\beta)}, \quad \text{for any measurable } B \subset \mathbb{R}^p.
\]

This definition reduces to the usual posterior distribution if \(\alpha = 1\). By a slight abuse of terminology, we often refer to both the usual and fractional posteriors by the term ‘posterior’.

For two parameter values \(\beta_1\) and \(\beta_2\), the root-average-squared Hellinger metric is defined by \(H_n(\beta_1, \beta_2) = \{n^{-1} \sum_{i=1}^n H^2(f_{\hat{\beta}_1,i}, f_{\hat{\beta}_2,i})\}^{1/2}\), where \(H(f_{\hat{\beta}_1,i}, f_{\hat{\beta}_2,i}) = \{f(\sqrt{f_{\hat{\beta}_1,i}^2 + \sqrt{f_{\hat{\beta}_2,i}^2}}) \, d\nu\}^{1/2}\) is the Hellinger distance between \(f_{\hat{\beta}_1,i}\) and \(f_{\hat{\beta}_2,i}\), \(i = 1, \ldots, n\). The notations \(\mathbb{E}_0\) and \(\mathbb{P}_0\) denote the expectation and probability operators with the true parameter \(\beta_0\), respectively. We also define the uniform compatibility number \(\phi_1\) and the smallest scaled singular value \(\phi_2\) for a square matrix \(W \in \mathbb{R}^{n \times n}\),

\[
\phi_1(s; W) = \inf_{\beta_1 \leq |S| \leq s} \frac{\|WX\beta_1|S_\beta^{1/2}\|_1}{n^{1/2} \|\beta_1\|_1}, \quad \phi_2(s; W) = \inf_{\beta_1 \leq |S| \leq s} \frac{\|WX\beta_1\|_2}{n^{1/2} \|\beta_1\|_2}.
\]

The compatibility numbers are required to obtain the \(\ell_1\)- and \(\ell_2\)-contraction rates. Similar definitions are widely used in the Bayesian high-dimensional literature (e.g., Castillo et al., 2015; Martin et al., 2017; Belitser and Ghosal, 2020). Unlike in the linear regression setup, however, our definitions have an extra scaling factor \(W\) that is appropriate for generalized linear models.
2.2. Prior specification

The choice of the prior distribution is crucial to obtain a good posterior contraction rate in high- and infinite-dimensional models. In this subsection, we specify a class of prior distributions that induces the desired contraction rate for the high-dimensional regression coefficients \( \beta \).

As is customary in the literature, we first choose a value for the dimension \( s \) from a prior \( \pi_p \), and then choose a subset \( S \subseteq \{1, \ldots, p\} \) of cardinality \( s \) randomly. Next, we select \( \beta_S \) from a density \( g_S \) on \( \mathbb{R}^s \) while \( \beta_{S^c} \) is set to zero. The resulting prior for \((S, \beta)\) is

\[
(S, \beta) \mapsto \left(\frac{p}{s}\right)^{-1} \pi_p(s) g_S(\beta_S) \delta_0(\beta_{S^c}),
\]

where \( \delta_0 \) is the Dirac measure at zero on \( \mathbb{R}^{p-s} \) with the dimension suppressed. For the prior \( \pi_p \) on a model size, we consider a distribution such that for some constants \( A_1, A_2, A_3, A_4 > 0 \),

\[
A_1 p^{-A_3} \pi_p(s - 1) \leq \pi_p(s) \leq A_2 p^{-A_4} \pi_p(s - 1), \quad s = 1, \ldots, \pi,
\]

\[
\pi_p(s) = 0, \quad s = \pi + 1, \ldots, p,
\]

for some predetermined bound \( \pi \). Examples of priors satisfying (3) can be found in Castillo and van der Vaart (2012) and Castillo et al. (2015). If \( g_S \) is well-defined for every \( S \subseteq \{1, \ldots, p\} \), there is no need to make a restriction and the bound \( \pi \) can be chosen to be \( p \) as in Castillo et al. (2015). In that case, the second line of (3) disappears. Otherwise, a restriction \( \pi < p \) may be required, coupled with a choice of \( g_S \); see Examples 3–4 in Section 4 and also the empirical Bayes priors in Martin et al. (2017) and Chae et al. (2019). Even in such a case, we need to consider a sequence \( \pi \) increasing at a suitable rate so that \( s_0 \leq \pi \) for large enough \( n \), to recover the true sparsity level.

The prior density \( g_S \) needs to be specified only for \( S \) such that \( s \leq \pi \). We consider a class of distributions satisfying certain conditions motivated by our targeted rate. For a given \( g_S \), let \( G_S \) be the probability measure with density \( g_S \), and \( Z_S \in \mathbb{R}^s \) be distributed as \( G_S \). Henceforth we write \( \xi = (h \circ b')^{-1} \) and \( \tau = \min_{1 \leq i \leq n} \tau_i \).

**Assumption 1.** Let \( \gamma_n(\beta) = 1 + \max\{ (h^{-1})(X_i^T \beta)\xi(X_i^T \beta) : 1 \leq i \leq n \} \). There exist constants \( m_1 > 0, m_2 > 0 \) such that the prior density \( g_S \) satisfies

\[
G_{S_0} \left\{ \| X_{S_0} (Z_{S_0} - \beta_0, S_0) \|_\infty^2 \leq \frac{m_1 s_0 \log p}{\tau \gamma_n(\beta_0) n} \right\} \geq e^{-m_2 s_0 \log p}
\]

for sufficiently large \( n \).

**Assumption 2.** Let \( \psi(L) = 1 + \sup_{|\eta| \leq L} (h^{-1}(\eta))\xi(\eta) \). For sufficiently large constants \( m_3, m_4 \) and some \( q \in [1, \infty] \), there exist invertible matrices \( \mathcal{V} = \{ V_S \in \mathbb{R}^{s \times s} : s \leq \pi \} \) with

\[
- \log p \lesssim \min_{S: s \leq m_3 s_0} \log \rho_{\min}(V_S^T V_S) \leq \max_{S: s \leq m_3 s_0} \log \rho_{\max}(V_S^T V_S) \lesssim \log p
\]

and a sequence \( L_n \) with \( \log L_n + \log \psi(L_n) \lesssim \log p \) such that the prior density \( g_S \) satisfies

\[
\max_{S: s \leq m_3 s_0} G_S \left\{ \| V_S Z_S \|_q \max_{1 \leq i \leq n} \| (V_S^T)^{-1} X_{S,i} \|_{q/(q-1)} > L_n \right\} \leq e^{-m_4 s_0 \log p}
\]

for sufficiently large \( n \).

Assumption 1 is needed to ensure sufficient prior mass on a Kullback-Leibler neighborhood around the true density, and is used for both the usual and fractional posteriors. We note that Assumption 1 is dependent on the true support \( S_0 \). Although the true sparsity class is unknown,
Posterior contraction in sparse generalized linear models

this is typically not an issue if the density \( g_S \) belongs to the same family for every \( S \subset \{1, \ldots, p\} \) such that \( s \leq \tilde{s} \); see Examples 1–4 in Section 4. Clearly, larger values of \( m_1 \) and \( m_2 \) make the assumption more easily satisfied. However, smaller values allow tighter theoretical bounds. Assumption 2 is to control the entropy and prior tail-decay, and is needed only for the usual posterior distribution. As will be discussed in Section 4, Assumption 2 may require stronger conditions than Assumption 1 in some instances of generalized linear models. The use of the fractional posterior may be useful in these cases. It is sufficient that Assumption 2 holds for some moderately large \( m_3 \) and \( m_4 \).

As illustrated by the examples and the remarks in Section 4, Assumptions 1–2 require a restriction on the magnitude of the true signal \( \beta_0 \) or the linear predictor \( X\beta_0 \). In high-dimensional linear regression, such a restriction is often circumvented by choosing an appropriate prior distribution (e.g., Castillo et al., 2015; Martin et al., 2017; Gao et al., 2020; Belitser and Ghosal, 2020). In our setting, we need the restriction for the additional complexity of generalized regression models. For this reason, our main results are not meant to be applied to sparse linear regression with Gaussian errors, where similar results under milder conditions are available. Our main contribution is focused on more complicated generalized linear models.

3. POSTERIOR CONTRACTION RATES

In this section, we provide unified results on contraction rates in high-dimensional generalized linear models under suitable assumptions on the design matrix \( X \) and the true regression coefficients \( \beta_0 \) for both the usual and fractional posteriors. In the main results below, we shall use the following notations. For a given prior \( \Pi \), constants \( M = \{m_1, \ldots, m_4\} \), \( q \in [1, \infty] \), invertible matrices \( \mathcal{V} \), and a sequence \( L_n \), let \( \Delta_1(M; \Pi) = \Delta_1(m_1, m_2; \Pi) \) and \( \Delta_2(M, q, \mathcal{V}, L_n; \Pi) = \Delta_2(m_3, m_4, q, \mathcal{V}, L_n; \Pi) \) be the sets of \( \beta_0 \in \mathbb{R}^p \) satisfying (4) and (6), respectively.

We first present a lemma giving a lower bound for the denominator of the posterior distribution with probability tending to one. The bound is directly related to our targeted contraction rate for the model. The assertion only requires a sufficient prior concentration near the true parameter value, which is provided by Assumption 1.

**Lemma 1 (Evidence Lower Bound).** For any given \( \alpha \in (0, 1] \), \( m_1 > 0 \), \( m_2 > 0 \), and \( \kappa_n = o(n) \), the constant \( K_0 = \alpha m_1 + m_2 + 1 + \delta \) with a sufficiently small \( \delta > 0 \) satisfies

\[
\inf_{\beta_0 \in \Delta_1(M; \Pi)} \mathbb{P}_0 \left\{ \int_{\mathbb{R}^p} A_n^\alpha(\beta)d\Pi(\beta) \geq \pi_p(s_0)e^{-K_0s_0\log p} \right\} \to 1. \tag{7}
\]

Lemma 1 plays a key role in the derivation of the posterior contraction rate. The fact that \( K_0 \) is dependent on \( m_1 \) and \( m_2 \) in Assumption 1 reveals that smaller \( m_1 \) and \( m_2 \) help obtain a sharper threshold. The result is used to derive our main results on the effective dimension and the posterior contraction rate in Theorems 1–2.

Theorem 1 states that the effective dimension is not much larger than the true one. The result allows us to restrict the attention to models of relatively small sizes, rather than the full dimension \( p \), so that we can effectively control the closeness of the two densities with respect to the root-average-squared Hellinger metric.
We note that the constant $K_1$ in the threshold is free of any assumption on the design matrix. A similar result for the contraction rate relative to the squared error loss of the linear predictor is made to achieve the optimality. Gao et al. (2020) removed the condition to obtain a threshold. Instead, the results in Section 4 of Castillo et al. (2015) can be compared with ours, where certain size restriction of the true linear predictor is shown in Theorem 3. In this sense, our results cannot be directly compared to those for sparse linear regression in Section 2 of Castillo et al. (2015). Instead, the results in Section 4 of Castillo et al. (2015) who only studied posterior contraction relative to the Hellinger metric. We shall recover the contraction rates for the contraction rate relative to the root-average-squared Hellinger metric. Whereas only an adequate prior concentration suffices for the fractional posterior contraction (Bhattacharya et al., 2019), derivation far away from the true parameter value (Birgé, 1983). The assertion thus follows by controlling the metric entropy of the sieve, measured with respect to the root-average-squared Hellinger metric, and verifying that a sieve grows sufficiently fast so that the residual prior probability is exponentially small (Ghosal and van der Vaart, 2007). These are assured by Assumption 2.

**Theorem 2** (Contraction rate, Hellinger). For any given $\alpha \in (0, 1]$, $m_1 > 0$, $m_2 > 0$, $m_3 \geq K_1$, $m_4 > K_0 + A_3$, $q \in [1, \infty]$, $\kappa_n = o(n)$, $V$ with (5), and $L_n$ with $\log L_n + \log \psi(L_n) \lesssim \log p$, there exists a constant $K_2 > 0$ such that

$$\sup_{\beta_0 \in \Delta_0(M; \Pi) : s_0 \log p \leq \kappa_n} \mathbb{E}_0 \Pi_0 \left\{ \beta : H_n(\beta, \beta_0) > K_2 n^{-1/2} (s_0 \log p)^{1/2} \right\} \rightarrow 0,$$

where $\Delta_0(M, q, V, L_n; \Pi)$ is $\Delta_1(M; \Pi)$ if $\alpha \in (0, 1)$, and is $\Delta_1(M; \Pi) \cap \Delta_2(M, q, V, L_n; \Pi)$ if $\alpha = 1$.

The theorem calibrates specific thresholds for $m_3$ and $m_4$ in Assumption 2. It is worthwhile to compare our results in Theorem 2 with the posterior contraction rates obtained in Jiang (2007). Corollary 1 of Jiang (2007) implies that if $s_0 = o(\log b_1 n)$ and $\log p \lesssim n^{b_2}$ for some $b_1 > 1$ and $b_2 \in (0, 1)$, then the contraction rate with respect to the Hellinger metric for the independent and identically distributed observations is given by $n^{-(1-b_2)/2} (\log n)^{b_2/2}$, which is also obtained by plugging in those bounds for the sequences in the rate given in Theorem 2. However, the implication of our rate is different, as we deal with independent but not identically distributed observations and use the root-average-squared Hellinger metric to quantify the rate. Similarly, we can also recover the rates in Corollary 2 of Jiang (2007). An important point to note that Jiang (2007) made a very strong assumption that $\|\beta_0\|_1 \lesssim 1$, which we do not need.

Because of the vagueness of the root-average-squared Hellinger metric used in Theorem 2, the assertion says nothing explicitly about the closeness of $\beta$ and $\beta_0$ in terms of a Euclidean-type distance. Our main purpose is to go significantly beyond Jiang (2007) and Liang et al. (2013) who only studied posterior contraction relative to the Hellinger metric. We shall recover the contraction rates for $\beta$ with respect to more concrete metrics, under the following additional
boundedness requirement on the sparsity of the true parameter $\beta_0$: for $W_0 = \text{diag}(w_{01}, \ldots, w_{0n})$ with $w_{0i}^2 = (h^{-1})'(X_i^T \beta_0)\xi'(X_i^T \beta_0)$,
\[ s_0^2(\log p)\|X\|^2 / \phi_2^2(K_1 s_0; W_0) = o(n), \] (8)
where $K_1 = K_0 + 1$. Since $\phi_1(s; W) \geq \phi_2(s; W)$ for any $s > 0$ by the Cauchy-Schwarz inequality, the compatibility number in (8) is removed if the smallest scaled singular value is bounded away from zero. The latter may be achieved under certain mild conditions with some (stochastic) size restriction on the true linear predictor $X \beta_0$; for example, see Lemma A.4 of Narisetty et al. (2019). In most generalized linear models, under the stronger condition that $\|X \beta_0\|_\infty \lesssim 1$ such that $w_0$ are bounded away from zero, the compatibility numbers are also usually bounded away from zero; see Example 7 of Castillo et al. (2015). The quantity $\|X\|_\infty$ is often assumed to be uniformly bounded. Then a sufficient condition for (8) is $s_0^2 \log p = o(n)$.

**THEOREM 3 (RECOVERY).** For any given $\alpha \in (0, 1]$, $m_1 > 0$, $m_2 > 0$, $m_3 \geq K_1$, $m_4 > K_0 + A_3$, $q \in [1, \infty]$, $\kappa_n = o(n)$, $V$ with (5), and $L_n$ with $\log L_n + \log \psi(L_n) \lesssim \log p$, there exists a constant $K_3 > 0$ such that
\[
\sup_{\beta_0 \in \Delta_0(M, q; V, L_n; \Pi) \cap \Delta_3(\kappa_n)} \mathbb{E}_0 \Pi_0 \left\{ \beta : \|\beta - \beta_0\|_1 > \frac{K_3 s_0(\log p)^2}{\phi_1(K_1 s_0; W_0)^2} \right\} \to 0,
\]
\[
\sup_{\beta_0 \in \Delta_0(M, q; V, L_n; \Pi) \cap \Delta_3(\kappa_n)} \mathbb{E}_0 \Pi_0 \left\{ \beta : \|\beta - \beta_0\|_2 > \frac{K_3 s_0(\log p)^2}{\phi_2(K_1 s_0; W_0)^2} \right\} \to 0,
\]
\[
\sup_{\beta_0 \in \Delta_0(M, q; V, L_n; \Pi) \cap \Delta_3(\kappa_n)} \mathbb{E}_0 \Pi_0 \left\{ \beta : \|W_0 X (\beta - \beta_0)\|_2 > \frac{K_3 s_0(\log p)^2}{\phi_3(K_1 s_0; W_0)^2} \right\} \to 0,
\]
where $\Delta_0(M, q, V, L_n; \Pi)$ is defined as in Theorem 2 and $\Delta_3(\kappa_n)$ is the set of $\beta_0 \in \mathbb{R}^p$ such that $s_0^2(\log p)\|X\|_\infty^2 / \phi_3^2(K_1 s_0; W_0) \leq \kappa_n$.

Similar to the dimension-recovery result in Theorem 1, the constant in the threshold in the third assertion of Theorem 3 does not impose any condition on the design matrix. As mentioned earlier, our results are thus more in line with those in Section 4 of Castillo et al. (2015) for sparse linear regression. Note that the compatibility conditions appear in the contraction rates relative to the $\ell_1$- and $\ell_2$-norms, which make the implication of the theorem somewhat vague. As noted above, the compatibility numbers can be bounded away from zero under some conditions. The compatibility numbers are then removed from the rates, leading to $s_0(\log p)^2/n^{1/2}$ and $\{s_0 \log p\}/n^{1/2}$ for the $\ell_1$- and $\ell_2$-contraction rates, respectively.

### 4. APPLICATIONS

#### 4.1. Examples of the prior density

Below we provide examples of a prior density satisfying Assumptions 1–2. If the prior distribution sufficiently spreads out, the bound for $\beta_0$ can be made comparable to that for sparse linear regression in Section 4 of Castillo et al. (2015). Proofs of the claims made here are provided in the online supplementary material.

**Example 1 (Uniform).** We set $\pi = p$. For $\lambda$ such that $1/(B_1 p^{B_2}) \leq \lambda \leq B_3$ with some constants $B_1, B_2, B_3 > 0$, suppose that $g_S(\beta_S) = \lambda / 4^s$ if $\|\beta_S\|_\infty \leq 2\lambda^{-1}$, and $g_S(\beta_S) = 0$ otherwise. If $\log \tau_n(\beta_0) \lesssim \log p$, $\log \|X\|_\infty \lesssim \log p$, and $\|\beta_0\|_\infty \leq \lambda^{-1}$, then Assumption 1 is satisfied. If there exists $L_n$ such that $\log L_n + \log \psi(L_n) \lesssim \log p$ and $\lambda L_n \|X\|_\infty^{-1} \geq D s_0$ for a sufficiently large $D$, then Assumption 2 is also satisfied.
Example 2 (Laplace). We set $\bar{\pi} = \pi$. Suppose that $g_S(\beta_S) = (\lambda/2)^{e^{-\lambda ||\beta_S||_1}}$ with $\lambda$ assumed to satisfy $1/(B_1 p^{B_2}) \leq \lambda \leq B_3 \log p$ for some constants $B_1, B_2, B_3 > 0$. If $\log \gamma_n(\beta_0) \lesssim \log p$, $\log ||X||_{\infty} \lesssim \log p$, and $||\beta_0||_{\infty} \lesssim \lambda^{-1} \log p$, then Assumption 1 is satisfied. If there exists $L_n$ such that $\log L_n + \log \psi(L_n) \lesssim \log p$ and $\lambda L_n ||X||_{-\infty}^{-1} \geq D\lambda_0 \log p$ for a sufficiently large $D$, then Assumption 2 is also satisfied.

Example 3 (Elliptical Laplace). Consider $\bar{\pi}$ satisfying $s_0 \leq \bar{\pi} \leq p$ and positive definite matrices $\Sigma_S \in \mathbb{R}^{s \times s}$ such that for every $S$ with $s \leq \bar{\pi}$, $p^{-a} \leq \rho_{\min}(\Sigma_S) \leq \rho_{\max}(\Sigma_S) \leq p^a$ for some constant $a > 0$. We choose $g_S(\beta_S) = \lambda^s \pi^{s/2} \det(\Sigma_S)^{-1/2} \{\Gamma(s/2 + 1)/\Gamma(s + 1)\} \exp(-\lambda ||\Sigma_S^{-1/2} \beta_S||_2^2)$ with $\lambda$ assumed to satisfy $1/(B_1 p^{B_2}) \leq \lambda \leq B_3 \rho_{\min}^1(\Sigma_S) s_0^{-1/2} \log p$ for some constants $B_1, B_2, B_3 > 0$. If $\log \gamma_n(\beta_0) \lesssim \log p$, $\log ||X||_{\infty} \lesssim \log p$, and $||\beta_0||_{\infty} \lesssim \lambda^{-1} \rho_{\min}^1(\Sigma_S) s_0^{-1/2} \log p$, then Assumption 1 is satisfied. If there exists $L_n$ such that $\log L_n + \log \psi(L_n) \lesssim \log p$ and $\lambda L_n ||X||_{-\infty}^{-1} \min\{\rho_{\max}^{-1}(\Sigma_S) : s \leq \lambda_0\} \geq D\lambda_0^{3/2} \log p$ for a sufficiently large $D$, then Assumption 2 is also satisfied.

Example 4 (Multivariate normal). Consider $\bar{\pi}$ satisfying $s_0 \leq \bar{\pi} \leq p$ and positive definite matrices $\Sigma_S \in \mathbb{R}^{s \times s}$ such that for every $S$ with $s \leq \bar{\pi}$, $p^{-a} \leq \rho_{\min}(\Sigma_S) \leq \rho_{\max}(\Sigma_S) \leq p^a$ for some constant $a > 0$. We choose $g_S(\beta_S) = \lambda^s \pi^{s/2} \det(\Sigma_S)^{-1/2} \exp(-\lambda ||\Sigma_S^{-1/2} \beta_S||_2^2)$ with $\lambda$ assumed to satisfy $1/(B_1 p^{B_2}) \leq \lambda \leq B_3 \rho_{\min}^1(\Sigma_S) \log p$ for some constants $B_1, B_2, B_3 > 0$; that is, the normal density with mean zero and covariance matrix $\lambda^{-1} \Sigma_S$. If $\log \gamma_n(\beta_0) \lesssim \log p$, $\log ||X||_{\infty} \lesssim \log p$, and $||\beta_0||_{\infty} \lesssim \lambda^{-1} \rho_{\min}^1(\Sigma_S)(\log p)^{1/2}$, then Assumption 1 is satisfied. If there exists $L_n$ such that $\log L_n + \log \psi(L_n) \lesssim \log p$ and $\lambda L_n ||X||_{-\infty}^{-2} \min\{\rho_{\max}^{-1}(\Sigma_S) : s \leq \lambda_0\} \geq D\lambda_0^{3/2} \log p$ for a sufficiently large $D$, then Assumption 2 is also satisfied.

It is possible to impose weaker restrictions on $\lambda$, but the present formulation is used to focus on the size restriction on $||\beta_0||_{\infty}$, which is a constant in the worst case scenario. Then the conditions in the above examples become comparable for the case with $\Sigma_S = I_s$, though the need for an upper bound dependent on $s_0$ in Example 3 makes it less appealing. We also present results under simpler conditions on $\beta_0$. For example, with a multivariate normal prior, the stated condition on $\beta_0$ may be replaced by a more complicated one $\lambda ||\Sigma_S^{-1/2} \beta_{0,s_0}||_2^2 \lesssim s_0 \log p$. A close examination of the proofs given in the supplementary material reveals it.

The uniform prior in Example 1 is less appealing from a practical point of view as it restricts the possible values of the true parameter. However, for a fixed value of $\lambda$, the uniform prior more easily satisfies Assumption 2. For linear regression, the Laplace prior in Example 2 was shown to be useful to remove a size restriction on $\beta_0$ (Castillo et al., 2015). Although we resort to some restriction on $\beta_0$, one may see that the Laplace prior in our setup requires conditions comparable with those using the other densities. In Examples 3–4, if the densities are well defined for every $S$, there is no need to consider the restriction $\bar{\pi} < p$, and one may choose $\bar{\pi} = p$. A typical example is given by $\Sigma_S = I_s$. Another common choice $\Sigma_S = (X_S^T X_S)^{-1}$ needs the bound $\bar{\pi} \leq n$. For sparse linear regression, Gao et al. (2020) removed the compatibility conditions used in Castillo et al. (2015) using an elliptical Laplace distribution with this choice. A normal prior with $\Sigma_S = (X_S^T X_S)^{-1}$, often called a $g$-prior, has been widely studied in the literature (e.g., Liang et al., 2008).

Examples 1–4 provide general conditions for Assumptions 1–2 to be satisfied with each prior. Since the assumptions are directly associated with the function $h^{-1} y' \times q'$, a more precise description is dependent on its order in a given model. For example, $h^{-1} y' \times q'$ grows at most polynomially in logistic regression, probit regression, gamma regression, and negative binomial
regression, while it grows at most exponentially in Poisson regression; see Examples 5–9 below.

We make the following two remarks for more details.

**Remark 1 (On Assumption 1 with Examples 1–4).** If $(h^{-1})' \times \xi'$ grows at most polynomially, we have $\gamma_n(\beta_0) \lesssim \log p$ as soon as $\log \|X\beta_0\|_{\infty} \lesssim \log p$, which is trivially satisfied if $\log \|X\|_{\infty} + \log \|\beta_0\|_{\infty} \lesssim \log p$. In this case, for every prior in Examples 1–4, Assumption 1 is satisfied with any value of $\lambda$ lying on its range, as any $\lambda$ gives rise to $\log \|\beta_0\|_{\infty} \lesssim \log p$. The smallest possible values of $\lambda$ are particularly appealing as they relax the conditions the most. On the other hand, if $(h^{-1})' \times \xi'$ grows exponentially, we need the stronger bound $\|X\beta_0\|_{\infty} \lesssim \log p$ to have $\gamma_n(\beta_0) \lesssim \log p$. This is certainly not an issue if $s_0 \lesssim \log p$, as we can reasonably assume that $\|X\|_{\infty} \|\beta_0\|_{\infty} \lesssim 1$. In this situation, to satisfy Assumption 1 with each prior, we may need to use the largest possible value of $\lambda$, which imposes the restriction that $\|\beta_0\|_{\infty}$ is bounded. Even if $\log p = o(s_0)$, the restriction $\|X\beta_0\|_{\infty} \lesssim \log p$ can be met if there is some offset relation between $X$ and $\beta_0$. Note that stronger assumptions are usually made in the literature on high-dimensional generalized linear models; e.g., $\|\beta_0\|_1 \lesssim 1$ or $\|X\beta_0\|_{\infty} \lesssim 1$ (Ghosal, 1997; Jiang, 2007; Abramovich and Grinshtein, 2016).

**Remark 2 (On Assumption 2 with Examples 1–4).** If $(h^{-1})' \times \xi'$ grows at most polynomially, $L_n$ can be chosen to be a polynomial in $p$ with a sufficiently large exponent. Then for every prior in Examples 1–4, the boundedness requirement for Assumption 2 is easily satisfied with any $\lambda$ in its range, and hence Assumption 2 automatically follows under Assumption 1. If $(h^{-1})' \times \xi'$ grows at most exponentially, $L_n$ should be chosen to satisfy $L_n \lesssim \log p$, in which case the requirement for Assumption 2 is more restrictive. In this situation, the largest possible values of $\lambda$ are especially helpful to make the restriction as mild as possible. In all examples above, Assumption 2 is then satisfied as soon as $s_0 \|X\|_{\infty} \lesssim \log p$, with $\Sigma_S = I_s$ for Examples 3–4. Hence the condition $s_0 \lesssim \log p$ is essentially required, while Assumption 1 may be satisfied even if $\log p = o(s_0)$.

### 4.2. Examples of generalized linear models

We now study posterior contraction properties in specific examples of high-dimensional generalized linear models. Based on Remarks 1–2, we only need to check the order of $(h^{-1})' \times \xi'$ for each model. Then the results in Section 3 can be applied with an appropriate prior density.

**Example 5 (Logistic regression).** Note that $b(x) = \log(1 + e^x)$ and $h(x) = \log(x/(1 - x))$, and thus $\xi(x) = x$. Since $(h^{-1})'(x) = e^x/(1 + e^x)^2$, $(h^{-1})' \times \xi'$ is uniformly bounded.

**Example 6 (Probit regression).** Note that $b(x) = \log(1 + e^x)$ and $h(x) = \Phi^{-1}(x)$, and thus $(h^{-1})'(x) = \phi(x)$ and $\xi(x) = \log[\Phi(x)/(1 - \Phi(x))]$, where $\phi$ and $\Phi$ are the density and distribution function of the standard normal distribution, respectively. Note also that $\xi'(x) = [1/\Phi(x) + 1/(1 - \Phi(x))]\phi(x)$ grows linearly by the Mills ratio. Hence, $(h^{-1})' \times \xi'$ is uniformly bounded.

**Example 7 (Gamma regression with the log-link).** For gamma regression with a known shape parameter, the natural parameter is the negative rate parameter. Hence, $b(x) = -\log(-x)$ and $h(x) = \log x$, which give $(h^{-1})'(x) = e^x$ and $\xi(x) = -e^{-x}$. Hence, $(h^{-1})' \times \xi'$ is a constant.

**Example 8 (Negative binomial regression with the log-link).** With a given number of failures $r$, negative binomial regression for the number of successes possesses $b(x) = -r \log(1 - e^x)$, and hence $b'(x) = re^x/(1 - e^x)$. Since $h(x) = \log x$, we obtain $(h^{-1})'(x) = e^x$ and $\xi(x) = -\log(re^{-x} + 1)$ which gives $\xi'(x) = r/(r + e^x)$. Hence, $(h^{-1})' \times \xi'$ is uniformly bounded.
Example 9 (Poisson regression). Since \( b(x) = e^x \) and \( h(x) = \log x \), we have \( (h^{-1})'(x) = e^x \) and \( \xi(x) = x \). Thus, \( (h^{-1})' \times \xi' \) grows at most exponentially.

For Examples 5–8, any prior distribution in Examples 1–4 can be used to satisfy Assumptions 1–2, with the smallest values of \( \lambda \) to make the size restriction on \( \beta_0 \) as mild as possible. The main results in this study are thus easily applied and the desired posterior contraction rates are obtained for both the usual and fractional posteriors under the same conditions.

For Example 9, Assumption 1 is satisfied as soon as \( \|X\beta_0\|_\infty \lesssim \log p \), which is achieved more easily with the largest values of \( \lambda \) in the prior densities in Examples 1–4. Moreover, Remark 2 shows that Assumption 2 requires that a restriction on the true model size \( s_0 \) be practically necessary for every prior. Hence, the fractional posterior contraction rates obtained under milder conditions than the usual posterior contraction are also of considerable interest.

5. DISCUSSION

Although this study focuses on theoretical aspects, we provide a brief discussion on the computation of the posterior. Since a prior is usually not conjugate in generalized linear models, integrating out the parameter to compute the marginal posterior of \( S \) may not be possible. Some approximation methods such as the Laplace approximation may be useful in a relatively low-dimensional problem. Standard techniques such as the reversible-jump Markov chain Monte Carlo can be useful to fully explore the model space without resorting to approximations (Green, 1995). Some examples of the generalized linear models have latent variable expressions that facilitate the computation (e.g., Albert and Chib, 1993; Polson et al., 2013). One may refer to Dunson and Johndrow (2020) for a comprehensive overview of Markov chain Monte Carlo.

It is also worth comparing our rates with those studied in the frequentist literature. In high-dimensional generalized linear models with canonical link functions, Abramovich and Grinshtein (2016) obtained the convergence rate \( s_0 \log(ep/s_0) \) for their penalized estimator relative to the Kullback-Leibler risk under certain limited situations; see their Corollary 1. It can be seen that from the proof of their Lemma 1 (and also the proof of our Lemma 1), the Kullback-Leibler divergence of \( f_\beta \) from \( f_{\beta_0} \) can be matched with \( \|X(\beta - \beta_0)\|_2^2 \) up to constants under their condition that \( b''(X_T^T \beta_0) \) is bounded away from zero. Hence our rates match theirs up to logarithmic factors, but their results are based on generally stronger conditions than ours.

ACKNOWLEDGEMENT

Research is partially supported by Faculty Research and Professional Development Grant from the College of Sciences, North Carolina State University. The authors are grateful to Professors Ryan Martin and Anindya Roy for carefully reading the manuscript and suggestions for improvements in the presentation.

SUPPLEMENTARY MATERIAL

Supplementary material available at Biometrika online includes proofs of Lemma 1, Theorems 1–2, and the assertions in Examples 1–4.
Here and in the supplementary material, we write \( \eta_i = X_i \beta \) and \( \eta_{0i} = X_i \beta_0 \). We also write \( \theta_{0i} = \xi(\eta_{0i}) \) for the natural parameter with the true \( \beta_0 \). To prove Theorem 3, we first need the following lemma.

**Lemma A1.** Let \( \zeta_i(\eta_i) = H^2(f_{\beta,i}, f_{\beta_0,i}) \). Then there exist constants \( \delta_1, \delta_2 > 0 \) such that \( \zeta_i(\eta_i) \geq \zeta''(\eta_{0i}) \min \{ \delta_1(\eta_i - \eta_{0i})^2, \delta_2 \} \).

**Proof.** Note that it suffices to show that

(i) \( \zeta_i(\eta_i)/\zeta''(\eta_{0i}) \) is strongly convex only near the minimizer \( \eta_i = \eta_{0i} \) with its global minimum zero: \( \zeta_i(\eta_{0i}) = \zeta''(\eta_{0i}) = 0 \) and the second derivative of \( \zeta_i(\eta_i) \) is 1 at \( \eta_i = \eta_{0i} \);

(ii) \( \zeta_i(\eta_i)/\zeta''(\eta_{0i}) \) (or equivalently \( \zeta_i(\eta_i) \)) is strictly quasi-convex on \( \mathbb{R} \): \( \zeta_i(tx + (1 - t)y) < \max \{ \zeta_i(x), \zeta_i(y) \} \) for all \( x, y \in \mathbb{R} \) and \( t \in (0, 1) \).

Observe that (i) is easily checked, so we only need to verify (ii). For this, we can instead work with the function \( g_i(t) \) defined below. For any given \( \eta_i \neq \eta_{0i} \), let \( \tilde{\eta}_i^t = t\eta_i + (1 - t)\eta_{0i}, \ t \in [0, 1] \), and define

\[
g_i(t) = - \log \left( 1 - \frac{H^2(f_{\beta,i}, f_{\beta_0,i})}{2} \right) = \frac{1}{\tau_i} \left[ \frac{b_i(\xi(\eta_i^t) + b_i(\xi(\eta_{0i}))}{2} - b_i \left( \frac{\xi(\eta_i^t) + \xi(\eta_{0i})}{2} \right) \right],
\]

where \( f_{\beta,i} \) is the density with \( \tilde{\eta}_i^t \) as the parameter. Since \( g_i(0) = g_i^t(0) = 0 \) clearly, it suffices to show that \( g_i(t) \) is strictly increasing for every \( t \in (0, 1) \). Observe that by direct calculations,

\[
g_i'(t) = \left[ \frac{\xi'(\eta_i^t)}{2\tau_i} \right] b_i(\eta_i^t) - \left[ \frac{\xi'(\eta_i^t)}{2\tau_i} \right] (\eta_i - \eta_{0i}).
\]

Since \( \xi \) is strictly increasing, we have that \( \xi'(\eta_i^t) > 0 \) and \( \text{sgn} \{ \xi'(\eta_i^t) \} = \text{sgn}(\eta_i - \eta_{0i}) \). Note also that \( b' \) is strictly increasing as \( b''(x) > 0 \) for every \( x \in \Theta \). This gives that \( g_i'(t) > 0 \) for \( t \in (0, 1) \), which leads to the desired assertion. \( \square \)

**Proof of Theorem 3.** We now need to consider the set \( \mathcal{A}_n^\beta = \{ \beta \in \mathbb{R}^p : s_\beta \leq K_1 s_{\beta_0}, H_n(\beta, \beta_0) \leq K_2 \epsilon_n \} \) for \( \epsilon_n = \{ n/\log p \}/n^{1/2} \) as its complement is excluded by Theorems 1–2. For \( \delta_1, \delta_2 > 0 \) in Lemma A1, let \( \mathcal{I}_n = \{ i \leq n : \delta_1(\eta_i - \eta_{0i})^2 \geq \delta_2 \} \). Then, on \( \mathcal{A}_n^\beta \),

\[
K_2^2 \epsilon_n^2 \geq H_n^2(\beta, \beta_0) \geq \frac{\delta_1}{n} \sum_{i \in \mathcal{I}_n} \zeta_i''(\eta_{0i})(\eta_i - \eta_{0i})^2 + \frac{\delta_2}{n} \sum_{i \in \mathcal{I}_n} \zeta_i(\eta_{0i}).
\]

Using \( - \log \{ 1 - \zeta_i(\eta_i)/2 \} = g_i(1) \) with \( g_i \) in (A1), we can check that \( \zeta_i''(\eta_{0i}) = b''(\theta_{0i}) \{ \xi'(\eta_{0i}) \}^2 / 4 = w_{\beta_0}^2/4 \). It follows that on \( \mathcal{A}_n^\beta \),

\[
\frac{1}{n} \sum_{i \in \mathcal{I}_n} \zeta_i''(\eta_{0i})(\eta_i - \eta_{0i})^2 \geq \frac{1}{n} \sum_{i \in \mathcal{I}_n} \zeta_i''(\eta_{0i})(\eta_i - \eta_{0i})^2 - \frac{1}{n} \sum_{i \in \mathcal{I}_n} \zeta_i'(\eta_{0i}) \max_{1 \leq i \leq n} (\eta_i - \eta_{0i})^2 \geq \frac{\phi_1^2(K_1 s_{\beta_0}; W_{\beta_0})}{4K_1 s_{\beta_0}^2} \| \beta - \beta_0 \|^2 - \frac{K_2^2 \epsilon_n^2}{\delta_2} \| X \|^2 \| \beta - \beta_0 \|^2,
\]

where the inequality follows from the definition of \( \phi_1 \) and (A2). Since \( s_0^2(\log p)/\| X \|^2 > \phi_1^2(K_1 s_{\beta_0}; W_{\beta_0}) = o(n) \) on \( \Delta_\beta(\kappa_n) \) with any \( \kappa_n = o(n) \), the last display is further bounded below by a constant multiple of \( \phi_1^2(K_1 s_{\beta_0}; W_{\beta_0})/\| \beta - \beta_0 \|^2 \). This implies the first assertion of the theorem.

We now verify the other assertions. Using (A2) and (A3), note that on \( \mathcal{A}_n^\beta \),

\[
K_2^2 \epsilon_n^2 \geq H_n^2(\beta, \beta_0) \geq \frac{\delta_1}{n} \sum_{i \in \mathcal{I}_n} \zeta_i''(\eta_{0i})(\eta_i - \eta_{0i})^2 - \frac{\delta_1 K_2^2 \epsilon_n^2}{\delta_2} \| X \|^2 \| \beta - \beta_0 \|^2.
\]

By rearranging the terms, we obtain that \( \| W_0 X (\beta - \beta_0) \|^2 \leq n c_0^2 + n c_0^2 \| X \|^2 \| \beta - \beta_0 \|^2 \). Note that we have \( \| \beta - \beta_0 \|^2 \leq s_0^2(\log p)/\{ n \phi_1^2(K_1 s_{\beta_0}; W_{\beta_0}) \} \) by the first assertion of the theorem. Since
\[ s_0^2(\log p) \| X \|_2^2 / \phi_1^2(\kappa_0) \leq o(n) \] for any \( \alpha > 0 \), we see that \( \| W_0 X (\beta - \beta_0) \|^2 \) is bounded by a multiple of \( n \phi_2^2 \). This concludes the third assertion of the theorem. The second assertion then follows from the definition of \( \phi_2 \). \qed

REFERENCES


Li, Z., Song, Q., and Yu, K. (2013). Bayesian subset modeling for high-dimensional generalized linear models. 


Bernoulli, 26(3):2353–2382.


