ABSTRACT

JEONG, SEONGHYUN. Frequentist Properties of Bayesian Procedures for High-Dimensional Sparse Regression. (Under the direction of Subhashis Ghosal.)

High-dimensional sparse models have received a lot of attention during the last two decades, but their theoretical properties from the Bayesian point of view have been much less studied than those under the frequentist framework. This dissertation aims to fill these gaps by studying the frequentist properties of the Bayesian procedures for various high-dimensional sparse regression models.

Chapter 2 deals with a unified approach to Bayesian sparse regression when a possibly infinite-dimensional nuisance parameter is involved in a model. A mixture of a point mass at zero and a Laplace distribution is used for the prior distribution on sparse regression coefficients and an appropriate prior distribution is assigned to a nuisance parameter to yield nearly optimal posterior contraction. A shape approximation to the posterior distribution is characterized to show a semiparametric Bernstein-von Mises-type theorem and model selection consistency. Numerous examples are discussed including regression with partially sparse coefficients, multiple response models with missing components, multivariate measurement error models, mixed effects models, partial linear models, and nonparametric heteroskedastic regression models.

Chapter 3 considers high-dimensional logit models for categorical response variables under group sparsity. Similar to Chapter 2, a mixture prior consisting of a point mass at zero and a Laplace-type distribution is assigned to high-dimensional regression coefficients. Our study improves existing results in various directions. The procedure exhibits nearly optimal posterior contraction for the group sparse modeling. An approximation to the posterior distribution is also characterized to achieve model selection consistency. The distributional approximation also leads to a Bernstein-von Mises theorem for uncertainty quantification through credible sets with guaranteed frequentist coverage.

Chapter 4 focuses on posterior contraction rates for the high-dimensional generalized linear models with sparse priors. Again, a mixture prior is considered for high-dimensional regression coefficients, but a proper class of continuous distributions is considered for the slab-part in place of one single distribution. Our procedure incorporates non-canonical and canonical link functions; hence some standard, but non-canonical, models can also be considered, such as probit regression or gamma regression with the logarithmic link. The induced posterior contraction rates are nearly optimal. High-dimensional generalized linear models can be treated under this framework and their posterior contraction rates can be derived explicitly using the developed theory.
Frequentist Properties of Bayesian Procedures for High-Dimensional Sparse Regression

by
Seonghyun Jeong

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Statistics

Raleigh, North Carolina
2019

APPROVED BY:

____________________________________  ______________________________________
Arnab Maity                             Ryan Martin

____________________________________  ______________________________________
Yichao Wu                                Subhashis Ghoshal
                                           Chair of Advisory Committee
DEDICATION

To my parents.
BIOGRAPHY

Seonghyun Jeong was born and grew up in Seoul, South Korea. He received a B.A. in 2012 and an M.A. in 2014 from Yonsei University, Seoul. During his undergraduate studies, he served in the Republic of Korea Army for two years. Upon completion of his master’s study, he spent a year as a lecturer at Yonsei University. In 2015, he joined the Department of Statistics at North Carolina State University to pursue his Ph.D. in Statistics.
Although it is a cliché, I would like to begin by saying that this dissertation could not have been written without the help of many people. Pursuing a Ph.D. was an arduous, but rewarding, four-year journey. Along the way, I have indeed been helped by many professors, colleagues, and friends. And now, thanks to their support, the journey is about to end.

First and foremost, I would like to extend my deepest gratitude to my advisor Dr. Subhashis Ghoshal for his guidance, direction, and leadership. Without his continued support, this dissertation could never have been completed. Not to mention his expertise and academic success, I would like to especially pay homage to his patience in dealing with laymen, which is, in my opinion, a quality many experts do not have. The times I spent with him and the experiences I gained as his advisee will be invaluable assets throughout my academic career. There is no doubt that I owe him a debt I never can repay.

I was lucky enough to have another mentor, Dr. Ryan Martin, who never minded carving time out from his hectic schedule for me. He has always given me a great deal of valuable advice on both personal and professional aspects. I especially appreciate the help and advice he gave me concerning applying for academic positions and making my decision.

I am also grateful to my dissertation committee members, Dr. Arnab Maity and Dr. Yichao Wu. Their valuable participation in the advisory committee allowed me to view my dissertation topics from a more objective perspective and raise interesting questions from the non-Bayesian point of view. I thank them for being willing to invest time in me despite their busy schedules.

I would be remiss in my thanks if I failed to mention my master’s advisor, Dr. Taeyoung Park. I appreciate the generous support that he has continued to provide since my graduate studies in Korea. I also thank him for the collaborations we have built across the countries. The experiences and knowledge I gained by working with him have undoubtedly been helpful to me in terms of remaining in academia. I look forward to many more years of collaboration.

Among my many peers at NC State, I should mention my dear friend and housemate, Leonardo Cella. He has always been a good friend and colleague. The constructive discussions and titbits of gossip we shared every weekend were good refreshers. I look forward to the day when we achieve our goals in the respective areas. In addition, I would also like to thank all my friends and faculty members in the Department of Statistic at NC State. In particular, I want to express my gratitude to Eunah Cho for being there. She obviously made a major contribution to this dissertation.

Last but not at least, I wish to thank my parents for their unconditional love. I owe all my achievements to them. Without their extraordinary support, I would not be here now and none of these works would have been possible.
# TABLE OF CONTENTS

## Chapter 1 Introduction

1.1 Bayesian high-dimensional regression .......................... 1

1.2 Frequentist properties of Bayesian procedures .................. 2

1.2.1 Posterior contraction rates .................................. 2

1.2.2 Bernstein-von Mises theorems ............................... 3

1.2.3 Model selection consistency ................................. 4

1.3 Outline ...................................................... 4

## Chapter 2 A general framework for Gaussian sparse regression

2.1 Introduction .................................................. 5

2.2 Setup and prior specification .................................. 8

2.2.1 Notation .................................................. 8

2.2.2 Prior for the high-dimensional coefficients .................. 9

2.3 Main results ................................................ 10

2.3.1 Posterior contraction rates ................................ 10

2.3.2 Bernstein-von Mises and selection consistency ............. 16

2.4 Applications ............................................... 21

2.4.1 Regression with partially sparse coefficients ............... 21

2.4.2 Multiple response models with missing components ......... 22

2.4.3 Multivariate measurement error models .................... 23

2.4.4 Sparse regression with parametric correlation structure ... 24

2.4.5 Mixed effects models ...................................... 25

2.4.6 Graphical structure with sparse precision matrices ....... 26

2.4.7 Partial linear models ...................................... 27

2.4.8 Nonparametric heteroskedastic regression models ........... 28

2.5 Proofs ....................................................... 29

2.5.1 Proofs for the main results ................................ 29

2.5.2 Proofs for the applications ................................ 47

2.5.3 Auxiliary results ......................................... 67

## Chapter 3 Bayesian sparse regression for categorical responses

3.1 Introduction .................................................. 69

3.2 Setup and prior specification .................................. 72

3.2.1 Notation .................................................. 72

3.2.2 Prior specification ......................................... 74

3.3 Posterior contraction rates ................................... 76

3.3.1 Effective dimensionality ................................... 77

3.3.2 Posterior contraction with restricted design ............... 78

3.3.3 Posterior contraction for general prediction ............... 80

3.4 Bernstein-von Mises theorem and selection consistency ....... 81

3.4.1 Shape approximation to the posterior distribution ......... 81

3.4.2 Selection consistency ...................................... 85
3.4.3 Frequentist coverage of credible sets ........................................... 86
3.5 Proofs ......................................................................................... 87
   3.5.1 Preliminary results ................................................................. 87
   3.5.2 Proofs for the main results ..................................................... 88
   3.5.3 Proofs for the lemmas ............................................................. 99

Chapter 4 Posterior contraction in sparse generalized linear models ........ 110
   4.1 Introduction ............................................................................. 110
   4.2 Setup and prior specification ..................................................... 112
      4.2.1 Notation ........................................................................... 112
      4.2.2 Model assumption ............................................................... 113
      4.2.3 Prior specification ............................................................... 114
   4.3 Posterior contraction rates ........................................................ 116
   4.4 Proofs ..................................................................................... 120

References ..................................................................................... 129
Chapter 1

Introduction

1.1 Bayesian high-dimensional regression

Identifying weak and strong signals from a set of coefficients in regression models has been of great interest. Despite its early popularity for simple models, the study of model selection in high-dimensions started only in the 2000s, when the storage and processing of massive data became possible in practice. Recently, model selection has received more attention in terms of high-dimensional rather than low-dimensional problems, as sparsity classes are necessary for sensible estimation of high-dimensional models. From the frequentist perspective, there is a substantial body of literature on this topic; see Fan and Lv (2010), Bühlmann and van de Geer (2011), and the references therein for a complete overview.

While Bayesian model selection for classical low-dimensional problems has a long history, Bayesian methodologies for sparse estimation in high-dimensional regression models were developed much later; see Bondell and Reich (2012), Johnson and Rossell (2012), and Narisetty and He (2014) for consistent Bayesian model selection methods in high-dimensional linear models. Extensive theoretical investigations involving frequentist asymptotic properties of posterior distributions of high-dimensional models have been conducted only very recently and are limited to relatively simple settings. In a pioneering work, Castillo et al. (2015) considered the simplest form of sparse linear regression with a known error variance. They showed that for such a model class, the Bayesian procedure based on a mixture of point-mass at zero and a Laplace distribution prior possesses a convergence rate that is comparable to the lasso rate except for compatibility conditions (Bühlmann and van de Geer, 2011). They also established a Bernstein-von Mises theorem and derived the selection consistency of their Bayesian procedure. Martin et al. (2017) and Belitser and Ghosal (2019) studied an equivalent model with a known error variance, but using an empirical conjugate normal prior that induces the minimax posterior contraction rates. In particular, oracle inequalities derived in Belitser and Ghosal (2019)
give rise to optimal-size credible sets with guaranteed frequentist coverage. Song and Liang (2017) considered a variety of shrinkage priors that produce posterior asymptotic properties comparable to those obtained in Castillo et al. (2015), although uniformity is possible only on certain norm-bounded subsets of the parameter space. In Song and Liang (2017), the unappealing assumption of a known error variance was relaxed by putting an inverse gamma prior on it. Posterior contraction rates in some high-dimensional nonlinear regression models are considered in Atchadé (2017).

This dissertation studies frequentist asymptotic properties of various sparse regression models under the Bayesian framework. The primary objective is to characterize the theoretical properties of the posterior distributions of high-dimensional sparse regression coefficients, but nuisance parameters can also be of interest if they are present in a model. As for the frequentist properties of the posterior distributions, posterior contraction rates, Bernstein-von Mises theorems, and model selection consistency are considered, all of which are discussed briefly in the following section.

1.2 Frequentist properties of Bayesian procedures

1.2.1 Posterior contraction rates

For every $n \geq 1$, consider an observation $X^{(n)}$ in a sample space $(X^{(n)}, X^{(n)})$ with distribution $P_{\theta}^{(n)}$ parameterized by $\theta$ that lies on a topological parameter space $\Theta$. With a given prior distribution on $\theta$, let $\Pi(\cdot|X^{(n)})$ be a version of the posterior distribution, which we call ‘the’ posterior distribution in this section. A posterior contraction rate at the true parameter $\theta_0$ with respect to the semimetric $d$ is a sequence $\epsilon_n$ such that for any $M_n \to \infty$,

$$\Pi(\theta : d(\theta, \theta_0) \geq M_n \epsilon_n | X^{(n)}) \to 0, \text{ in } P_{\theta_0}^{(n)}-probability.$$  

The fastest decreasing $\epsilon_n$ satisfying this definition is of particular interest. For a slightly stronger definition, an arbitrarily increasing sequence $M_n$ is often replaced by a sufficiently large constant. Posterior contraction rates are Bayesian analogues of frequentist rates of convergence; the two are closely related. In fact, it is known that if a posterior contraction rate $\epsilon_n$ is attained, there exists an estimator that exhibits the same rate of convergence. While many studies have been done on specific statistical models, Ghosal et al. (2000) and Ghosal and van der Vaart (2007) are especially useful for general theories. For a comprehensive review of the topic, we refer the reader to Chapter 8 of Ghosal and van der Vaart (2017).

The Hellinger metric or the root-average-squared Hellinger metric is one of the most popular choices for the semimetric $d$ due to its exponentially powerful test functions (Ghosal et al., 2000; Ghosal and van der Vaart, 2007). It is still of interest to choose $d$ such that the closeness of the
parameter is well interpreted (e.g., Giné and Nickl, 2011; Castillo, 2014). In this dissertation, since we are primarily interested in high-dimensional regression coefficients with a Euclidean parameter space, the most preferred metric is the \( \ell_q \)-norm, \( 1 \leq q < \infty \). Our results show that the rates are adaptive to an unknown sparsity level. Nonetheless, we mainly formulate pointwise results for simplicity. See Chapter 8 of Giné and Nickl (2016) and Chapter 10 of Ghosal and van der Vaart (2017) for adaptive inference.

### 1.2.2 Bernstein-von Mises theorems

In many regular parametric statistical models, a good estimator \( \hat{\theta} \) of the Euclidean parameter \( \theta \) is asymptotically normally distributed with mean \( \theta \) and covariance matrix proportional to the inverse Fisher information \( I_0 \). A Bayesian analogue to this optimal property is called a Bernstein-von Mises theorem, which states that the posterior distribution is well approximated by the normal distribution with the corresponding mean and variance. Following Le Cam and Yang (1990), we formulate this property precisely as

\[
\left\| \mathbb{P}(\theta \in \cdot | X^{(n)}) - N(\hat{\theta}, (nI_0)^{-1}) \right\|_{TV} \to 0, \quad \text{in } P_{\theta_0}^{(n)}-probability,
\]

where \( \| \cdot \|_{TV} \) denotes the total variation metric and \( \hat{\theta} \) is an efficient estimator. One of the main benefits of characterizing Bernstein-von Mises theorems is that they guarantee frequentist coverage of credible sets; that is, remaining uncertainty on the parameter can be correctly quantified through the posterior distribution. We refer to Chapter 10 of van der Vaart (2000) for further discussion on the parametric Bernstein-von Mises theorems.

In nonparametric models, neither the frequentist asymptotic normality nor the Bernstein-von Mises theorem is fully valid (Freedman, 1999). Instead, versions of Bernstein-von Mises theorems for low-dimensional functionals or Euclidean parameters in semiparametric models, called semiparametric Bernstein-von Mises theorems, are usually of interest. There is now a substantial body of work on this topic (e.g., Bickel and Kleijn, 2012; Castillo, 2012; Castillo and Nickl, 2013, 2014). For a complete review of semiparametric Bernstein-von Mises theorems, we refer to Chapter 12 of Ghosal and van der Vaart (2017) and the references contained therein.

The form of Bernstein-von Mises theorems we consider is similar to (1.1), but differs in that the posterior distribution of high-dimensional regression coefficients is approximated by a product of a normal distribution and a point mass rather than a single normal distribution. This difference is essential in high-dimensional settings as we assume that the number of parameters increases faster than the sample size. If a nuisance parameter is present in a model, the marginal posterior distribution of regression coefficients is the main substance for approximation, as in semiparametric Bernstein-von Mises theorems for strictly semiparametric models.
1.2.3 Model selection consistency

Throughout this dissertation, high-dimensional regression coefficients are assumed to be sparse, where ‘sparsity’ means that only a few of the coefficients are nonzero and that many of them are (exactly or close to) zero. In addition to the frequentist properties discussed thus far, we are also interested in how well the posterior distribution recovers the true sparsity class. Denoting the support of \( \theta = (\theta_j) \) by \( S_\theta \), i.e., \( S_\theta = \{ j : \theta_j \neq 0 \} \), the desired property can be written as

\[
\Pi(\theta : S_\theta \neq S_{\theta_0}|X^{(n)}) \to 0, \quad \text{in } P_{\theta_0}^{(n)}\text{-probability},
\]

which is referred to as the model selection consistency. Unlike posterior contraction rates and Bernstein-von mises theorems, model selection consistency is not a general asymptotic objective for arbitrary statistical models but instead is tailored for variable selection problems. High-dimensional models are the archetypes for which this property is particularly of interest, as sparsity classes are essential for sensible estimation. A frequentist counterpart has also received a great deal of attention in the literature (e.g., Zhao and Yu, 2006).

1.3 Outline

The remainder of this dissertation is organized as follows. Chapter 2 considers a general framework for sparse linear regression models in the presence of nuisance parameters. Nuisance parameters can be finite-, high-, or infinite-dimensional depending on the setup. Our primary interest lies in the frequentist properties of the posterior distribution of high-dimensional regression coefficients, but we also study posterior contraction rates for nuisance parameters, which is particularly interesting if the nuisance parameters are not finite-dimensional objects. In Chapter 3, we study high-dimensional logit models for categorical response variables. In particular, group sparsity is considered in order to account for the nature of multinomial regression models. This chapter extends the recent findings on the frequentist properties of Bayesian procedures for logistic regression in many directions. Chapter 4 studies posterior contraction rates for high-dimensional generalized linear models. A suitable class of sparse prior distributions is considered for obtaining the desired posterior contraction rates. The models with both non-canonical and canonical link functions are of interest.
Chapter 2

A general framework for Gaussian sparse regression

2.1 Introduction

It is natural to consider generalizations of simple sparse regression models to account for the intricate characteristics of certain data. For example, in a longitudinal or spatial setting, the response variables may form groups with correlation while independence is assumed between groups. In the partial linear models setting, a baseline effect is included and modeled nonparametrically. Further extensions adding more complications in modeling will be considered in the examples below.

We formulate a general framework to treat sparse regression models in a unified way as follows. Let $\eta$ be possibly an infinite-dimensional nuisance parameter taking values in a set $\mathbb{H}$. For each $\eta \in \mathbb{H}$ and an integer $m_i \in \{1, \ldots, m\}$ for some $m \geq 1$, suppose that there are a vector $\xi_{\eta,i} \in \mathbb{R}^{m_i}$ and a positive definite matrix $\Delta_{\eta,i} \in \mathbb{R}^{m_i \times m_i}$ which define a regression model for a response variable $Y_i \in \mathbb{R}^{m_i}$ against covariates $X_i \in \mathbb{R}^{m_i \times p}$ given by

$$Y_i = X_i \theta + \xi_{\eta,i} + \varepsilon_i, \quad \varepsilon_i \sim \text{ind } N(m_i(0, \Delta_{\eta,i})), \quad i = 1, \ldots, n,$$

(2.1)

where $\theta \in \mathbb{R}^p$ is a vector of regression coefficients. Here $m_i$ (and $m$) can increase with $n$. We are primarily interested in the high-dimensional situation where $p$ can be much larger than $n$, but $\theta$ is assumed to be sparse, with many coordinates zero. The form in (2.1) clearly includes sparse linear regression with a known error variance (Castillo et al., 2015; Martin et al., 2017; Belitser and Ghosal, 2019) and with an unknown error variance (Song and Liang, 2017). Indeed, the model (2.1) includes many more interesting examples considered below.

Example 2.1 (Regression with partially sparse coefficients). For a response variable $Y_i \in \mathbb{R}^m$
with some fixed $\overline{m} \geq 1$ and covariates $X_i \in \mathbb{R}^{\overline{m} \times p}$ and $Z_i \in \mathbb{R}^{\overline{m} \times q}$, consider a regression model, $Y_i = X_i \theta + Z_i \beta + \varepsilon_i$, $\varepsilon_i \ind N_{\overline{m}}(0, \Sigma)$, $i = 1, \ldots, n$, where $\theta \in \mathbb{R}^p$ is a sparse coefficient vector with $p > n$, $\beta \in \mathbb{R}^q$ is a vector of nonzero regression coefficients with possibly increasing $q \geq 1$, and $\Sigma \in \mathbb{R}^{\overline{m} \times \overline{m}}$ is a positive definite covariance matrix. In this model, the relatively low-dimensional vector $\beta$ is not assumed to be sparse, but it needs to be estimated along with other parameters. The model belongs to the form of (2.1) by letting $\xi_{\eta,i} = Z_i \beta$ and $\Delta_{\eta,i} = \Sigma$ with $\eta = (\beta, \Sigma)$.

**Example 2.2** (Multiple response models with missing components). We consider a general multiple response model with missing values, which is very common in practice. Suppose that for each $i$, a vector of $\overline{m}$ responses with covariance matrix $\Sigma$ are supposed to be observed, but for the $i$th group (or subject) only $m_i$ entries are actually observed with the rest missing. Letting $Y_i \in \mathbb{R}^{m_i}$ be the $i$th observations and $Y_i^{\text{aug}} \in \mathbb{R}^{\overline{m}}$ be the augmented vector of $Y_i$ and missing entries, we can write $Y_i = E_i^{T}Y_i^{\text{aug}}$ and $\text{Cov}(Y_i) = E_i^{T}\Sigma E_i$, where $E_i \in \mathbb{R}^{\overline{m} \times m_i}$ is the submatrix of the $\overline{m} \times \overline{m}$ identity matrix with the $j$th column included if the $j$th element of $Y_i^{\text{aug}}$ is observed, $j = 1, \ldots, \overline{m}$. Assuming that the mean of $Y_i$ is only $X_i \theta$ for covariates $X_i \in \mathbb{R}^{m_i \times p}$ and sparse coefficients $\theta \in \mathbb{R}^p$ with $p > n$, the model of interest can be written as $Y_i = X_i \theta + \varepsilon_i$, $\varepsilon_i \ind N_{m_i}(0, E_i^{T}\Sigma E_i)$, $i = 1, \ldots, n$. The model belongs to the class described by (2.1) with $\xi_{\eta,i} = 0_{m_i}$ and $\Delta_{\eta,i} = E_i^{T}\Sigma E_i$ with $\eta = \Sigma$.

**Example 2.3** (Multivariate measurement error models). Suppose that a scalar response variable $Y_i^* \in \mathbb{R}$ is connected to fixed covariates $X_i^* \in \mathbb{R}^p$ with $p > n$ and random covariates $Z_i \in \mathbb{R}^q$ with fixed $q \geq 1$, through the following linear additive relationship: $Y_i^* = \alpha + X_i^{*T}\theta + Z_i^{T}\beta + \varepsilon_i^*$, $Z_i \ind N_q(\mu, \Sigma)$, $\varepsilon_i^* \ind N(0, \sigma^2)$. While $X_i^*$ is fully observed without noise, we observe a surrogate $W_i$ of $Z_i$ as $W_i = Z_i + \tau_i$, $\tau_i \ind N_q(0, \Psi)$, where to ensure identifiability, $\Psi$ is assumed to be known. This type of model is called a measurement error model or an errors-in-variables model; see Fuller (1987) and Carroll et al. (2006) for a complete overview. By direct calculations, the joint distribution of $(Y_i^*, W_i)$ is given by

$$
\left( \begin{array}{c} Y_i^* \\ W_i \end{array} \right) \ind N_{q+1} \left( \begin{array}{c} \alpha + X_i^{*T}\theta + \mu^{T}\beta \\ \mu \end{array} \right), \left( \begin{array}{rr} \beta^{T}\Sigma\beta + \sigma^2 & \beta^{T}\Sigma \\ \Sigma\beta & \Sigma + \Psi \end{array} \right)
$$

which is of the form (2.1) by writing $Y_i^T = (Y_i^*, W_i^T)$, $X_i^T = (X_i^{*T}, 0_{p \times q})$, $\xi_{\eta,i}^T = (\alpha + \mu^{T}\beta, \mu^{T})$, and $\Delta_{\eta,i} = \left( \begin{array}{c} \beta^{T}\Sigma\beta + \sigma^2 \\ \Sigma\beta \end{array} \right)$ with $\eta = (\alpha, \beta, \mu, \sigma^2, \Sigma)$.

**Example 2.4** (Sparse regression with parametric correlation structure). For a response variable $Y_i \in \mathbb{R}^{m_i}$ we consider a standard regression model given by $Y_i = X_i \theta + \varepsilon_i$, $\varepsilon_i \ind N_{m_i}(0, \Sigma_i)$, $i = 1, \ldots, n$, but for flexibility, $m_i$ is considered to be possibly increasing. As $m_i$ can increase with $n$ and vary across $i$, some restriction on $\Sigma_i$ is required for sensible estimation. For a known parametric correlation structure $G_i$ and a fixed dimensional Euclidean parameter $\alpha$, we model
the covariance matrix as $\Sigma_i = \sigma^2 G_i(\alpha)$ using a variance parameter $\sigma^2$ and a correlation matrix $G_i(\alpha) \in \mathbb{R}^{m_i \times m_i}$. Then the model belongs to (2.1) by writing $\xi_{\eta,i} = 0_{m_i}$ and $\Delta_{\eta,i} = \sigma^2 G_i(\alpha)$ with $\eta = (\alpha, \sigma^2)$.

**Example 2.5** (Mixed effects models). For a response variable $Y_i \in \mathbb{R}^{m_i}$ and covariates $X_i^* \in \mathbb{R}^p$ with $p > n$ and $Z_i^* \in \mathbb{R}^q$ with fixed $q \geq 1$, consider a mixed effect model given by $Y_i = X_i \theta + Z_i b_i + \varepsilon_i^*, b_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \Psi), \varepsilon_i^* \overset{\text{iid}}{\sim} \mathcal{N}_{m_i}(0, \sigma^2 I_{m_i})$, where $\Psi \in \mathbb{R}^{q \times q}$ is a positive definite matrix. Then the marginal law of $Y_i$ is given by $Y_i = X_i \theta + \varepsilon_i, \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}_{m_i}(0, \sigma^2 I_{m_i} + Z_i \Psi Z_i^T)$. We assume that $\sigma^2$ is known. The model belongs to the class (2.1) by letting $\xi_{\eta,i} = 0_{m_i}$ and $\Delta_{\eta,i} = \sigma^2 I_{m_i} + Z_i \Psi Z_i^T$ with $\eta = \Psi$.

**Example 2.6** (Graphical structure with sparse precision matrices). For a response variable $Y_i \in \mathbb{R}^m$ and covariates $X_i \in \mathbb{R}^{m \times p}$ with increasing $m \geq 1$ and $p > n$, consider a model given by $Y_i = X_i \theta + \varepsilon_i, \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}_{m}(0, \Omega^{-1})$, $i = 1, \ldots, n$, where $\theta$ is a sparse coefficient vector and the precision matrix $\Omega \in \mathbb{R}^{m \times m}$ is a positive definite matrix. Along with $\theta$, we also impose sparsity on the off-diagonal entries of $\Omega$ accounts for a graphical structure between observations. More precisely, if an off-diagonal entry is zero, it implies the conditional independence between the two concerned entries of $\varepsilon_i$ given the remaining ones, and we suppose that most off-diagonal entries are actually zero even though we do not know their locations. The model is then seen to be a special case of (2.1) by letting $\xi_{\eta,i} = 0_m$ and $\Delta_{\eta,i} = \Omega^{-1}$ with $\eta = \Omega$.

**Example 2.7** (Partial linear models). Consider a partial linear model given by $Y_i = X_i^T \theta + g(z_i) + \varepsilon_i, \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), i = 1, \ldots, n$, where $Y_i \in \mathbb{R}$ is a response variable, $X_i \in \mathbb{R}^p$ is a vector of covariates with $p > n$, $\theta \in \mathbb{R}^p$ is a sparse coefficient vector, $g : [0, 1] \mapsto \mathbb{R}$ is a univariate function, and $z_i \in [0, 1]$ is a scalar predictor. This model is expressed in the form (2.1) by writing $\xi_{\eta,i} = g(z_i)$ and $\Delta_{\eta,i} = \sigma^2$ with $\eta = (g, \sigma^2)$.

**Example 2.8** (Nonparametric heteroskedastic regression models). For a response variable $Y_i \in \mathbb{R}$ and covariates $X_i \in \mathbb{R}^p$, a linear regression model with a nonparametric heteroskedastic error is given by $Y_i = X_i^T \theta + \varepsilon_i, \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, v(z_i)), i = 1, \ldots, n$, where $\theta$ is a sparse coefficient vector, $v : [0, 1] \mapsto (0, \infty)$ is a univariate function, and $z_i \in [0, 1]$ is a one-dimensional variable associated with the $i$th observation that controls the variance of $Y_i$ through the variance function $v$. Then the model belongs to the class (2.1) by letting $\xi_{\eta,i} = 0$ and $\Delta_{\eta,i} = v(z_i)$ with $\eta = v$.

In this chapter, we study a unified theory of posterior asymptotics in the high-dimensional sparse regression models described by the form (2.1). To the best of our knowledge, there is no available study that considers a general modeling framework of sparse regression as in (2.1) even from the frequentist perspective. The results on complicated high-dimensional regression models are only available at model-specific levels and cannot be universally used for different classes of model structure; for example, see Müller and van de Geer (2015) for the frequentist
convergence rates that are tailored for sparse partial linear models. On the other hand, our approach is a unified theoretical treatment of the general model structure in (2.1) under the Bayesian framework. We establish general theorems on nearly optimal posterior contraction rates, a Bernstein-von Mises theorem via shape approximation to the posterior distribution of \( \theta \), and model selection consistency. We then apply the general theorems to derive specific results for the examples given above.

The general theory of posterior contraction using the root-average-squared Hellinger metric on the joint density (Ghosal and van der Vaart, 2007) is not very useful in this context, since to recover rates in terms of the metric of interest on the regression coefficients, some boundedness conditions are needed. To deal with this issue, we construct an exponentially powerful likelihood ratio test in small pieces that are sufficiently separated from the true parameters in terms of the average Rényi divergence of order \( 1/2 \) (which coincides with the average negative log-affinity). This test provides posterior contraction relative to the corresponding divergence. The posterior contraction rates of \( \theta \) and \( \eta \) can then be recovered in terms of the metrics of interest under milder conditions on the parameter space. Our results imply that the posterior contraction rates are adaptive to unknown sparsity level.

For a Bernstein-von Mises theorem and selection consistency, stronger conditions are required than those used for posterior contraction, in line with the existing literature (e.g., Castillo et al., 2015; Martin et al., 2017). A key to these arguments is to find a suitable orthogonal projection that satisfies the required conditions, which is typically straightforward if the support of a prior for \( \xi_{\eta,i} \) is a linear space. The complexity of the covariance matrix, measured by its metric entropy, also has an important role in deriving a Bernstein-von Mises theorem and selection consistency.

The rest of this chapter is organized as follows. In Section 2.2, some notations are introduced and a prior distribution on sparse regression coefficients is specified. Section 2.3 provides our main results on the posterior contraction, the Bernstein-von Mises phenomenon, and selection consistency of the posterior distribution. In Section 2.4, these general theorems are applied to the examples considered in this section to derive posterior asymptotic properties in each specific example. All technical proofs are provided in Section 2.5.

### 2.2 Setup and prior specification

#### 2.2.1 Notation

Here we describe the notations we use throughout this chapter. For a vector \( \theta = (\theta_j) \in \mathbb{R}^p \) and a set \( S \subset \{1, \ldots, p\} \) of indices, we write \( S_{\theta} = \{ j : \theta_j \neq 0 \} \) to denote the support of \( \theta \), \( s := |S| \) (or \( s_{\theta} := |S_{\theta}| \)) to denote the cardinality of \( S \) (or \( S_{\theta} \)), and \( \theta_S = \{ \theta_j : j \in S \} \) and
$\theta_{S^c} = \{\theta_j : j \notin S\}$ to separate components of $\theta$ using $S$. In particular, the support of the true parameter $\theta_0$ and its cardinality are written as $S_0$ and $s_0 := |S_0|$, respectively. As for norms of a vector $\theta$, $||\theta||_q = (\sum_j |\theta_j|^q)^{1/q}$, $1 \leq q < \infty$, stands for the $\ell_q$-norms and $||\theta||_{\infty} = \max_j |\theta_j|$ denotes the maximum norm. For a matrix $X = (x_{ij})$, we write $\rho_{\min}(X)$ and $\rho_{\max}(X)$ for the minimum and maximum eigenvalues, respectively. The column space of $X$ is denoted by $\text{span}(X)$. Let $||X||_{sp} = \rho_{\max}^{1/2}(X^T X)$ stand for the spectral norm and $||X||_F = (\sum_{i,j} x_{ij}^2)^{1/2}$ stand for the Frobenius norm of $X$. We also define a matrix norm $||X||_s = \max_j ||X_j||_2$ for the $j$th column of $X$, which is used for compatibility conditions. For sequences $a_n$ and $b_n$, $a_n \lesssim b_n$ (or $b_n \gtrsim a_n$) stands for $a_n \leq C b_n$ for some constant $C > 0$ independent of $n$, and $a_n \asymp b_n$ means $a_n \lesssim b_n \lesssim a_n$. These inequalities are also used for relations involving constant sequences.

For given parameters $\theta$ and $\eta$, we write the joint density as $p_{\theta,\eta} = \prod_{i=1}^n p_{\theta_i,\eta_i}$ for $p_{\theta_i,\eta_i}$ the density of the $i$th observation vector $Y_i$. In particular, the true joint density is expressed as $p_0 = \prod_{i=1}^n p_{0,i}$ for $p_{0,i} := p_{\theta_0,\eta_0, i}$ with the true parameters $\theta_0$ and $\eta_0$. The notation $\mathbb{E}_0$ denotes the expectation operator with the true density $p_0$. For two probability measures $P$ and $Q$, let $\|P - Q\|_{\text{TV}}$ denote the total variation between $P$ and $Q$. For two $n$-variate densities $f := \prod_{i=1}^n f_i$ and $g := \prod_{i=1}^n g_i$ of independent variables, denote the average Rényi divergence (of order $1/2$) by $R_n(f, g) = -n^{-1} \sum_{i=1}^n \log \int f_i^{1/2} g_i^{1/2}.$

For any $\eta_1, \eta_2 \in \mathbb{H}$, we define the squared pseudo-metric $d^2_A(\eta_1, \eta_2) = d^2_{A,n}(\eta_1, \eta_2) + d^2_{B,n}(\eta_1, \eta_2)$ for the two squared pseudo-metrics

$$d^2_{A,n}(\eta_1, \eta_2) = \frac{1}{n} \sum_{i=1}^n ||\xi_{n,i} - \xi_{n,j}||^2_2, \quad d^2_{B,n}(\eta_1, \eta_2) = \frac{1}{n} \sum_{i=1}^n ||\Delta_{\eta_1,i} - \Delta_{\eta_2,i}||^2_F.$$  

For compatibility conditions, the uniform compatibility number $\phi_1$ and the smallest scaled singular value $\phi_2$ are defined as

$$\phi_1(s) = \inf_{0 \leq |S| \leq s} \frac{||X\theta||_2|S\theta|^{1/2}}{||X||_s ||\theta||_1}, \quad \phi_2(s) = \inf_{0 \leq |S| \leq s} \frac{||X\theta||_2}{||X||_s ||\theta||_2}.$$ 

We write $Y^{(n)} = (Y_1^n, \ldots, Y_n^n)^T$ for the observation vector, $n_s = \sum_{i=1}^n m_i$ for the dimension of $Y^{(n)}$, and $\Theta = \mathbb{R}^p$ for the parameter space of $\theta$. Lastly, for a (pseudo-)metric space $(\mathcal{F}, d)$, let $N(\epsilon, \mathcal{F}, d)$ denote the $\epsilon$-covering number, the minimal number of $\epsilon$-balls that cover $\mathcal{F}$.

### 2.2.2 Prior for the high-dimensional coefficients

In this subsection, we specify a prior distribution for the high-dimensional regression coefficients $\theta$. A prior for $\eta$ also needs to be specified, but it is deferred to Section 2.3 as it should be designed together with the main results to guarantee desirable posterior asymptotic properties, while the prior for $\theta$ specified here is always good for such a purpose.
We first select a dimension $s$ from a prior $\pi_p$, and then randomly choose $S \subset \{1, \ldots, p\}$ for given $s$. A nonzero part $\theta_S$ of $\theta$ is then selected from a prior $g_S$ on $\mathbb{R}^s$ while $\theta_{S^c}$ is fixed to zero. The resulting prior specification for $(S, \theta)$ is formulated as
\begin{equation}
(S, \theta) \mapsto \frac{\pi_p(s)}{p} g_S(\theta_S) \delta_0(\theta_{S^c}), \tag{2.2}
\end{equation}
where $\delta_0$ is the Dirac measure at zero on $\mathbb{R}^{p-s}$ with suppressed dimensionality. For the prior $\pi_p$ on the model dimensions, we consider a prior satisfying the following: for some constants $A_1, A_2, A_3, A_4 > 0$,
\begin{equation}
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, p. \tag{2.3}
\end{equation}
Examples of priors satisfying (2.3) can be found in Castillo and van der Vaart (2012) and Castillo et al. (2015). For the prior $g_S$, the $s$-fold product of the Laplace density is considered, where the regularization parameter is allowed to vary with $p$ and $\|X\|_*$, i.e.,
\begin{equation}
g_S(\theta_S) = \prod_{j \in S} \frac{\lambda}{2} \exp(-\lambda |\theta_j|), \quad \frac{\|X\|_*}{L_1 p L_2} \leq \lambda \leq \frac{L_3 \|X\|_*}{\sqrt{n}}, \tag{2.4}
\end{equation}
for some constants $L_1, L_2, L_3 > 0$. The order of $\lambda$ is important in that it determines the boundedness requirement of the true signal $\theta_0$ (see the condition (C3) below). A particularly interesting case is that $\lambda$ is set to the lower bound $\|X\|_*/(L_1 p L_2)$. Then the boundedness condition becomes very mild by choosing $L_2$ sufficiently large. The upper bound is motivated such that the boundedness condition still makes a practical sense. However, it can actually be relaxed if the true signal is known to be small enough, though we do not pursue this generalization in this chapter. In Section 2.3.2, we shall see that values of $\lambda$ that do not increase too fast are in fact necessary for distributional approximation and selection consistency.

We expect that prior distributions possessing suitable tail properties should also work similarly. In particular, one can refer to Chapter 4 of this dissertation for posterior contraction with some class of prior distributions.

2.3 Main results

2.3.1 Posterior contraction rates

The prior for a nuisance parameter $\eta$ should be chosen to complete the prior specification. Once we assign the prior for the full parameters, the posterior distribution $\Pi(\cdot | Y^{(n)})$ is defined by Bayes’ rule. How the prior for $\eta$ is chosen is crucial to obtain desirable asymptotic properties.
of the posterior distribution. In this subsection, we shall examine such conditions on the prior distribution for a nuisance parameter and study the posterior convergence rates for both \( \theta \) and \( \eta \).

The prior for \( \eta \) is put on a subspace \( \mathcal{H} \subset \mathbb{H} \). In many instances, we take \( \mathcal{H} = \mathbb{H} \), especially when a nuisance parameter is finite dimensional, but the flexibility of a subspace may be beneficial in infinite-dimensional situations. We need to choose \( \mathcal{H} \) to satisfy certain conditions.

(C1) There exists a nondecreasing sequence \( a_n = o(n) \) such that

\[
a_n \max_{1 \leq i \leq n} \| \Delta_{\eta',i} - \Delta_{\eta_0,i} \|_F^2 =: e_n \to 0, \quad \text{for some } \eta' \in \mathcal{H},
\]

\[
\max_{1 \leq i \leq n} \| \Delta_{\eta_1,i} - \Delta_{\eta_2,i} \|_F^2 \leq a_n d^2_{B,n}(\eta_1, \eta_2), \quad \eta_1, \eta_2 \in \mathcal{H}.
\]

(C2) For some sequence \( \bar{\epsilon}_n \) such that \( a_n \bar{\epsilon}_n^2 \to 0 \) and \( n \bar{\epsilon}_n^2 \to \infty \) with \( a_n \) satisfying (C1),

\[
\log \Pi (\eta \in \mathcal{H} : d_n(\eta, \eta_0) \leq \bar{\epsilon}_n) \gtrsim -n \bar{\epsilon}_n^2.
\]

The first condition of (C1) implies that we have a good approximation to the true parameter value in the parameter set \( \mathcal{H} \). This holds trivially if there exists \( \eta' \in \mathcal{H} \) such that \( \Delta_{\eta',i} = \Delta_{\eta_0,i} \) for every \( i \leq n \), which is obviously true if \( \eta_0 \in \mathcal{H} \). The second condition of (C1) means that in \( \mathcal{H} \), the maximum Frobenius norms of the difference between covariance matrices can be controlled by the average Frobenius norm multiplied by the sequence \( a_n \). Clearly, this holds with \( a_n = 1 \) if \( \Delta_{\eta,i} \) is the same for every \( i \leq n \). By the triangle inequality, we see that (C1) explicitly implies that

\[
\max_{1 \leq i \leq n} \| \Delta_{\eta,i} - \Delta_{\eta_0,i} \|_F^2 \lesssim e_n + a_n d^2_{B,n}(\eta, \eta_0), \quad \eta \in \mathcal{H}, \tag{2.5}
\]

which is used throughout the chapter. The condition (C2) is typically called the prior concentration condition, which requires a prior to put sufficient mass around the true parameter \( \eta_0 \), measured by the pseudo-metric \( d_n \). As in other infinite-dimensional situations, such a closeness is translated into the closeness in terms of the Kullback-Leibler divergence and variation (see Lemma 2.1 in Section 2.5 for more details).

As noted in Section 2.1, the true parameters should be restricted to certain norm-bounded subset of the parameter space, which is clarified as follows.

(C3) The true signal satisfies \( \| \theta_0 \|_\infty \lesssim \lambda^{-1} \log p \).

(C4) The eigenvalues of the true covariance matrix satisfy

\[
1 \lesssim \min_{1 \leq i \leq n} \rho_{\min}(\Delta_{\eta_0,i}) \leq \max_{1 \leq i \leq n} \rho_{\max}(\Delta_{\eta_0,i}) \lesssim 1.
\]
The condition (C3) is required to apply the general strategies for posterior contraction to our modeling framework containing nuisance parameters. If nuisance parameters are not present, one can directly handle the model and such a restriction may be removed (e.g., Castillo et al., 2015; Gao et al., 2015). One may refer to Song and Liang (2017) for a similar condition to ours, where a nuisance parameter is an unknown error variance. Still, the condition is mild if \( \lambda \) is chosen to decrease at a proper order. In particular, if \( \lambda \) is matched to the lower bound \( 1/(L_1 p^{L_2}) \), the condition becomes \( \| \theta_0 \|_\infty \lesssim (p^{L_2} \log p)/\| X \|_{*} \) which is very mild if \( L_2 \) is sufficiently large. Even if the upper bound \( L_3 \| X \|_{*}/\sqrt{n} \) is chosen, the condition is not restrictive as the right hand side of the condition can be made nondecreasing as long as \( \| X \|_{*} \) is increasing at a suitable order. We also mention that the condition (C3) is actually slightly stronger than what it needs to be. It suffices if we have \( \lambda \| \theta_0 \|_1 \leq (s_0 \log p) \lor n\bar{\epsilon}^2_n \) for \( \bar{\epsilon}_n \) satisfying (C2), but the given bound is adopted for ease of interpretation for the true signal. The condition (C4) implies that the eigenvalues of the true covariance matrix are bounded below and above. The lower and upper bounds are required for a lot of technical details, including the construction of an exponentially powerful test in Lemma 2.2.

To study posterior contraction, we first need to examine a dimensionality property of the support of \( \theta \). The following theorem shows that the posterior distribution is concentrated on models of relatively small sizes.

**Theorem 2.1** (Dimension). Suppose that (C1)–(C4) are satisfied. Then there exists a constant \( K_1 \) such that

\[
\mathbb{E}_0 \Pi \left( \theta : s_\theta > K_1 \left\{ s_0 \lor \frac{n\bar{\epsilon}^2_n}{\log p} \right\} \bigg| Y^{(n)} \right) \to 0.
\]

Compared to the literature (e.g., Castillo et al., 2015; Martin et al., 2017; Belitser and Ghosal, 2019), the rate in Theorem 2.1 is floored by the extra term \( n\bar{\epsilon}^2_n / \log p \). This arises from the presence of a nuisance parameter in the model formulation. To minimize its impact, a prior on \( \eta \) should be chosen such that (C2) holds for as small \( \bar{\epsilon}_n \) as possible; a suitable choice induces the (nearly) optimal contraction rate.

Using the basic results in Theorem 2.1, the next theorem shows that the posterior distribution contracts at the truth with respect to the average Rényi divergence. The theorem requires additional assumptions on a prior.

**C5** There exists a subset \( \mathcal{H}_n \subset \mathcal{H} \) such that for \( s_* := s_0 \lor (n\bar{\epsilon}^2_n / \log p) \) with \( \bar{\epsilon}_n \) satisfying (C2), a sufficiently large \( B > 0 \), and some sequences \( \gamma_n \) and \( \epsilon_n \geq \sqrt{s_* \log(p \lor m \lor \gamma_n)} / n \)
The condition (2.6) requires that for every \( i \leq n \), the minimum eigenvalue of \( \Delta_{\eta,i} \) is not too small on a sieve \( \mathcal{H}_n \). Although \( \gamma_n \) can be any positive sequence, a sequence increasing exponentially fast makes the entropy in (2.7) too large, resulting in a suboptimal rate \( \epsilon_n \). If \( \gamma_n \) can be chosen to be smaller than \( p \) and \( \bar{m} \), then this does not lead to any deterioration of rate in \( \epsilon_n \). It should be noted that the entropy condition (2.7) is actually stronger than what is really needed. Scrutinizing the proof of the theorem, one can see that the entropy appearing in the theorem is obtained using pieces that are smaller than those giving the exponentially powerful test in Lemma 2.2 in Section 2.5. However, the covering number with those pieces looks more complicated and the form in (2.7) suffices for all examples in the present chapter. Lastly, the condition (2.8) implies that the outside of a sieve \( \mathcal{H}_n \) should possess sufficiently small prior mass to kill the factor \( s_* \log p \) arising from the lower bound of the denominator of the posterior distribution (see Lemma 1). In fact, conditions similar to (C2), (2.7) and (2.8) are also required for the prior of \( \theta \). By reading the proof, it is easy to see that the prior (2.2) explicitly satisfies the analogous conditions on an appropriately chosen sieve.

**Theorem 2.2** (Contraction rate, Rényi). Suppose that (C1)–(C5) are satisfied. Then there exists a constant \( K_2 \) such that

\[
E_0 \Pi \left( R_n(p_{\theta,\eta}, p_0) > K_2 \epsilon_n^2 \right) \rightarrow 0.
\]

Although Theorem 2.2 provides the basic results for posterior contraction, it does not give precise interpretations for the parameters \( \theta \) and \( \eta \) themselves, because of the abstruse expression of the average Rényi divergence. The contraction rates with respect to more concrete metrics are recovered under some additional conditions. Under the additional assumption \( a_n \epsilon_n^2 \rightarrow 0 \), it can be shown that Theorem 2.1 and Theorem 2.2 explicitly imply that for the set

\[
\mathcal{A}_n = \left\{ (\theta, \eta) \in \Theta \times \mathcal{H} : s_\theta \leq K_1 s_* \ , \sum_{i=1}^n \| X_i (\theta - \theta_0) + \xi_{\eta,i} - \xi_{\eta_0,i} \|_2^2 + nd_{B_{\mathcal{H}}(\eta, \eta_0)}^2(n) \leq M_1 \epsilon_n^2 \right\},
\]

with a sufficiently large constant \( M_1 \), the posterior mass of \( \mathcal{A}_n \) goes to one in probability (see
the proof of Theorem 2.3). To complete the recovery, we need to separate the sum of squares of the mean into $\|X(\theta - \theta_0)\|_2$ and $nd^2_{A,n}(\eta, \eta_0)$, which requires an additional condition. The conditions required for the recovery are clarified as follows.

(C5*) The condition (C5) holds with some $\epsilon_n \geq \sqrt{s_\star \log(p \vee m \vee \gamma_n)/n}$ such that $a_n \epsilon_n^2 \to 0$ with $a_n$ satisfying (C1).

(C6) For $s_\star$ and $\epsilon_n$ satisfying (C5*), there exists $\eta_\star \in H_n$ such that
\[
\lim \inf_{n \geq 1} \inf_{(\theta, \eta) \in A_n} \frac{\sum_{i=1}^n (\theta - \theta_0)^T X_i^T (\xi_{\eta,i} - \xi_{\eta_\star,i})}{\|X(\theta - \theta_0)\|_2^2 + nd^2_{A,n}(\eta, \eta_\star)} > -\frac{1}{2}, \quad d_{A,n}(\eta_\star, \eta_0) \lesssim \epsilon_n.
\]

By expanding the quadratic term for the mean in $A_n$, one can see that the separation is possible if (C6) is satisfied. Clearly, (C6) is trivially satisfied if the model has only $X\theta$ for its mean, in which we take $\xi_{\eta,i} - \xi_{\eta_\star,i} = 0$ for every $i \leq n$. In many cases where there exists $\eta' \in H$ such that $d_{A,n}(\eta', \eta_0) = 0$, we can often take $\eta_\star = \eta'$ for the second inequality of (C6) to hold automatically.

The following theorem shows that the posterior distribution of $\theta$ and $\eta$ contracts to the truth at some rates, relative to more easily comprehensible metrics than the average Rényi divergence.

**Theorem 2.3 (Recovery).** Suppose that (C1)–(C4), (C5*), and (C6) are satisfied. Then, there exists a constant $K_3$ such that
\[
E_0 \Pi \left( \frac{s_\star n \epsilon_n^2}{\phi_2^2(K_1 s_\star + s_0)} | Y^{(n)} \right) \to 0,
\]
\[
E_0 \Pi \left( \frac{n \epsilon_n^2}{\phi_2^2(K_1 s_\star + s_0)} | Y^{(n)} \right) \to 0,
\]
\[
E_0 \Pi \left( \frac{\|X(\theta - \theta_0)\|_2^2}{K_3 n \epsilon_n^2} | Y^{(n)} \right) \to 0,
\]
\[
E_0 \Pi \left( d_n(\eta, \eta_0) > K_3 \epsilon_n | Y^{(n)} \right) \to 0.
\] (2.9)

**Remark 2.1.** The separation condition (C6) can be left as an assumption to be satisfied, but can also be verified by a stronger condition on the design matrix without resorting to the values of the parameters. Suppose that for some integer $q \geq 1$, there exists a matrix $Z_i \in \mathbb{R}^{m_i \times q}$ such that $\xi_{\eta,i} = Z_i h(\eta)$ for every $\eta \in H$, with some map $h : H \mapsto \mathbb{R}^q$. Since we can write $\xi_{\eta,i} - \xi_{\eta_\star,i} = Z_i (h(\eta) - h(\eta_\star))$ for any $\eta, \eta_\star \in H$, the Cauchy-Schwarz inequality indicates that
the first inequality of \((C6)\) is implied by
\[
\liminf_{n \geq 1} \inf_{(\theta, n) \in \Theta \times H; s_0 \leq K_1s} \frac{(\theta - \theta_0)^T X^T Z(h(\eta) - h(\eta_0))}{\|X(\theta - \theta_0)\|_2 \|Z(h(\eta) - h(\eta_0))\|_2} > -1,
\]
for \(Z = (Z_1^T, \ldots, Z_n^T)^T\). The left hand side is always between \(-1\) and 1 by the Cauchy-Schwarz inequality, and is exactly equal to \(-1\) or 1 if and only if the two vectors are linearly dependent. A sufficient condition for the preceding display is thus \(\min\{\rho_{\min}\{[X_S, Z]^T[X_S, Z]\} : s \leq K_1s + s_0\} \geq 1\) since the linear dependence cannot happen under such a condition due to the inequality \(s_\theta - \theta_0 \leq s_\theta + s_0 \leq K_1s + s_0\) for \(\theta\) such that \(s_\theta \leq K_1s\). This sufficient condition is not restrictive at all if \(q = o(n)\) as we already have \(K_1s + s_0 = o(n)\). Since there typically exists \(\eta_0 \in H\) satisfying the second inequality of \((C6)\) as long as \(H\) provides a good approximation for the true parameter \(\eta_0\), the condition \((C6)\) can be easily satisfied if the sufficient condition is met.

In most instances, the sequence \(\gamma_n\) in \((C5)\) can be chosen such that \(\log \gamma_n \lesssim \log p\). This is trivially satisfied if \(\gamma_n\) is some polynomial in \(n\) as in the examples in this chapter. If \(p\) is known to increase much faster than \(n\), e.g., \(\log p \asymp n^c\) for some \(c \in (0, 1)\), then \(\gamma_n\) needs not be a polynomial in \(n\) and the condition can be met more easily with a sequence that grows even faster. Note also that we typically have \(\log m \lesssim \log p\) in most cases. Moreover, it is often possible to choose \(\epsilon_n \asymp \sqrt{(s_\star \log p)/n}\), which is commonly guaranteed by choosing an appropriate sieve \(\mathcal{H}_n\) and a prior. For the optimal recovery of \(\theta\) and \(\eta\), we make the above assumptions precise in the following modifications of \((C5)\) and \((C6)\).

\((C5)^\dagger\) The condition \((C5)\) holds with \(\gamma_n\) and \(\epsilon_n \asymp \sqrt{(s_\star \log p)/n}\) such that \(\log \gamma_n \lesssim \log p\), \(s_\star \log m \lesssim n\epsilon_n^2\), and \(a_n\epsilon_n^2 \to 0\) with \(a_n\) satisfying \((C1)\).

\((C6)^*\) The condition \((C6)\) holds with \(s_\star\) and \(\epsilon_n\) satisfying \((C5)^\dagger\).

**Corollary 2.1** (Optimal recovery). *Suppose that \((C1)-(C4)\), \((C5)^\dagger\), and \((C6)^*\) are satisfied. Then, there exists a constant \(K_4\) such that*

\[
\begin{align*}
\mathbb{E}_0 \Pi \left( \|\theta - \theta_0\|_1 > K_4 \frac{s_\star \sqrt{\log p}}{\phi_1(K_1s + s_0)\|X\|_s} \right| Y^{(n)} \right) & \to 0, \\
\mathbb{E}_0 \Pi \left( \|\theta - \theta_0\|_2 > K_4 \frac{\sqrt{s_\star \log p}}{\phi_2(K_1s + s_0)\|X\|_s} \right| Y^{(n)} \right) & \to 0, \\
\mathbb{E}_0 \Pi \left( \|X(\theta - \theta_0)\|_2 > K_4 \frac{s_\star \log p}{\sqrt{n}} \|X\|_s \right| Y^{(n)} \right) & \to 0, \\
\mathbb{E}_0 \Pi \left( d_n(\eta; \eta_0) > K_4 \frac{s_\star \log p}{n} \right| Y^{(n)} \right) & \to 0.
\end{align*}
\]
The thresholds for contraction depend upon the compatibility conditions, which make their implication somewhat vague. As $K_1s_* + s_0$ is much smaller than $n_*$, it is not unreasonable to assume that $\phi_1(K_1s_*+s_0)$ or $\phi_2(K_1s_*+s_0)$ is bounded away from zero, whence the compatibility number is removed from the rates. We refer to Example 7 of Castillo et al. (2015) for more discussion. In the next subsection, we will see that this restriction is actually necessary for shape approximation or selection consistency.

### 2.3.2 Bernstein-von Mises and selection consistency

The posterior contraction properties of $\theta$ established in the previous subsection explicitly imply that the posterior distribution of $\theta$ contracts to the truth with respect to the $\ell_1$- or $\ell_2$-norm. However, this does not indicate model selection consistency, the property that the posterior distribution of the model $S_\theta$ is concentrated at the true support $S_0$ with probability tending to one. To show this, we need additional arguments which are based on distributional approximation to the posterior distribution.

It is worth noting that selection consistency can often be verified without distributional approximation (see, for example, Song and Liang (2017)). This is possible only when we know a lot more about the marginal posterior distribution of $\theta$. However, in our general formulation, information on the exact marginal posterior of $\theta$ is rarely available due to an arbitrary structure of a nuisance parameter $\eta$ and its prior distribution. Hence using a shape approximation is a natural solution to the problem, which may require some extra conditions on the parameter space and the priors on $\theta$ and $\eta$. First of all, we assume the following condition on the uniform compatibility number.

(C7) For $s_*$ satisfying (C5*), $\phi_1(K_1s_* + s_0)$ is bounded away from zero.

This condition is weaker than assuming that the smallest scaled singular value $\phi_2(K_1s_*+s_0)$ is bounded away from zero, as we have $\phi_1(s) \geq \phi_2(s)$ for any $s > 0$ by the Cauchy-Schwarz inequality. In this sense, our condition is weaker than those in Theorem 4 of Castillo et al. (2015). The condition (C7) is not restrictive as (C5*) requires $s_* = o(n)$ (refer to Example 7 of Castillo and van der Vaart (2012) for a discussion).

The assumption on the prior for $\theta$ is made only through the regularization parameter $\lambda$. As in Castillo et al. (2015), $\lambda$ should not increase too fast and should satisfy $\lambda \epsilon_n \sqrt{n s^*} / \|X\|_* \rightarrow 0$ for distributional approximation. In our setup, the range of $\lambda$ induces a sufficient condition for this: $s_\epsilon^2 \rightarrow 0$. However, this sufficient condition is weaker than one that will be made later in this section, and hence the “small lambda regime” is automatically met by a stronger condition for the entire procedure for distributional approximation (see (C12) and the following paragraph).
To describe the conditions on the prior for \( \eta \), we define additional notations used throughout this subsection. In what follows, we write

\[
\hat{X} = \begin{pmatrix}
\Delta_{\eta_0,1}^{-1/2} X_1 \\
\vdots \\
\Delta_{\eta_0,n}^{-1/2} X_n
\end{pmatrix},
\tilde{\xi}_\eta = \begin{pmatrix}
\Delta_{\eta_0,1}^{-1/2} \tilde{\xi}_{\eta,1} \\
\vdots \\
\Delta_{\eta_0,n}^{-1/2} \tilde{\xi}_{\eta,n}
\end{pmatrix},
U = \begin{pmatrix}
\Delta_{\eta_0,1}^{-1/2} (Y_1 - X_1 \theta_0 - \xi_{\eta_0,1}) \\
\vdots \\
\Delta_{\eta_0,n}^{-1/2} (Y_n - X_n \theta_0 - \xi_{\eta_0,n})
\end{pmatrix},
\]

and \( \tilde{\Delta}_\eta \) to denote the collection of \( \Delta_{\eta,i} \) for \( i = 1, \ldots, n \). In particular, \( \hat{X}_S \) denotes the submatrix of \( \hat{X} \) with columns chosen by an index set \( S \). Moreover, for sufficiently large constants \( M_2 \) and \( M_3 \), define the sets

\[
\tilde{\Theta}_n = \{ \theta \in \Theta : s_\theta \leq K_1 s_*, \| \theta - \theta_0 \|_1^2 \leq M_2 s_\epsilon n \epsilon_n^2 / \| X \|_*^2 \},
\tilde{\mathcal{H}}_n = \{ \eta \in \mathcal{H} : d_{A,n}^2 (\eta, \eta_0) \leq M_3 s_\epsilon \epsilon_n^2, d_{B,n}^2 (\eta, \eta_0) \leq M_3 \epsilon_n^2 \},
\]

with the sequences \( s_* \) and \( \epsilon_n \) satisfying (C5\textsuperscript{*}) so that the posterior distribution can be concentrated on these sets with probability tending to one by Theorem 2.3, combined by other conditions. Let \( \Phi(\eta) = (\tilde{\xi}_\eta, \tilde{\Delta}_\eta) \). For a given \( \theta \), we choose a bijective map \( \eta \mapsto \tilde{\eta}_n(\theta, \eta) \) such that \( \Phi(\tilde{\eta}_n(\theta, \eta)) = (\tilde{\xi}_\eta + H \hat{X} (\theta - \theta_0), \tilde{\Delta}_\eta) \) for some orthogonal projection \( H \) which may depend on the true parameter values, but not on \( \theta \) and \( \eta \). The projection \( H \) plays a key role in distributional approximation and should be appropriately chosen to satisfy the followings.

(C8) For \( s_* \) and \( \epsilon_n \) satisfying (C5\textsuperscript{*}), the orthogonal projection \( H \) satisfies

\[
s_* \epsilon_n^2 \sup_{\eta \in \tilde{\mathcal{H}}_n} \| (I - H)(\tilde{\xi}_\eta - \tilde{\xi}_{\eta_0}) \|_2^2 \to 0.
\]

(C9) The conditional law \( \Pi_{n,\theta} \) of \( \tilde{\eta}_n(\theta, \eta) \) given \( \theta \), induced by the prior, is absolutely continuous relative to its distribution \( \Pi_{n,\theta_0} \) at \( \theta = \theta_0 \) (which is the same as the prior for \( \eta \)), and the Radon-Nikodym derivative \( d\Pi_{n,\theta} / d\Pi_{n,\theta_0} \) satisfies for \( s_* \) and \( \epsilon_n \) satisfying (C5\textsuperscript{*}),

\[
\sup_{\theta \in \tilde{\Theta}_n} \sup_{\eta \in \tilde{\mathcal{H}}_n} \left| \log \frac{d\Pi_{n,\theta}}{d\Pi_{n,\theta_0}} (\eta) \right| \to 0.
\]

(C10) For \( s_* \) satisfying (C5\textsuperscript{*}), there exists a constant \( D_1 > 0 \) such that

\[
\liminf_{n \to \infty} \min_{s : s \leq K_1 s_*} \inf_{v \in \mathbb{R}^s : \| v \|_2 = 1} \left\{ D_1 \| (I - H) \hat{X} v \|_2 - \| \hat{X} v \|_2 \right\} > 0.
\]

The condition (C8) is required to make the remainder of the approximation tend to zero with probability tending to one. The condition (C9) implies that the prior on \( \eta \) is nearly invariant under a shift in certain directions, and hence it justifies the use of the shifting map \( \eta \mapsto \tilde{\eta}_n(\theta, \eta) \).
The condition (C10) implies that \( \|u\|_2 \lesssim \|(I - H)u\|_2 \leq \|u\|_2 \) for every \( u \in \text{span}(\tilde{X}_S) \) with \( S \) such that \( s \leq K_1 s_* \), as the second inequality trivially holds by the fact that \( I - H \) is an orthogonal projection. This also allows \( \tilde{X}_S^T(I - H)\tilde{X}_S \) to be positive definite for every \( S \) such that \( s \leq K_1 s_* \), so that a shape approximation by normal distributions can be well defined.

It is worth noting that for regression models where no additional mean part \( \xi, \eta, i \) exists, \( H \) can simply be chosen to be the zero matrix, whence the conditions (C8) and (C9) are trivially satisfied and (C10) is translated into the condition of full-column rank of \( X_S \) for \( S \) such that \( s \leq K_1 s_* \).

Lastly, we also assume the followings.

(C11) The set \( \mathcal{H}_n \) is a separable space with the pseudo-metric \( d_{B,n} \).

(C12) For a sufficiently large \( D_2 > 0 \), \( a_n \) and \( e_n \) satisfying (C1), and \( s_* \) and \( \epsilon_n \) satisfying (C5\#),

\[
\begin{align*}
  s_* n \epsilon_n^2 & + a_n \epsilon_n^2 + \sqrt{n \epsilon_n^2 a_n s_* \log p} \int_0^{D_2 \epsilon_n} \sqrt{\log N (\delta, \mathcal{H}_n, d_{B,n})} d\delta \to 0.
\end{align*}
\]

Similar to (C8), these conditions are also required to make the remainder of the approximation tend to zero. In the display in (C12), the integral term comes from the expected supremum of a separable Gaussian process, exploiting the Gaussian likelihood of the model and the separability of \( \mathcal{H}_n \) with the standard deviation metric. The condition (C11) is crucial for this reason, and is trivially satisfied if \( \mathcal{H} \) is a separable pseudo-metric space with \( d_{B,n} \). Since we usually put a prior on \( \eta \) in an explicit way, the condition (C11) is rarely violated in practice. We also mention that (C12) is a sufficient condition for the small lambda regime, as the first term of the condition is larger than \( ns_* \epsilon_n^2 \). From this, it is also easy to see that a necessary condition for (C12) is \( s_*^5 \log^3 p = o(n) \), which is also a sufficient condition in many finite dimensional models.

Under the assumptions above, the posterior distribution of \( \theta \) is approximated by \( \Pi^\infty \) given by

\[
\Pi^\infty(\theta \in \cdot | Y^{(n)}) = \sum_{S: s \leq K_1 s_*} \hat{w}_S N \left( \theta_S | \hat{\theta}_S, (\tilde{X}_S^T(I - H)\tilde{X}_S)^{-1} \right) \otimes \delta_0 (\theta_{S^c}),
\]

where \( \hat{\theta}_S \) is the least squares solution \( \hat{\theta}_S = (\tilde{X}_S^T(I - H)\tilde{X}_S)^{-1} \tilde{X}_S^T(I - H)(U + \tilde{X}\theta_0) \) and the weights satisfy

\[
\hat{w}_S \propto \frac{\pi_p(s)}{p} \left( \frac{\lambda}{2} \right)^s (2\pi)^{s/2} \det \left( \tilde{X}_S^T(I - H)\tilde{X}_S \right)^{-s/2} \exp \left\{ \frac{1}{2\lambda} \|I - H\tilde{X}_S^T\theta_S\|^2 \right\}.
\]
Another way to express $\Pi^\infty$, for any measurable $B \subset \mathbb{R}^p$, is

$$
\Pi^\infty(\theta \in B \mid Y^{(n)}) = \frac{\sum_{S:s \leq K_1 s_*} \pi_p(s)(p_s)^{-1} \left( \lambda/2 \right)^s \int_B \Lambda_n^*(\theta) d\{\kappa(\theta_S) \otimes \delta_0(\theta_S^c)\}}{\sum_{S:s \leq K_1 s_*} \pi_p(s)(p_s)^{-1} \left( \lambda/2 \right)^s \int_{\mathbb{R}^p} \Lambda_n^*(\theta) d\{\kappa(\theta_S) \otimes \delta_0(\theta_S^c)\}},
$$

where $\kappa$ denotes the Lebesgue measure and

$$
\Lambda_n^*(\theta) = \exp \left\{ -\frac{1}{2} \left\| (I - H) \tilde{X}(\theta - \theta_0) \right\|^2_2 + U^T (I - H) \tilde{X}(\theta - \theta_0) \right\}.
$$

(2.13)

It can be easily checked that both the expressions are equivalent. The results are summarized in the following theorem.

**Theorem 2.4 (Distributional approximation).** Suppose that (C1)–(C4), (C5*), and (C6)–(C12) are satisfied for some orthogonal projection $H$. Then

$$
E_0 \left\| \Pi(\theta \in \cdot \mid Y^{(n)}) - \Pi^\infty(\theta \in \cdot \mid Y^{(n)}) \right\|_{TV} \to 0.
$$

(2.14)

**Remark 2.2.** Similar to Remark 2.1, suppose that there exists a matrix $Z_i \in \mathbb{R}^{m_i \times q}$ such that $\xi_{\eta,i} = Z_i h(\eta)$ for every $\eta \in \mathcal{H}$ with some map $h : \mathcal{H} \mapsto \mathbb{R}^q$. Then, a general strategy to choose $H$ is to set $H = \tilde{Z} (\tilde{Z}^T \tilde{Z})^{-1} \tilde{Z}^T$ for $\tilde{Z} = (Z_1 T \Delta_{\eta_1,1}^{-1/2}, \ldots, Z_n T \Delta_{\eta_n,1}^{-1/2})^T$. In this case, the condition (C8) is satisfied if there exists $\eta_* \in \mathcal{H}$ such that $s_* \xi_*^2 \epsilon_n^2 d_{A,n}(\eta_*, \eta_0) \to 0$, which is trivially the case if $d_{A,n}(\eta', \eta_0) = 0$ for some $\eta' \in \mathcal{H}$. If not, since $s_* \xi_*^2 \epsilon_n^2$ can increase fast, we may need a good decay for $d_{A,n}(\eta_*, \eta_0)$. This may result in suboptimal contraction rates; see Section 2.4.7 for example. Also, similar to Remark 2.1, a sufficient condition for (C10) is $\min \{ \rho_{\min}([X_S, Z]^T [X_S, Z]) : s \leq K_1 s_* \} \gtrsim 1$ as pre-multiplication of a positive definite matrix by $X_S$ and $Z$ is an isometry.

**Remark 2.3.** In many instances, for every $\delta > 0$, the entropy in (C12) is typically bounded above by a multiple of $0 \vee r_n \log(b_n \epsilon_n / \delta)$ for some sequences $r_n$ and $b_n$, especially when the part of $\eta$ involved with $d_{B,n}$ is an $r_n$-dimensional Euclidean parameter, whence the term in (C12) is simplified. Note that in this case, the integral is bounded by a multiple of

$$
\int_0^{(D_2 \wedge b_n) \epsilon_n} \sqrt{r_n \log \frac{b_n \epsilon_n}{\delta}} \, d\delta
$$

$$
= (D_2 \wedge b_n) \epsilon_n \sqrt{r_n \log \frac{b_n}{D_2 \wedge b_n}} + b_n \epsilon_n \sqrt{r_n} \int_0^\infty e^{-t^2} \, dt.
$$

If $b_n$ is increasing, the right hand side is bounded by a multiple of $\epsilon_n \sqrt{r_n \log b_n}$ by the tail probability of a normal distribution, while it is bounded by a multiple of $\epsilon_n b_n \sqrt{r_n}$ for nonincreasing
This simplification can be useful in many applications.

The shape approximation to the posterior distribution facilitates obtaining the next theorem which shows that the posterior distribution is concentrated on subsets of the true support with probability tending to one. The result is then used as the basis of selection consistency. The theorem requires an additional condition on the prior as follows.

(C13) For \( s_* \) satisfying (C5*), assume that \( A_4 > 1 \) and \( s_* \lesssim p^a \) with \( a < A_4 - 1 \).

**Theorem 2.5** (Selection, no supersets). Suppose that (C1)–(C4), (C5*), and (C6)–(C13) are satisfied for some orthogonal projection \( H \). Then

\[
E_0 \Pi \left( \theta : S_\theta \supset S_0, S_\theta \neq S_0 \mid Y^{(n)} \right) \to 0. \tag{2.15}
\]

Since the coefficients that are too close to zero cannot be identified by any selection strategy, some threshold for the true nonzero coefficients are needed for detection. We make the following threshold, the so-called beta-min condition.

(C14) For \( \epsilon_n \) and \( s_* \) satisfying (C5*), assume that

\[
\min_{\theta_0,j \neq 0} |\theta_0,j|^2 > \frac{K_3n\epsilon_n^2}{\phi_2^2(K_1s_* + s_0)\|X\|^2}. 
\]

Since Theorem 2.3 implies that the posterior distribution of the support of \( \theta \) includes that of the true support with probability tending to one, selection consistency is an easy consequence of Theorem 2.5 under the beta-min condition (C14). Moreover, this improves the distributional approximation in (2.14) so that the posterior distribution can be approximated by a single component of the mixture; that is, the Bernstein-von Mises theorem holds for the parameter component \( \theta_{S_0} \). The arguments here are summarized in the following two corollaries, whose proofs are straightforward and thus are omitted.

**Corollary 2.2** (Selection consistency). Suppose that (C1)–(C4), (C5*), and (C6)–(C14) are satisfied for some orthogonal projection \( H \). Then

\[
E_0 \Pi \left( \theta : S_\theta \neq S_0 \mid Y^{(n)} \right) \to 0. \tag{2.16}
\]
Corollary 2.3 (Bernstein-von Mises). Suppose that (C1)–(C4), (C5∗), and (C6)–(C14) are satisfied for some orthogonal projection $H$. Then

$$
\mathbb{E}_0 \| \Pi(\theta \in \cdot | Y^{(n)}) - N\left(\theta_{S_0} | \hat{\theta}_{S_0}, (\bar{X}_{S_0}^T (I - H) \bar{X}_{S_0} )^{-1} \right) \otimes \delta_{0}(\theta_{S_0}) \|_{TV} \to 0. \quad (2.17)
$$

2.4 Applications

In this section, we apply the main results established in the previous section to the examples considered in Section 2.1. The main objective is to obtain nearly optimal posterior contraction rates and selection consistency via shape approximation to the posterior distribution with the Bernstein-von Mises phenomenon. The results here imply that our unified theory works very well for all examples of our interest.

2.4.1 Regression with partially sparse coefficients

First we consider Example 2.1. While the prior in Section 2.2.2 is assigned to the sparse regression coefficients $\theta$, a normal prior and inverse Wishart prior are used for $\beta$ and $\Sigma$ to complete the prior specification for the nuisance parameters. Using the main results, asymptotic properties of the posterior distribution are obtained as in the following theorem.

Theorem 2.6. Let $s_* = s_0 \lor (q \log n / \log p)$. Assume that $s_0 \log p = o(n)$, $q \log n = o(n)$, $\log \|Z\|_{sp} \lesssim \log n$, $\|\beta_0\|_\infty \lesssim 1$, $1 \lesssim \rho_{min}(\Sigma_0) \lesssim \rho_{max}(\Sigma_0) \lesssim 1$, $\|\theta_0\|_\infty \lesssim \lambda^{-1} \log p$, and $\min\{\rho_{min}([X_S, Z]^T [X_S, Z]) ; s \leq Ds_* \} \gtrsim 1$ for a sufficiently large $D$. Then the following assertions hold.

(a) The posterior contraction rates for $\theta$ are given by (2.10).

(b) The posterior contraction rates for $\beta$ and $\Sigma$ are given by $\sqrt{(s_0 \log p \lor q \log n) / \rho_{min}(Z^T Z)}$ and $\sqrt{(s_0 \log p \lor q \log n) / n}$ with respect to the $\ell_2$- and Frobenius norms, respectively.

Assume further that $\|Z\|_{sp s_*} \sqrt{\log p / \rho_{min}(Z^T Z)} = o(1)$, $s_5^3 \log^3 p = o(n)$, and $\phi_1(Ds_*) \gtrsim 1$ for a sufficiently large $D$. Then the following assertions hold.

(c) For the projection $H = \tilde{Z}(\tilde{Z}^T \tilde{Z})^{-1} \tilde{Z}^T$ with $\tilde{Z}^T = (Z_1^T \Sigma_0^{-1}, \ldots, Z_n^T \Sigma_0^{-1})$, the distributional approximation in (2.14) holds.

(d) If $A_4 > 1$ and $s_4 \lesssim p^a$ for $a < A_4 - 1$, then the no-superset result in (2.15) holds.

(e) Under the beta-min condition as well as the conditions for (d), the selection consistency in (2.16) and the Bernstein-von Mises theorem in (2.17) hold.
If the error variance is assumed to be known, one might try to use the contraction results for the basic sparse regression in Castillo et al. (2015) while considering nonzero $\beta$ to also belong to the sparsity classes. This attempt gives the posterior contraction rate $\sqrt{(s_0 + q)\log p}$ relative to the prediction error (except for compatibility conditions), whereas the one obtained in (a) of Theorem 2.6 can be written as $\sqrt{s_0 + (q\log n)/\log p}\log p$. Hence, the latter improves the former if $\log n = o(\log p)$, i.e., the ultra high-dimensional cases. It is also worth noting that the rate for $\beta$ depends on the minimum eigenvalue $\rho_{\min}(Z^TZ)$. This is typically of order $n$ if $q$ is fixed, but can be slower for increasing $q$, resulting in a slower rate for $\beta$ than the case of fixed $q$.

2.4.2 Multiple response models with missing components

We apply the main results to Example 2.2. To recover posterior consistency of $\Sigma$ from the primitive results, it is required to assume that every entry of the response is jointly observed sufficiently many times. To be more specific, let $e_{ij}$ be 1 if the $j$th entry of $Y_i^{\text{aug}}$ is observed and be zero otherwise. Then the contraction rate of the $(j,k)$th element of $\Sigma$ is directly determined by the order of $n^{-1} \sum_{i=1}^{n} e_{ij}e_{ik}$. The ideal case is when this quantity is bounded away from zero, that is, the entries are jointly observed at a rate proportional to $n$. Then the recovery is possible without loss of information. If $n^{-1} \sum_{i=1}^{n} e_{ij}e_{ik}$ decays to zero, then the optimal recovery is not attainable, but consistent estimation may still be possible with slower convergence rates. With an inverse Wishart prior on $\Sigma$, the following theorem studies the posterior asymptotic properties of the given model.

**Theorem 2.7.** Let $s_* = s_0 \vee (\log n/\log p)$. Assume that $1 \lesssim \rho_{\min}(\Sigma_0) \leq \rho_{\max}(\Sigma_0) \lesssim 1$, $\|\theta_0\|\infty \lesssim \lambda^{-1}\log p$, and $\min_{j,k} n^{-1} \sum_{i=1}^{n} e_{ij}e_{ik} \gtrsim c_n^{-1}$ for some nondecreasing $c_n$ such that $c_n s_* \log p = o(n)$. Then the following assertions hold.

(a) The posterior contraction rates for $\theta$ are given by (2.10).

(b) The posterior contraction rate for $\Sigma$ is $\sqrt{c_n(s_0 \log p \vee \log n)/n}$ with respect to the Frobenius norm.

Assume further that $c_n(s_*^2 \vee \log c_n)(s_* \log p)^3 = o(n)$ and $\phi_1(Ds_*) \gtrsim 1$ for a sufficiently large $D$. Then the following assertions hold.

(c) For $H \in \mathbb{R}^{n* \times n*}$ the zero matrix, the distributional approximation in (2.14) holds.

(d) If $A_4 > 1$ and $s_* \lesssim p^a$ for $a < A_4 - 1$, then the no-superset result in (2.15) holds.

(e) Under the beta-min condition as well as the conditions for (d), the selection consistency in (2.16) and the Bernstein-von Mises theorem in (2.17) hold.
In view of the proof of Theorem 2.7, we see that the posterior contraction for each entry of \( \Sigma \) can be improved depending on the missingness mechanism. More precisely, if we have \( n^{-1} \sum_{i=1}^{n} e_{ij} e_{ik} \preceq c_{j,k,n}^{-1} \) for some sequence \( c_{j,k,n} \), the contraction rate for the \((j,k)\)th entry of \( \Sigma \) is obtained as \( \sqrt{(c_{j,k,n} s_{s} \log p)/n} \). Note that since \( c_{j,j,n} \preceq c_{j,k,n} \) for every \( j \) and \( k \), the rates for the off-diagonal entries cannot be better than those for the diagonal entries.

2.4.3 Multivariate measurement error models

We now consider Example 2.3. For convenience we write \( Y = (Y_{1}, \ldots, Y_{n})^{\top} \in \mathbb{R}^{n} \), \( W = (W_{1}^{T}, \ldots, W_{n}^{T})^{T} \in \mathbb{R}^{n \times q} \), and \( X = (X_{1}^{T}, \ldots, X_{n}^{T})^{T} \in \mathbb{R}^{n \times p} \) in what follows. In this subsection, we use the symbols \( \otimes \) and vec(\cdot) for the Kronecker product and the vectorization operator, respectively. For priors of the nuisance parameters, normal prior distributions are assigned for the scale parameters (\( \sigma \)) and \( \rho \) and an inverse gamma and inverse Wishart prior are used for the scale parameters (\( \sigma^{2} \) and \( \Sigma \)). The next theorem shows posterior asymptotic properties of the model. In particular, specific forms of their mean and variance for shape approximation are provided considering the modeling structure.

**Theorem 2.8.** Let \( s = s_{0} \vee (\log n/\log p) \). Assume that \( s_{0} \log p = o(n) \), \( |\alpha_{0}| \vee \|\beta_{0}\|_{\infty} \vee \|\mu_{0}\|\infty \leq 1 \), \( \sigma_{0}^{2} \times 1 \), \( 1 \leq \rho_{\min}(\Sigma_{0}) \leq \rho_{\max}(\Sigma_{0}) \leq 1 \), \( \|\theta_{0}\|_{\infty} \leq \lambda^{-1} \log p \), and \( \min\{\rho_{\min}(X_{0}^{*}, 1_{n})^{T}X_{0}^{*}, 1_{n}\} : s \leq Ds_{s} \} \geq 1 \) for a sufficiently large \( D \). Then the following assertions hold.

(a) The posterior contraction rates for \( \theta \) are given by (2.10).

(b) The contraction rates for \( \alpha \), \( \beta \), \( \mu \), and \( \sigma^{2} \) are \( \sqrt{(s_{0} \log p \vee \log n)/n} \) relative to the \( \ell_{2} \)-norms. The same rate is also obtained for \( \Sigma \) with respect to the Frobenius norm.

Assume further that \( s_{0}^{5} \log^{3} p = o(n) \) and \( \phi_{1}(Ds_{s}) \geq 1 \) for a sufficiently large \( D \). Then the following assertions hold.

(c) The distributional approximation in (2.14) holds with the mean vector

\[
\hat{\theta}_{S} = (X_{S}^{T} H^{*} X_{S}^{*})^{-1} X_{S}^{T} \left\{ H^{*} \left[ (Y^{*} - (\alpha_{0} + \mu_{0} \beta_{0}) 1_{n}) \right] \\
- \left( I_{n} \otimes (\beta_{0}^{T} \Sigma_{0}(\Sigma_{0} + \Psi)^{-1}) \right) \left( \text{vec}(W^{T}) - 1_{n} \otimes \mu_{0} \right) \right\}
\]

and the covariance matrix \( (\sigma_{0}^{2} + \beta_{0}^{T} \Sigma_{0}(\Sigma_{0} + \Psi)^{-1} \Psi \beta_{0})(X_{S}^{T} H^{*} X_{S}^{*})^{-1} \) for \( H^{*} = I_{n} - n^{-1} 1_{n} 1_{n}^{T} \).

(d) If \( A_{4} > 1 \) and \( s_{k} \leq p^{a} \) for \( a < A_{4} - 1 \), then the no-superset result in (2.15) holds.

(e) Under the beta-min condition as well as the conditions for (d), the selection consistency in (2.16) and the Bernstein-von Mises theorem in (2.17) hold.
One should note that the marginal law of $W_i$ is given by $W_i \sim N(\mu, \Sigma + \Psi)$. This gives a hope that the rates for $\mu$ and $\Sigma$ may actually be improved up to the parametric rate $n^{-1/2}$ (possibly up to some logarithmic factors). However, other parameters are connected to the high-dimensional coefficients $\theta$, so such a parametric rate may not be obtained for them.

2.4.4 Sparse regression with parametric correlation structure

Next, our main results are applied to Example 2.4. A correlation matrix $G_i(\alpha)$ should be chosen so that the conditions in the main theorems can be satisfied. Here we consider a compound-symmetric, a first order autoregressive, and a first order moving average correlation matrices: for $\alpha \in (b_1, b_2)$ with fixed boundaries $b_1$ and $b_2$ of the range, respectively, $\{G_i^{CS}(\alpha)\}_{j,k} = 1(j = k) + \alpha 1(j \neq k)$, $\{G_i^{AR}(\alpha)\}_{j,k} = \alpha |j-k|$, and $\{G_i^{MA}(\alpha)\}_{j,k} = 1(j = k) + \alpha 1(|j-k| = 1)$.

The range is chosen so that the corresponding correlation matrix can be positive definite, i.e., $(b_1, b_2) = (0, 1)$ for $G_i^{CS}(\alpha)$, $(b_1, b_2) = (-1, 1)$ for $G_i^{AR}(\alpha)$, and $(b_1, b_2) = (-1/2, 1/2)$ for $G_i^{MA}(\alpha)$. Again, an inverse gamma prior is assigned to $\sigma^2$. For a prior on $\alpha$, we consider a density

$$
\Pi(d\alpha) \propto \exp\left\{-\frac{1}{(\alpha - b_1)c_1(b_2 - \alpha)c_2}\right\}, \quad \alpha \in (b_1, b_2),
$$

for some $c_1, c_2 > 0$ such that $\Pi(\alpha < t) \lesssim \exp(-(t - b_1)^{-c_1})$ for $t > b_1$ close to $b_1$ and $\Pi(\alpha > t) \lesssim \exp(-(b_2 - t)^{-c_2})$ for $t < b_2$ close to $b_2$.

**Theorem 2.9.** Let $s_* = s_0 \vee (\log n / \log p)$. Assume that $s_0 \log p = o(n)$, $\overline{mn} \asymp n_*$, $\|\theta_0\|_\infty \lesssim \lambda^{-1} \log p$, $\sigma_0^2 \asymp 1$, $\alpha_0 \in [b_1 + \epsilon, b_2 - \epsilon]$ for some fixed $\epsilon > 0$. Suppose further that $\overline{m} \lesssim 1$ for the compound-symmetric correlation matrix and $\log \overline{m} \lesssim \log p$ for the autoregressive and moving average correlation matrices. Then the following assertions hold.

(a) For any correlation matrix, the posterior contraction rates for $\theta$ are given by (2.10).

(b) For the autoregressive and moving average correlation matrices, the posterior contraction rates for $\sigma^2$ and $\alpha$ are $\sqrt{(s_0 \log p \vee \log n)} / \overline{mn}$ with respect to the $\ell_2$-norms. For the compound-symmetric correlation matrix, their contraction rates are $\sqrt{(s_0 \log p \vee \log n)} / n$ relative to the $\ell_2$-norm.

Assume further that $s_*^5 \log^3 p = o(n)$ and $\phi_1(Ds_*) \gtrsim 1$ for a sufficiently large $D$. Then the following assertions hold.

(c) For $H \in \mathbb{R}^{n_* \times n_*}$ the zero matrix, the distributional approximation in (2.14) holds.

(d) If $A_4 > 1$ and $s_* \lesssim p^a$ for $a < A_4 - 1$, then the no-superset result in (2.15) holds.

(e) Under the beta-min condition as well as the conditions for (d), the selection consistency in (2.16) and the Bernstein-von Mises theorem in (2.17) hold.
As for the prior for $\alpha$, the property that the tail probabilities decay to zero exponentially fast is crucial for the optimal posterior contraction rates. It should be noted that many common probability distributions with compact supports may not be enough for this purpose (e.g., beta distributions).

The main difference between this example and those in the preceding subsections is that we consider possibly increasing $m_i$ here. Although we have the same form of contraction rates for $\theta$ as in previous examples, the implication is not the same due to a different order of $\|X\|_*$. For increasing $m_i$, it is expected to have $\|X\|_* \asymp \sqrt{n}$, which is commonly the case in regression settings. This is reduced to $\|X\|_* \asymp \sqrt{n}$ for the cases with fixed $m_i$, and hence increasing $m_i$ may help get faster convergence rates. While increasing dimensionality of $m_i$ is often a benefit for contraction properties of $\theta$, this may or may not be the case for the nuisance parameters since it depends on the dimensionality of $\eta$. In the example in this subsection, the dimension of the nuisance parameters are fixed although $m_i$ can increase, which makes their posterior contraction rates faster than those with fixed $m_i$. However, this may not be true if $\eta$ is increasing dimensional. For example, see the example in Section 2.4.6.

### 2.4.5 Mixed effects models

For the mixed effects models with sparse regression coefficients in Example 2.5, we assume that the maximum of $\|Z_i\|_{sp}$ is bounded, which is particularly mild if $\overline{m}$ is bounded. We also assume that $\sum_{i=1}^n \mathbf{1}(m_i \geq q) \asymp n$ and $\min \{\rho_{\min}(Z_i^T Z_i) : m_i \geq q\} \gtrsim 1$, that is, $m_i$ is likely to be larger than $q$ with fixed probability and $Z_i$ is a full rank. These conditions are required for (C1) to hold. We put an inverse Wishart prior on $\Psi$ as in other examples. The following theorem shows that the posterior asymptotic properties of the mixed effects models.

**Theorem 2.10.** Let $s_* = s_0 \lor (\log n/\log p)$. Assume that $s_0 \log p = o(n)$, $1 \asymp \rho_{\min}(\Psi_0) \leq \rho_{\max}(\Psi_0) \lesssim 1$, $\|\theta_0\|_{\infty} \lesssim \lambda^{-1} \log p$, $\sum_{i=1}^n \mathbf{1}(m_i \geq q) \asymp n$, $\min \{\rho_{\min}(Z_i^T Z_i) : m_i \geq q\} \gtrsim 1$, and $\max_i \|Z_i\|_{sp} \lesssim 1$. Then the following assertions hold.

(a) The posterior contraction rates for $\theta$ are given by (2.10).

(b) The posterior contraction rate for $\Psi$ is $\sqrt{(s_0 \log p \lor \log n)/n}$ with respect to the Frobenius norm.

Assume further that $s_*^5 \log^3 p = o(n)$ and $\phi_1(Ds_*) \gtrsim 1$ for a sufficiently large $D$. Then the following assertions hold.

(c) For $H \in \mathbb{R}^{n_\ast \times n_\ast}$ the zero matrix, the distributional approximation in (2.14) holds.

(d) If $A_4 > 1$ and $s_* \lesssim p^a$ for $a < A_4 - 1$, then the no-superset result in (2.15) holds.

(e) Under the beta-min condition as well as the conditions for (d), the selection consistency in (2.16) and the Bernstein-von Mises theorem in (2.17) hold.
Note that we assume that $\sigma^2$ is known, which is actually unnecessary at the modeling stage. The assumption was made to find a sequence $a_n$ satisfying (C1) with ease. This can be relaxed only with strong assumptions on $Z_i$. For example, if $q = 1$ and $Z_i$ is a one vector, then the model is equivalent to that with a compound-symmetric correlation matrix in Section 2.4.4 with some reparameterization, in which $\sigma^2$ can be treated as unknown.

### 2.4.6 Graphical structure with sparse precision matrices

For the graphical structure models in Example 2.6, we define an edge-inclusion indicator $\Upsilon = \{v_{jk} : 1 \leq j \leq k \leq m\}$ such that $v_{jk} = 1$ if $\omega_{jk} \neq 0$ and $v_{jk} = 0$ otherwise, where $\omega_{jk}$ is the $(j,k)$th element of $\Omega$. We put a prior with a density $f_1$ on $(0, \infty)$ to the nonzero off-diagonal entries and a prior with a density $f_2$ on $\mathbb{R}$ to the diagonal entries of $\Omega$, such that the support is truncated to a matrix space with restricted eigenvalues and entries. For the edge-inclusion indicator, we use a binomial prior with probability $\varpi$ when $|\Upsilon| := \sum_{j,k} v_{jk}$ is given, and assign a prior to $|\Upsilon|$ such that $\log \Pi(|\Upsilon| \leq \bar{r}) \lesssim -\bar{r} \log \bar{r}$. The prior specification is summarized as

$$
\Pi(\Omega|\Upsilon) \propto \prod_{j,k:v_{jk}=1} f_1(\omega_{jk}) \prod_{j=1}^m f_2(\omega_{jj}) \mathbb{1}_{\mathcal{M}_0^+(L)}(\Omega),
$$

$$
\Pi(\Upsilon) \propto \varpi^{|\Upsilon|}(1-\varpi)^{(\bar{m})-|\Upsilon|} \Pi(|\Upsilon| = \bar{r}), \quad \log \Pi(|\Upsilon| \leq \bar{r}) \lesssim -\bar{r} \log \bar{r},
$$

where $\mathcal{M}_0^+(L)$ is a collection of $m \times m$ positive definite matrices for a sufficiently large $L$, in which eigenvalues are between $[L^{-1}, L]$ and entries are also bounded by $L$ in absolute value.

**Theorem 2.11.** Let $s_* = s_0 \lor \{(\bar{m} + d) \log n / \log p\}$. Assume that $s_0 \log p = o(n)$, $\bar{m} \log n = o(n)$, $|\Upsilon_0| \leq d$ for some $d$ such that $d \log n = o(n)$, $\Omega_0 \in \mathcal{M}_0^+(cL)$ for some $0 < c < 1$, and $\|\theta_0\|_{\infty} \lesssim \lambda^{-1} \log p$. Then the following assertions hold.

(a) The posterior contraction rates for $\theta$ are given by (2.10).

(b) The posterior contraction rate of $\Omega$ is $\sqrt{(s_0 \log p \lor (\bar{m} + d) \log n)/n}$ with respect to the Frobenius norm. Assume further that $(s_* \lor \bar{m})^2 (s_* \log p)^3 = o(n)$ and $\phi_1(Ds_*) \gtrsim 1$ for a sufficiently large $D$. Then the following assertions hold.

(c) For $H \in \mathbb{R}^{n_x \times n_x}$ the zero matrix, the distributional approximation in (2.14) holds.

(d) If $A_4 > 1$ and $s_* \lesssim p^a$ for $a < A_4 - 1$, then the no-superset result in (2.15) holds.

(e) Under the beta-min condition as well as the conditions for (d), the selection consistency in (2.16) and the Bernstein-von Mises theorem in (2.17) hold.
Note that increasing $m$ is likely to improve the $\ell_2$-norm contraction rate for $\theta$ if $\|X\|_* \propto \sqrt{mn}$. In particular, this is clearly the case if $d \lesssim m$ and $\phi_2(Ds_\ast) \gtrsim 1$ for a sufficiently large $D$. However, as pointed out in Section 2.4.4, this is not the case for $\Omega$ as its dimension is also increasing.

If we assume that $\log n \lesssim \log m$, then the term $\sqrt{(m + d)(\log n)/n}$ in the rates arising from the sparse precision matrix $\Omega$ becomes $\sqrt{(m + d)(\log m)/n}$. The latter is comparable to the frequentist convergence rate of the graphical lasso in Rothman et al. (2008). Therefore, our rate is deemed to be optimal considering the additional complication due to the sparse regression coefficients.

### 2.4.7 Partial linear models

In this subsection, we use the main results for Example 2.7. For a bounded, convex subset $\mathcal{X} \subset \mathbb{R}$, define the $\alpha$-Hölder class $\mathcal{C}^\alpha(\mathcal{X})$ as the collection of functions $f : \mathcal{X} \to \mathbb{R}$ such that $\|f\|_{\mathcal{C}^\alpha} < \infty$, where

$$
\|f\|_{\mathcal{C}^\alpha} = \max_{0 \leq k \leq \alpha} \sup_{x \in \mathcal{X}} |f^{(k)}(x)| + \sup_{x,y \in \mathcal{X}, x \neq y} \frac{|f^{(\lfloor \alpha \rfloor)}(x) - f^{(\lfloor \alpha \rfloor)}(y)|}{|x - y|^\alpha},
$$

with the $k$th derivative $f^{(k)}$ of $f$ and $\lfloor \alpha \rfloor$ the largest integer that is strictly smaller than $\alpha$. We assume that the true function $g_0$ belongs to $\mathcal{C}^\alpha[0,1]$ for with $\alpha > 0$. For posterior contraction, any $\alpha > 0$ would suffice, while the condition $\alpha > 3/2$ is required for the Bernstein-von Mises theorem and selection consistency to hold.

We put a prior on $g$ through B-splines. The function is expressed as a linear combination of $J$-dimensional B-spline basis terms $B_J$ of order $q \geq \alpha$, i.e., $g_\beta(z) = \beta^T B_J(z)$, and a normal prior distribution is assigned to $\beta$. Let $\|f\|_{\infty} = \sup_{z \in [0,1]} |f(z)|$ and $\|f\|_{2,n} = (n^{-1} \sum_{i=1}^n |f(z_i)|^2)^{1/2}$ denote the sup-norm and empirical $L_2$-norm, respectively.

The asymptotic properties of the posterior distribution are provided in the following theorem. The number of basis terms should be chosen to balance the conditions in Theorem 2.3. However, the number of basis terms that induces the optimal rates does not suffice for the conditions for the distributional approximation, and we need more basis terms for selection consistency, while sacrificing the optimal contraction rates.

**Theorem 2.12.** Let $s_\ast = s_0 \lor \{(\log n)^{2\alpha/(2\alpha + 1)}n^{1/(2\alpha + 1)}/\log p\}$ and let $W = (B_J(z_1), \ldots, B_J(z_n))^T \in \mathbb{R}^{n \times J}$ for $J = J_n \asymp (n/\log n)^{1/(2\alpha + 1)}$. Assume that $s_0 \log p = o(n)$, $\sigma_0^2 \asymp 1$, $\|\theta_0\|_{\infty} \lesssim \lambda^{-1} \log p$, $\min \{\rho_{\min}(\{X_S, W\}^T\{X_S, W\}) : s \leq Ds_\ast\} \gtrsim 1$ for a sufficiently large $D$, and $g_0 \in \mathcal{C}^\alpha[0,1]$ with $\alpha > 0$. Then the following assertions hold.

(a) The posterior contraction rates for $\theta$ are given by (2.10).
(b) The contraction rates for \( g \) and \( \sigma^2 \) are \( \sqrt{(s_0 \log p)/n} \vee (\log n/n)^{\alpha/(2\alpha+1)} \) with respect to the \( \| \cdot \|_{2,n^*} \) and \( \ell_2 \)-norms, respectively.

Moreover, instead of \( s_* \) defined above, let \( s_* = s_0 \vee \{ (\log n)[M_n n s_0^2 (\log p)^{1-2\alpha}]^{1/2} \} \vee \{ (\log n)^{2\alpha} \times M_n n (\log p)^{1-2\alpha} \}^{1/2(\alpha-1)} \) for an arbitrary \( M_n \to \infty \). Also, let \( W_* = (B_{J_*}(z_1), \ldots, B_{J_*}(z_n))^T \in \mathbb{R}^{n \times J_*} \) for \( J_* = J_n \propto (M n s_0^2 \log p)^{1/2} \vee (M_n n (\log^2 n/\log p))^{1/2(\alpha-1)} \). Assume that \( s_*^5 \log^3 p = o(n) \), and \( \min \{ \rho_{\min}([X_S W_*] [X_S W_*]) : s \leq D s_* \} \geq 1 \) and \( \phi_1(D s_*) \geq 1 \) for a sufficiently large \( D \). Then the following assertions hold.

(c) The distributional approximation in (2.14) holds for \( H = W_*(W_*^T W_*)^{-1} W_*^T \).

(d) If \( A_4 > 1 \) and \( s_* \lesssim p^a \) for \( a < A_4 - 1 \), then the no-superset result in (2.15) holds.

(e) Under the beta-min condition as well as the conditions for (d), the selection consistency in (2.16) and the Bernstein-von Mises theorem in (2.17) hold.

Here we elaborate more on the different choices of the number of basis terms for the two purposes. Unlike the previous examples in Section 2.4.1–2.4.6, the true nuisance mean part \( g_0 \) of this example does not belong to the support of the prior, which prevents the condition (C8) being satisfied trivially. It can be checked that \( H = W(W^T W)^{-1} W^T \) with \( W \) chosen in Theorem 2.12 does not make (C8) hold, which means that for the Bernstein-von Mises theorem and selection consistency, we need a better approximation for \( \| (I - H) \xi_{s_0} \|_2 \) with more basis terms to dominate the factor \( s_*^2 \log p \), sacrificing the optimal contraction rates obtained in (a)–(b) of Theorem 2.12. This result is opposed to the findings in partial linear models with fixed dimensional regression coefficients (Bickel and Kleijn, 2012). We do not know whether this is a fundamental limitation for our high-dimensional regression models or a technical artifact.

Note also that \( J_* \) is dependent on \( s_0 \) while \( J \) is free of \( s_0 \). This means that the procedure is not adaptive to the unknown sparsity level anymore. The adaptivity can be attained by considering a suitable bound for \( s_0 \); for example, as we assume that \( s_0 \log p = o(n) \), one can replace \( s_0 \) by \( n/\log p \) in the construction of \( J_* \). The required modifications for the conditions are straightforward. We do not provide the details here.

Using \( s_* \) used in (c)–(e) of Theorem 2.12, it can be seen that to satisfy \( s_*^5 \log^3 p = o(n) \), we need \( \alpha > 3/2 \). Note also that although this \( s_* \) yields suboptimal posterior contraction rates, it approaches \( s_0 \vee (\log n/\log p) \) as \( \alpha \) increases while \( n \) and \( p \) are held fixed. Hence, the loss of optimality is tolerable for sufficiently smooth functions.

### 2.4.8 Nonparametric heteroskedastic regression models

Lastly, we consider Example 2.8. Let the true function \( v_0 \) belong to \( \mathcal{C}^\alpha[0,1] \) with assumption that \( v_0 \) is strictly positive. We shall see that \( a > 1/2 \) suffices for posterior contraction, but for
the Bernstein-von Mises theorem and selection consistency, we need $\alpha > 2$. As in the partial linear models in the previous subsection, we put a prior on $v$ through a B-spline basis of order $q \geq \alpha$, i.e., $v(z) = \beta^T B_J(z)$. We also put an independent inverse Gaussian prior on each entry of $\beta$.

**Theorem 2.13.** Let $s_\star = s_0 \lor \{(\log n)^{2\alpha/(2\alpha+1)} n^{1/(2\alpha+1)}/\log p\}$. Assume that $s_0 \log p = o(n)$, $\|\theta_0\|_{\infty} \lesssim \lambda^{-1} \log p$, and that the true function $v_0$ is strictly positive on $[0,1]$ and belongs to $C^{\alpha}[0,1]$ with $\alpha > 1/2$. We choose $J = J_n \asymp (n/\log n)^{1/(2\alpha+1)}$. Then the following assertions hold.

(a) The posterior contraction rates for $\theta$ are given by (2.10).

(b) The posterior contraction rate for $v$ is $\sqrt{(s_0 \log p)/n} \lor (\log n/n)^{\alpha/(2\alpha+1)}$ with respect to the $\|\cdot\|_{2,n}$-norm.

Assume further that $J(s_\star^2 \lor J)(s_\star \log p)^3 = o(n)$ and $\phi_1(Ds_\star) \gtrsim 1$ for a sufficiently large $D$. Then the following assertions hold.

(c) The distributional approximation in (2.14) holds with $H$ the $n \times n$ zero matrix.

(d) If $A_4 > 1$ and $s_\star \lesssim p^a$ for $a < A_4 - 1$, then the no-superset result in (2.15) holds.

(e) Under the beta-min condition as well as the conditions for (d), the selection consistency in (2.16) and the Bernstein-von Mises theorem in (2.17) hold.

We put an inverse Gaussian prior on each entry of $\beta$ due to the property that its tail probability decays to zero exponentially fast. The exponentially decaying tail probability is essential to obtain the optimal contraction rate. Note that standard choices such as gamma and inverse gamma distributions do not satisfy this property.

By investigating the proof, it can be seen that we need $\alpha > 1/2$ to satisfy the condition (C1) for posterior contraction. It can also be checked that the condition $J(s_\star^2 \lor J)(s_\star \log p)^3 = o(n)$ for (c)–(e) implies $\alpha > 2$. Hence, we need more smoothness than that in the partial linear models for the theorem to hold. However, unlike in Theorem 2.12, the same number of basis terms are used in (a)–(b) and (c)–(e) of Theorem 2.13. This allows us to establish the optimal posterior contraction and selection consistency simultaneously, using the techniques developed in this chapter.

### 2.5 Proofs

#### 2.5.1 Proofs for the main results

In this section, we provide proofs for the main theorems given in Section 2.3. We first describe the additional notations used for the proofs. For a matrix $X$, we write $\rho_1(X) \geq \rho_2(X) \geq \cdots$ for
the eigenvalues of $X$ in decreasing order. The notation $\Lambda_n(\theta, \eta) = \prod_{i=1}^{n}(p_{\theta,\eta,i}/p_{0,i})(Y_i)$ stands for the likelihood ratio of $p_{\theta,\eta}$ and $p_0$. Let $\mathbb{E}_{\theta,\eta}$ denote the expectation operator with the density $p_{\theta,\eta}$ and let $\mathbb{P}_0$ denote the probability operator with the true density. For two densities $f$ and $g$, let $K(f, g) = \int f \log(f/g)$ and $V(f, g) = \int f|\log(f/g) - K(f, g)|^2$ stand for the Kullback-Leibler divergence and variation, respectively. Using some constants $\rho, \epsilon, g, p, \theta, \eta$, let $\Delta = \sum_{i=1}^{m} (1 - \rho^*_{i,k})^2 + \|\Delta_{\eta,i}^{-1}(X_i(\theta - \theta_0) + \xi_{\eta,i} - \xi_{0,i})\|_2$, where $\rho^*_{i,k}, k = 1, \ldots, m_i$, are the eigenvalues of $\Delta_{\eta,i}^{-1}$. For $\mathbb{I}_{\eta,\delta} = \{1 \leq i \leq n : \sum_{k=1}^{m_i} (1 - \rho^*_{i,k})^2 \geq \delta\}$ with small $\delta > 0$ and $|\mathbb{I}_{\eta,\delta}|$ the cardinality of $\mathbb{I}_{\eta,\delta}$, we see that on $\mathcal{B}_n$,

$$2a_n\epsilon_n^2 \geq \frac{a_n}{n} \sum_{i=1}^{n} (1 - \rho^*_{i,k})^2 \geq \frac{a_n\delta|\mathbb{I}_{\eta,\delta}|}{n} + \frac{a_n}{n} \sum_{i \notin \mathbb{I}_{\eta,\delta}} \sum_{k=1}^{m_i} (1 - \rho^*_{i,k})^2.$$

and

$$\frac{a_n}{n} \sum_{i \notin \mathbb{I}_{\eta,\delta}} \sum_{k=1}^{m_i} (1 - \rho^*_{i,k})^2 \geq \frac{a_n}{n} \sum_{i \notin \mathbb{I}_{\eta,\delta}} \sum_{k=1}^{m_i} (1 - \rho^*_{i,k})^2 \geq \frac{a_n}{\rho_0^*} \sum_{i \notin \mathbb{I}_{\eta,\delta}} \|\Delta_{\eta,i} - \Delta_{\eta,0,i}\|_F^2.$$
where the first inequality follows by the relation \(|1-x| \propto |1-x^{-1}|\) as \(x \to 1\) and the second inequality holds by (i) of Lemma 2.6 in Appendix. Since \(a_n |\mathcal{I}_{n,\delta}|/n \leq a_n \epsilon^2\) by (2.20), it follows that for some constants \(C_1, C_2 > 0\),

\[
\frac{a_n}{n} \sum_{i \notin \mathcal{I}_{n,\delta}} \|\Delta_{\eta,i} - \Delta_{\eta_0,i}\|^2_F \geq \frac{a_n d_B^2(n, \eta, \eta_0)}{n} \max_{1 \leq i \leq n} \|\Delta_{\eta,i} - \Delta_{\eta_0,i}\|^2_F \\
\geq (C_1 - C_2 a_n \epsilon^2) \max_{1 \leq i \leq n} \|\Delta_{\eta,i} - \Delta_{\eta_0,i}\|^2_F - \epsilon_n.
\]

Hence we have \(a_n \epsilon^2_n + \epsilon_n \geq \max_i \|\Delta_{\eta,i} - \Delta_{\eta_0,i}\|^2_F\) on \(\mathcal{B}_n\), which implies that \(\max_{i,k} |1 - \rho_{i,k}^*|\) is small for all sufficiently large \(n\), by (i) of Lemma 2.6 and the inequality \(|1-x| \propto |1-x^{-1}|\) as \(x \to 1\). Hence, \(\log \rho_{i,k}^*\) can be expanded in the powers of \((1 - \rho_{i,k}^*)\) to get \(- \log \rho_{i,k}^* - (1 - \rho_{i,k}^*) \sim (1 - \rho_{i,k}^*)^2/2\) for every \(i\) and \(k\). Furthermore, since \(\max_{i,k} |1 - \rho_{i,k}^*|\) is sufficiently small, we obtain that \(\sum_{k=1}^{m_i} (1 - \rho_{i,k}^*)^2 \leq \sum_{k=1}^{m_i} (1 - 1/\rho_{i,k}^*)^2 \leq \|\Delta_{\eta,i} - \Delta_{\eta_0,i}\|^2_F\) by (i) of Lemma 2.6, and that \(\|\Delta_{\eta,i}^1\|^2_{sp} \leq \|\Delta_{\eta_0,i}^1\|^2_{sp} \leq 1\) by the restriction on the eigenvalues of \(\Delta_{\eta,i}\). Combining these, it follows that on \(\mathcal{B}_n\), both \(n^{-1} \sum_{i=1}^{n} K(p_{0,i}, p_{\theta,\eta,i})\) and \(n^{-1} \sum_{i=1}^{n} V(p_{0,i}, p_{\theta,\eta,i})\) are bounded above by a constant multiple of \(n^{-1}\|X(\theta - \theta_0)\|_2^2 + d_n^2(\eta, \eta_0)\). Hence, there exists a constant \(C_3\) such that

\[
P(\mathcal{B}_n) \geq \Pi \left\{(\theta, \eta) \in \Theta \times \mathcal{H} : \frac{1}{n} \|X\|_2^2 \|\theta - \theta_0\|^2 + d_n^2(\eta, \eta_0) \leq C_3 \epsilon^2 \right\} \\
\geq \Pi \left\{\theta \in \Theta : \epsilon_n \geq \frac{1}{n} \|X\|_2^2 \|\theta - \theta_0\|^2 \leq \frac{C_3}{\epsilon_n} \right\} \Pi \left\{\eta \in \mathcal{H} : d_n^2(\eta, \eta_0) \leq \frac{C_3}{\epsilon_n} \right\},
\]

by the inequality \(\|X\theta\|_2 \leq \sum_{j=1}^{p} \theta_j \|X_j\|_2 \leq \|X\|_1 \|\theta\|_1\). The logarithm of the second term on the rightmost side is bounded below by a constant multiple of \(-n\epsilon^2\) by (C2). To find the lower bound for the first term, we shall first work with the case \(s_0 \geq 1\), and then show that the same lower bound is obtained even when \(s_0 = 0\).

Now, assume that \(s_0 \geq 1\) and let \(\Theta_{0,n} = \{\theta_{S_0} \in \mathbb{R}^{s_0} : n^{-1/2} \|X\|_1 \|\theta_{S_0} - \theta_{0,S_0}\|_1 \leq \epsilon\}\) for \(\epsilon > 0\) to be chosen later, then

\[
P\{\theta \in \Theta : n^{-1/2} \|X\|_1 \|\theta - \theta_0\|_1 \leq \epsilon\} \geq \frac{\pi_p(s_0)}{p_{s_0}} \int_{\theta_{0,n}} g_{S_0}(\theta_{S_0}) d\theta_{S_0} \\
\geq \frac{\pi_p(s_0)}{p_{s_0}} e^{-\lambda \|\theta_0\|_1} \int_{\theta_{0,n}} g_{S_0}(\theta_{S_0} - \theta_{0,S_0}) d\theta_{S_0},
\]

by the inequality \(g_{S_0}(\theta_{S_0}) \geq e^{-\lambda \|\theta_0\|_1} g_{S_0}(\theta_{S_0} - \theta_{0,S_0})\). Using the relation (6.2) of Castillo et al.
by the inequality (2.15) and the assumption on the prior in (2.4), the integral on the rightmost side satisfies
\[
\int_{\Theta_{0,n}} g_{S_0}(\theta S_0 - \theta_0 S_0) d\theta S_0 \geq e^{-\lambda \epsilon \sqrt{n}/\|X\|_\star} \left( \frac{\lambda \epsilon \sqrt{n}/\|X\|_\star}{s_0!} \right)^{s_0} \geq e^{-\lambda \epsilon \epsilon \sqrt{n}/L_1 p^{L_2}}/s_0!,
\]
for \( s_0 > 0 \), and thus the rightmost side of (2.22) is bounded below by
\[
\pi_p(s_0)(\epsilon \sqrt{n})^{s_0} \exp \{ -\lambda \|\theta_0\|_1 - (L_1 + 1) s_0 \log p - s_0 \log L_1 \},
\]
by the inequality \( (p/s_0) s_0! \leq p^{s_0} \). Choosing \( \epsilon = \sqrt{C_3 \epsilon_n} \) for \( C_4 = C_3/2 \), the first term on the rightmost side of (2.21) satisfies
\[
\Pi \left\{ \theta \in \Theta : \frac{1}{n} \|X\|_\star^2 \|\theta - \theta_0\|^2 \leq C_4 \epsilon_n^2 \right\} \geq \pi_p(s_0)(n \epsilon_n^2)^{s_0/2} \exp \{ -\lambda \|\theta_0\|_1 - L_3 \epsilon_n \sqrt{C_4} - (L_1 + 1) s_0 \log p + s_0 \log (\sqrt{C_4}/L_1) \}.
\]
Note that \( n \epsilon_n^2 > 1 \) and \( s_0 + \epsilon_n + s_0 \log p \leq s_0 \log p \) if \( s_0 > 0 \), and thus the last display implies that there exists a constant \( C_5 > 0 \) such that
\[
\Pi(B_n) \geq \pi_p(s_0) \exp \{ -C_5 (\lambda \|\theta_0\|_1 + s_0 \log p + n \epsilon_n^2) \}.
\]
If \( s_0 = 0 \), the first term of (2.21) is clearly bounded below by \( \pi_p(0) \), so that the same lower bound for \( \Pi(B_n) \) in the last display is also obtained since we have \( \lambda \|\theta_0\|_1 + s_0 \log p = 0 \). Finally, the lemma follows from (2.19).

**Proof of Theorem 2.1.** For the set \( \mathcal{B} = \{ (\theta, \eta) : s_\theta > \bar{s} \} \) with any integer \( \bar{s} \geq s_0 \), we have
\[
\Pi(\mathcal{B}) = \sum_{s= \bar{s} + 1}^p \pi_p(s) \leq \pi_p(s_0) \sum_{s= \bar{s} + 1}^p \left( \frac{A_2}{p A_4} \right)^{s-s_0} \leq \pi_p(s_0) \left( \frac{A_2}{p A_4} \right)^{\bar{s}+1-s_0} \sum_{j=0}^{\infty} \left( \frac{A_2}{p A_4} \right)^j.
\]
Let \( \mathcal{E}_n \) be the event in (2.18), then by Lemma 2.1,
\[
\mathbb{E}_0 \Pi(\mathcal{B} | Y^{(n)}) 1_{\mathcal{E}_n} = \mathbb{E}_0 \left[ \int \int \Lambda_n(\theta, \eta) d\Pi(\theta, \eta) \right] 1_{\mathcal{E}_n} \leq \pi_p(s_0)^{-1} \exp \{ C_1 (s_0 \log p + n \epsilon_n^2) \} \Pi(\mathcal{B}) \leq \exp \left\{ (\bar{s} + 1 - s_0)(\log A_2 - A_4 \log p) + C_1 (s_0 \log p + n \epsilon_n^2) \right\},
\]
for some constant \( C_1 \) and sufficiently large \( p \). For a sufficiently large constant \( C_2 \), choose the
largest integer that is smaller than \( C_2(s_0 \lor (n^2/\log p)) \) for \( s \). Replacing \( s + 1 \) by \( C_2(s_0 \lor (n^2/\log p)) \) in the last display, it is easy to see that the rightmost side goes to zero. The proof is complete since \( P_0(\mathcal{E}_n^c) \rightarrow 0 \) by Lemma 2.1.

The following lemma shows that a small piece of the alternative centered at any \((\theta_1, \eta_1) \in \Theta \times \mathcal{H}\) are locally testable with exponentially small errors, provided that the center is sufficiently separated from the truth with respect to the average Rényi divergence. Theorem 2.2 for posterior contraction relative to the average Rényi divergence will then be proved by showing that the number of those pieces is controlled by the target rate. We write \( p_1 \) for the density with \((\theta_1, \eta_1)\), and \( E_1 \) and \( P_1 \) for the expectation and probability with \( p_1 \), respectively.

**Lemma 2.2.** For some sequence \( \gamma'_n > 0 \) and given \((\theta_1, \eta_1) \in \Theta \times \mathcal{H}\) such that \( R_n(p_0, p_1) \geq \delta_n^2 \) with some \( \delta_n = o(\sqrt{m}) \), define

\[
F_{1,n} = \left\{(\theta, \eta) \in \Theta \times \mathcal{H} : \frac{1}{n} \sum_{i=1}^{n} \|X_i(\theta - \theta_1) + \xi_{\eta,i} - \xi_{\eta_1,i}\|_2^2 \leq \frac{\delta_n^2}{16 \gamma'_n}, \quad d_{B,n}(\eta, \eta_1) \leq \frac{\delta_n^2}{2m \gamma'_n \sqrt{a_n}}, \quad \max_{1 \leq i \leq n} \|\Delta_{\eta,i}^{-1}\|_{sp} \leq \gamma'_n \right\},
\]

(2.23)

where \( a_n \) is a sequence satisfying (C1). Then there exists a test \( \bar{\varphi}_n \) such that

\[
E_0 \bar{\varphi}_n \leq e^{-n\delta_n^2}, \quad \sup_{(\theta, \eta) \in F_{1,n}} E_{\theta,\eta}(1 - \bar{\varphi}_n) \leq e^{-n\delta_n^2/16}.
\]

**Proof.** For given \((\theta_1, \eta_1) \in \Theta \times \mathcal{H}\) such that \( R_n(p_0, p_1) \geq \delta_n^2 \), consider the most powerful test \( \bar{\varphi}_n = 1\{\Lambda_n(\theta_1, \eta_1) \geq 1\} \) given by the Neyman-Pearson lemma. It is then easy to see that

\[
E_0 \bar{\varphi}_n = P_0 \left( \sqrt{\Lambda_n(\theta_1, \eta_1)} \geq 1 \right) \leq \int \sqrt{p_0 p_1} \leq e^{-n\delta_n^2}, \quad E_1(1 - \bar{\varphi}_n) = P_1 \left( \sqrt{\Lambda_n(\theta_1, \eta_1)} \leq 1 \right) \leq \int \sqrt{p_0 p_1} \leq e^{-n\delta_n^2}.
\]

(2.24)

The first inequality of the lemma is a direct consequence of the first line of the preceding display. For the second inequality of the lemma, note that by the Cauchy-Schwarz inequality, we have

\[
\{E_{\theta,\eta}(1 - \bar{\varphi}_n)\}^2 \leq E_1(1 - \bar{\varphi}_n) E_1((p_{\theta,\eta}/p_1)(Y^{(n)}))^2.
\]

Thus, by the second line of (2.24), it suffices to show \( E_1((p_{\theta,\eta}/p_1)(Y^{(n)}))^2 \leq e^{7n\delta_n^2/8} \) for every
$(\theta, \eta) \in F_{1,n}$. Note that for $\Delta_{\eta,i}^* = \Delta_{\eta,i}^{-1/2} \Delta_{\eta,i} \Delta_{\eta,i}^{-1/2}$,

$$\max_{1 \leq i \leq n} \|\Delta_{\eta,i}^* - I\|_{sp} \leq \max_{1 \leq i \leq n} \|\Delta_{\eta,i}^{-1}\|_{sp} \|\Delta_{\eta,i} - \Delta_{\eta,i}^{-1}\|_{sp} \leq \max_{1 \leq i \leq n} \|\Delta_{\eta,i}^{-1}\|_{sp} \sqrt{\omega_n d_{B,n}(\eta, \eta_1)} \leq \frac{\delta_n^2}{2m},$$

on the set $F_{1,n}$. Since the leftmost side of the display is further bounded below by $\max_i |\rho_k(\Delta_{\eta,i}^*) - 1|$ for every $k \leq m_i$, we have that

$$1 - \frac{\delta_n^2}{2m} \leq \min_{1 \leq i \leq n} \rho_{min}(\Delta_{\eta,i}) \leq \max_{1 \leq i \leq n} \rho_{max}(\Delta_{\eta,i}) \leq 1 + \frac{\delta_n^2}{2m}. \quad (2.25)$$

Since $\delta_n^2/m \to 0$, the display implies that $2\Delta_{\eta,i}^* - I$ is nonsingular for every $i \leq n$, and hence it can be shown that on $F_{1,n},$

$$E_1(p_{\theta, \eta}/p_1)^2 = \prod_{i=1}^n \left\{ \det(\Delta_{\eta,i}^*)^{1/2} \det(2I - \Delta_{\eta,i}^{-1})^{1/2} \right\} \times \exp \left\{ \sum_{i=1}^n \|\Delta_{\eta,i} - I\|^{-1/2} \Delta_{\eta,i}^{-1/2} (X_i(\theta - \theta_1) + \xi_{\eta,i} - \xi_{\eta,i})_2 \right\}. \quad (2.26)$$

To bound this, note that $\det(\Delta_{\eta,i}^*)^{1/2} \det(2I - \Delta_{\eta,i}^{-1})^{1/2}$ is equal to

$$\prod_{k=1}^{m_i} \left\{ \frac{\rho_k(\Delta_{\eta,i})}{2 - \rho_k^{-1}(\Delta_{\eta,i})} \right\}^{1/2} \leq \left( \frac{1 - \delta_n^4/4m^2}{1 - \delta_n^2/m} \right)^{m_i/2} \leq \left( 1 + \frac{3\delta_n^2}{2m} \right)^{m_i/2} \leq e^{3\delta_n^2/4}, \quad (2.27)$$

where the first inequality holds by (2.25), the second inequality holds by the inequality $(1 - x^2)/(1 - 2x) \leq 1 + 3x$ for small $x > 0$, and the last inequality holds by the inequality $x + 1 \leq e^x$.

Now, for every $(\theta, \eta) \in F_{1,n}$, observe that the exponent in (2.26) is bounded above by

$$\max_{1 \leq i \leq n} \|\Delta_{\eta,i} - I\|_{sp} \max_{1 \leq i \leq n} \|\Delta_{\eta,i}^{-1}\|_{sp} \sum_{i=1}^n \|X_i(\theta - \theta_1) + \xi_{\eta,i} - \xi_{\eta,i}\|_2 \leq \frac{\delta_n^2}{8},$$

since $\max_i\|\Delta_{\eta,i}^* - I\|_{sp} \leq 2$ for large $n$. Combined with (2.26) and (2.27), the display completes the proof.

**Proof of Theorem 2.2.** Let $\Theta_n = \{\theta \in \Theta : s_{\theta} \leq K_1 s_n\}$ and $R_{n}(\theta, \eta) = R_{n}(p_{\theta, \eta}, \theta_0)$. Then for every $\epsilon > 0$,

$$E_0 \Pi \left( (\theta, \eta) \in \Theta \times \mathcal{H} : \sqrt{R_{n}(\theta, \eta)} > \epsilon \mid Y^{(n)} \right) \leq E_0 \Pi \left( (\theta, \eta) \in \Theta_n \times \mathcal{H} : \sqrt{R_{n}(\theta, \eta)} > \epsilon \mid Y^{(n)} \right) + E_0 \Pi \left( \Theta_n^c \mid Y^{(n)} \right), \quad (2.28)$$

34
where the second term on the right hand side goes to zero by Theorem 2.1. Hence, it suffices to show that the first term goes to zero for \(\epsilon > 0\) chosen to be the threshold in the theorem. Now, let \(\Theta_n^* = \{\theta \in \Theta : s_\theta \leq K_1 s_* \text{ and } \|\theta - \theta_0\|_\infty \leq pL_{2+2}/\|X\|_*\}\) and define \(F_{1,n}\) as in (2.23) with \(\gamma_n' = \gamma_n\) and \(\delta_n = \epsilon_n\). Then Lemma 2.2 implies that small pieces of the alternative densities can be tested with exponentially small errors as long as the center is \(\epsilon_n\)-separated from the true parameter values relative to the average Rényi divergence. To complete the proof, we shall show that the minimal number \(N_n^*\) of those small pieces that are needed to cover \(\Theta_n^* \times \mathcal{H}_n\) is controlled appropriately in terms of \(\epsilon_n\), and that the prior mass of \(\Theta_n \setminus \Theta_n^*\) and \(\mathcal{H} \setminus \mathcal{H}_n\) decreases fast enough to balance the denominator of the posterior distribution. (For more discussion on a construction of a test using metric entropies, see Section D.2 and Section D.3 of Ghosal and van der Vaart (2017).)

Note that for any small \(\delta > 0\),

\[
N(\delta, \Theta_n^*, \|\cdot\|_\infty) \leq \left(\frac{p}{K_1 s_*}\right) \left(\frac{3pL_{2+2}}{\delta \|X\|_*}\right)^{K_1 s_*} \leq \left(\frac{3pL_{2+3}}{\delta \|X\|_*}\right)^{K_1 s_*}
\]

and thus we obtain

\[
\log N \left(\frac{1}{6m\gamma_n n p\|X\|_*}, \Theta_n^*, \|\cdot\|_\infty\right) \leq s_*(\log m + \log \gamma_n + \log p) \leq n\epsilon_n^2.
\]

Using the last display and the entropy condition (2.7), the right hand side of (2.29) is bounded above by a constant multiple of \(n\epsilon_n^2\). Hence, by Lemma D.3 of Ghosal and van der Vaart (2017), for every \(\epsilon > \epsilon_n\), there exists a test \(\varphi_n\) such that for some \(C_1 > 0\), \(E_0\varphi_n \leq 2 \exp(C_1 n\epsilon_n^2 - n\epsilon^2)\) and \(E_{\theta,\eta}(1 - \varphi_n) \leq \exp(-n\epsilon^2/16)\) for every \((\theta, \eta) \in \Theta_n^* \times \mathcal{H}_n\) such that \(\sqrt{R_n^*(\theta, \eta)} > \epsilon\). Note
that under the condition (2.3) on the prior distribution, we have $-\log \pi_p(s_0) \lesssim s_0 \log p$ since $\pi_p(0)$ is bounded away from zero. Hence, for $E_n$ the event in (2.18) and some constant $C_2 > 0$, the first term on the right hand side of (2.28) is bounded by

$$
E_0 \Pi \left( (\theta, \eta) \in \Theta_n \times H : \sqrt{R_n(\theta, \eta)} > \epsilon \right) \Pi_0(1 - \varphi_n) + E_0(\varphi_n + 1_{E_n^c})
$$

$$
\leq \sup_{(\theta, \eta) \in \Theta_n \times H : \sqrt{R_n(\theta, \eta)} > \epsilon} E_{\theta, \eta}(1 - \varphi_n) + \Pi(\Theta_n \setminus \Theta_n^*) + \Pi(H \setminus H_n) \right) e^{C_2 s_n \log p}
$$

$$
+ E_0 \varphi_n + P_0 E_n^c,
$$

where the term $P_0 E_n^c$ converges to zero by Lemma 2.1. Choosing $\epsilon = C_3 \nu_n$ for a sufficiently large $C_3$, we have

$$
E_0 \varphi_n \to 0, \quad \sup_{(\theta, \eta) \in \Theta_n \times H : \sqrt{R_n(\theta, \eta)} > \epsilon} E_{\theta, \eta}(1 - \varphi_n) e^{C_2 s_n \log p} \to 0.
$$

Furthermore, $\Pi(H \setminus H_n) e^{C_2 s_n \log p}$ goes to zero by the condition (2.8). Now, to show that $\Pi(\Theta_n \setminus \Theta_n^*)$ goes to zero exponentially fast, note that $\Pi(\Theta_n \setminus \Theta_n^*)$ is equal to

$$
\Pi \left\{ \theta \in \Theta : s_0 \leq K_1 s_n, \| \theta - \theta_0 \|_\infty > p^{L_2 + 2} / \| X \|_* \right\}
$$

$$
= \sum_{S : s \leq K_1 s_n} \pi_p(s) \int_{\{\theta_S : \| \theta_S - \theta_{0,S} \|_\infty \| \theta_{0,S} \|_\infty > p^{L_2 + 2} / \| X \|_* \}} g_S(\theta_S) d\theta_S
$$

$$
\leq e^{\lambda \| \theta_0 \|_1} \sum_{S : s \leq K_1 s_n} \left( \frac{A_2 p^{-A_4}}{\| \theta_0 \|_1} \right) \int_{\{\theta_S : \| \theta_S - \theta_{0,S} \|_\infty > p^{L_2 + 2} / \| X \|_* \}} g_S(\theta_S - \theta_{0,S}) d\theta_S,
$$

by the inequalities $g_S(\theta_S) \leq e^{\lambda \| \theta_0 \|_1} g_S(\theta_S - \theta_{0,S})$ and $\pi_p(s) \leq (A_2 p^{-A_4})^s \pi_p(0)$ for every $S$. Since the tail probability of the Laplace distribution is given by $\int_{|x| > t} 2^{-1} \lambda e^{-\lambda |x|} dx = \exp(-\lambda t)$ for every $t > 0$, the rightmost side of the last display is bounded above by a constant multiple of

$$
e^{-\lambda \| \theta_0 \|_1} \sum_{S = 1}^{K_1 s_n} s e^{-\lambda p^{L_2 + 2} / \| X \|_* + \lambda \| \theta_0 \|_\infty} \left( \frac{A_2}{p^{A_4}} \right)^s \lesssim s e^{2 \lambda \| \theta_0 \|_1 - \lambda p^{L_2 + 2} / \| X \|_*}.
$$

Since $\lambda \| \theta_0 \|_1 \lesssim s_* \log p$ and $\lambda p^{L_2 + 2} / \| X \|_* \gtrsim p^2$ by (2.4), the right hand side of the last display is bounded above by $e^{-C_4 p^2}$ for some $C_4 > 0$, and thus $\Pi(\Theta_n \setminus \Theta_n^*) e^{C_2 s_n \log p}$ goes to zero since $s_* \log p = o(p^2)$. Finally, we conclude that the left hand side of (2.28) with $\epsilon = C_3 \nu_n$ goes to zero.

\[\square\]

**Proof of Theorem 2.3.** By Theorem 2.2, we obtain the contraction rate of the posterior distribution with respect to the average Rényi divergence $R_n(p_{\theta, \eta}, p_0)$ between $p_{\theta, \eta}$ and $p_0$ given
by

\[ R_n(p_{\theta, \eta}, p_0) = - \frac{1}{n} \sum_{i=1}^{n} \log \left\{ \frac{(\det \Delta_{\eta,i})^{1/4}(\det \Delta_{\eta_0,i})^{1/4}}{\det((\Delta_{\eta,i} + \Delta_{\eta_0,i})/2)^{1/2}} \right\} \]

\[ + \frac{1}{4n} \sum_{i=1}^{n} \| (\Delta_{\eta,i} + \Delta_{\eta_0,i})^{-1}(X_i(\theta - \theta_0) + \xi_{\eta,i} - \xi_{\eta_0,i}) \|_2^2. \]

Define

\[ g^2(\Delta_{\eta,i}, \Delta_{\eta_0,i}) = 1 - \frac{(\det \Delta_{\eta,i})^{1/4}(\det \Delta_{\eta_0,i})^{1/4}}{\det((\Delta_{\eta,i} + \Delta_{\eta_0,i})/2)^{1/2}}. \]

(2.30)

Then Theorem 2.2 implies that by the last display,

\[ \epsilon_n^2 \geq - \frac{1}{n} \sum_{i=1}^{n} \log(1 - g^2(\Delta_{\eta,i}, \Delta_{\eta_0,i})) \geq \frac{1}{n} \sum_{i=1}^{n} g^2(\Delta_{\eta,i}, \Delta_{\eta_0,i}), \]

(2.31)

where the second inequality holds by the inequality \( \log x \leq x - 1 \). Note that by combining (i) and (ii) of Lemma 2.6 in Appendix, we obtain \( g^2(\Delta_{\eta,i}, \Delta_{\eta_0,i}) \gtrsim \| \Delta_{\eta,i} - \Delta_{\eta_0,i} \|_F^2 \) if the left hand side is small. Thus, using the same approach in the proof of Lemma 2.1, (2.31) is further bounded below by

\[ C_1 d_{B,n}(\eta, \eta_0) - C_2 \epsilon_n^2 \max_{1 \leq i \leq n} \| \Delta_{\eta,i} - \Delta_{\eta_0,i} \|_F^2 \geq (C_1 - C_3 a_n \epsilon_n^2) d_{B,n}(\eta, \eta_0) - C_3 \epsilon_n \epsilon_n^2, \]

(2.32)

for some constants \( C_1, C_2, C_3 > 0 \). Since \( C_1 - C_3 a_n \epsilon_n^2 \) is bounded away from zero and \( \epsilon_n \) is decreasing, (2.31) and (2.32) imply that \( \epsilon_n \gtrsim d_{B,n}(\eta, \eta_0) \). Now, it is easy to see that by (2.5),

\[ \max_{1 \leq i \leq n} \| \Delta_{\eta,i} + \Delta_{\eta_0,i} \|_{sp}^2 \leq 2 \max_{1 \leq i \leq n} \| \Delta_{\eta,i} - \Delta_{\eta_0,i} \|_{sp}^2 + 8 \max_{1 \leq i \leq n} \| \Delta_{\eta_0,i} \|_{sp}^2 \leq \epsilon_n + a_n d_{B,n}(\eta, \eta_0) + 1, \]

which is bounded since \( \epsilon_n + a_n \epsilon_n^2 = o(1) \). Hence, we see that for \( \eta_s \) satisfying (C6), \( n^{-1} \| X(\theta - \theta_0) \|_2^2 + d_{A,n}(\eta, \eta_0) \) is bounded by a constant multiple of

\[ \frac{1}{n} \| X(\theta - \theta_0) \|_2^2 + d_{A,n}(\eta, \eta_s) + d_{A,n}(\eta_s, \eta_0) \]

\[ \lesssim \frac{1}{n} \sum_{i=1}^{n} \| X_i(\theta - \theta_0) + \xi_{\eta,i} - \xi_{\eta_0,i} \|_2^2 + d_{A,n}(\eta_s, \eta_0) \]

\[ \lesssim \frac{1}{n} \sum_{i=1}^{n} \| (\Delta_{\eta,i} + \Delta_{\eta_0,i})^{-1}(X_i(\theta - \theta_0) + \xi_{\eta,i} - \xi_{\eta_0,i}) \|_2^2 + d_{A,n}(\eta_s, \eta_0). \]

The display implies that \( \| X(\theta - \theta_0) \|_2^2 + nd_{A,n}(\eta, \eta_0) \lesssim n \epsilon_n^2 \) by Theorem 2.2 and (C6). Combining
the results verifies the third and fourth assertions of the theorem. For the remainder, observe that $s_\theta - \theta_0 \leq s_\theta + s_0 \leq K_1 s_\ast + s_0 \lesssim s_\ast \text{ for } \theta \text{ such that } s_\theta \leq K_1 s_\ast$. Therefore by Theorem 2.1, the first and the second assertions readily follow from the definitions of $\phi_1$ and $\phi_2$.

To prove the shape approximation in Theorem 2.4 and the selection results in Theorem 2.5, we first obtain two lemmas; the first shows that the remainder of the Bernstein-von Mises type theorem goes to zero in $P_{0^\ast}$-probability, while the second implies that with a point mass prior for $\theta$ at $\theta_0$, we also obtain the same posterior contraction rate for $\eta$ as in Theorem 2.3.

**Lemma 2.3.** Suppose that (C1), (C4), (C8), and (C12) are satisfied for some orthogonal projection $H$. Then, for $\Lambda_n^\ast(\theta)$ in (2.13) and $\Lambda_n^\ast(\theta, \eta) = (p_{\theta, \eta}/p_{\theta_0, \eta_0(\theta, \eta)})Y^{(n_h)}$ with the corresponding $H$, we have that

$$
\mathbb{E}_0 \sup_{\theta \in \Theta_n} \sup_{\eta \in \mathcal{H}_n} |\log \Lambda_n^\ast(\theta, \eta) - \log \Lambda_n^\ast(\theta)| \to 0.
$$

**Proof.** Let $\Delta_\eta \in \mathbb{R}^{n \times n^\ast}$ be the block-diagonal matrix formed by stacking $\Delta_{\eta_0,i}^1 \Delta_{\eta_0,i}^{-1} \Delta_{\eta_0,i}^{1/2}$, $i = 1, \ldots, n$. Observe that

$$
\log \Lambda_n^\ast(\theta, \eta) = -\frac{1}{2} \|\Delta_\eta^{1/2}(I - H)\tilde{X}(\theta - \theta_0)\|^2_2
+ (\theta - \theta_0)^T \tilde{X}^T(I - H)\Delta_{\eta}^\ast(U - (\hat{\xi}_\eta - \hat{\eta}_{0\eta}) - H\tilde{X}(\theta - \theta_0)),
$$

and hence it suffices to show the following three assertions:

$$
\sup_{\theta \in \Theta_n} \sup_{\eta \in \mathcal{H}_n} |(\theta - \theta_0)^T \tilde{X}^T(I - H)(I - \Delta_\eta^\ast)(I - H)\tilde{X}(\theta - \theta_0)| \to 0, \quad (2.33)
$$

$$
\sup_{\theta \in \Theta_n} \sup_{\eta \in \mathcal{H}_n} |(\theta - \theta_0)^T \tilde{X}^T(I - H)\Delta_\eta^\ast(\hat{\xi}_\eta - \hat{\eta}_{0\eta} + H\tilde{X}(\theta - \theta_0))| \to 0, \quad (2.34)
$$

$$
\mathbb{E}_0 \sup_{\theta \in \Theta_n} \sup_{\eta \in \mathcal{H}_n} |(\theta - \theta_0)^T \tilde{X}^T(I - H)(I - \Delta_\eta^\ast)U| \to 0. \quad (2.35)
$$

First, note that the left side of (2.33) is bounded above by a constant multiple of

$$
\sup_{\theta \in \Theta_n} \sup_{\eta \in \mathcal{H}_n} \|I - \Delta_\eta^\ast\|_{a'p} \|\tilde{X}(\theta - \theta_0)\|^2_2 \lesssim \sup_{\theta \in \Theta_n} \|X(\theta - \theta_0)\|^2_2 \sup_{\eta \in \mathcal{H}_n} \max_{1 \leq i \leq n} \|\Delta_{\eta,i}^{-1} - \Delta_{\eta_0,i}^{-1}\|_F. \quad (2.36)
$$

Using (i) of Lemma 2.6 and the inequality $|1 - x| \asymp |1 - x^{-1}|$ as $x \to 1$, we obtain that for
\[ \rho_{i,k}^* = \rho_k(\Delta_{\eta_0}^{1/2} \Delta_{\eta_i}^{-1} \Delta_{\eta_0,i}^{1/2}), \]

\[ \|\Delta_{\eta,i}^{-1} - \Delta_{\eta_0,i}^{-1}\|_F^2 \lesssim \sum_{k=1}^{m_i} (1 - \rho_{i,k}^*)^2 \lesssim \sum_{k=1}^{m_i} (1 - 1/\rho_{i,k}^*)^2 \lesssim \|\Delta_{\eta,i} - \Delta_{\eta_0,i}\|_F^2, \quad (2.37) \]

provided that the rightmost side is sufficiently small. Because \( \max_i \|\Delta_{\eta,i} - \Delta_{\eta_0,i}\|_F^2 \leq e_n + a_n d_{B,n}^2(\eta, \eta_0) \lesssim e_n + a_n \epsilon_n^2 \) on \( \mathcal{H}_n \), (2.37) holds. This implies that for all sufficiently large \( n \), the right hand side of (2.36) is bounded above by a constant multiple of

\[ \sup_{\eta \in \mathcal{H}_n} \|X\|_2 \|\theta - \theta_0\|_2^2 \sup_{\eta \in \mathcal{H}_n} \sqrt{e_n + a_n d_{B,n}^2(\eta, \eta_0)} \lesssim s_n \epsilon_n^2 \sqrt{e_n + a_n \epsilon_n^2}, \]

which goes to zero by (C12).

Next, the left side of (2.34) is equal to

\[ \sup_{\eta \in \mathcal{H}_n} \sup_{\eta \in \mathcal{H}_n} \left| (\theta - \theta_0)^T \hat{X}^T (I - H) \left\{ (\hat{\xi}_{\eta} - \hat{\xi}_{\eta_0}) - (I - \Delta_{\eta}^*) (\hat{\xi}_{\eta} - \hat{\xi}_{\eta_0} + H \hat{X}(\theta - \theta_0)) \right\} \right| \]

\[ \lesssim \sup_{\eta \in \mathcal{H}_n} \|X(\theta - \theta_0)\|_2 \|\|I - H(\tilde{\xi}_{\eta} - \tilde{\xi}_{\eta_0})\|_2 \]

\[ + \sup_{\eta \in \mathcal{H}_n} \left\{ \left( \|X(\theta - \theta_0)\|_2^2 + \|X(\theta - \theta_0)\|_2 \sqrt{n d_{A,n}(\eta, \eta_0)} \right) \max_{1 \leq i \leq n} \|\Delta_{\eta,i}^{-1} - \Delta_{\eta_0,i}^{-1}\|_F \right\}. \]

Using the same approach used in (2.37), the display is further bounded above by a constant multiple of

\[ \epsilon_n \sqrt{n s_n} \sup_{\eta \in \mathcal{H}_n} \|\|I - H(\tilde{\xi}_{\eta} - \tilde{\xi}_{\eta_0})\|_2 + s_n \epsilon_n^2 \sqrt{e_n + a_n \epsilon_n^2}, \]

which goes to zero by (C8) and (C12).

Now, note that (2.35) is bounded above by

\[ \sup_{\theta \in \Theta_n} \|\theta - \theta_0\|_1 \mathbb{E}_0 \sup_{\eta \in \mathcal{H}_n} \|\hat{X}^T (I - H)(I - \Delta_{\eta}^*) U\|_\infty, \quad (2.38) \]

by Hölder’s inequality. Let \( \bar{W}_{\eta,j} = \hat{X}^T_j (I - H)(I - \Delta_{\eta}^*) U \) for \( \hat{X}_{j} \in \mathbb{R}^{n*} \) the \( j \)th column of \( \hat{X} \). Then, by Lemma 2.2.2 of van der Vaart and Wellner (1996) applied with \( \psi(x) = e^{x^2} - 1 \), the expectation in (2.38) satisfies

\[ \mathbb{E}_0 \max_{1 \leq j \leq p} \sup_{\eta \in \mathcal{H}_n} |W_{\eta,j}| \leq \left\| \max_{1 \leq j \leq p} |W_{\eta,j}| \right\|_\psi \lesssim \sqrt{\log p} \max_{1 \leq j \leq p} \left\| \sup_{\eta \in \mathcal{H}_n} |W_{\eta,j}| \right\|_\psi, \quad (2.39) \]

where \( \|\cdot\|_\psi \) is the Orlicz norm for \( \psi \). For any \( \eta_1, \eta_2 \in \mathcal{H}_n \), define the standard deviation pseudo-
metric between $W_{\eta, j}$ and $W_{\eta' j}$ as

$$
  d_{\sigma,j}(\eta_1, \eta_2) := \sqrt{\text{Var}(W_{\eta_1 j} - W_{\eta' j})} = \|(\Delta_{\eta_1}^* - \Delta_{\eta' j}^*)(I - H)\hat{X}_{j}\|_2.
$$

Using the tail bound for normal distributions and Lemma 2.2.1 of van der Vaart and Wellner (1996), we see that $\|W_{\eta_1 j} - W_{\eta' j}\|_\psi \lesssim d_{\sigma,j}(\eta_1, \eta_2)$ for every $\eta_1, \eta_2 \in \mathcal{H}_n$. We shall show that $\mathcal{H}_n$ is a separable pseudo-metric space with $d_{\sigma,j}$ for every $j \leq p$. Then, under the true model $\mathbb{P}_0$, $\{W_{\eta j} : \eta \in \mathcal{H}_n\}$ is a separable Gaussian process for $d_{\sigma,j}$. Hence, by Corollary 2.2.5 of van der Vaart and Wellner (1996), for any fixed $\eta' \in \mathcal{H}_n$,

$$
  \left\| \sup_{\eta \in \mathcal{H}_n} |W_{\eta,j}| \right\|_\psi \lesssim \|W_{\eta', j}\|_\psi + \int_0^{\text{diam}_{\mathcal{H}_n}(\mathcal{H}_n)} \sqrt{\log N(e/2, \mathcal{H}_n, d_{\sigma,j})} de,
$$

(2.40)

where $\text{diam}_{\mathcal{H}_n}(\mathcal{H}_n) = \sup\{d_{\sigma,j}(\eta_1, \eta_2) : \eta_1, \eta_2 \in \mathcal{H}_n\}$. It is clear that $W_{\eta', j}$ possesses a normal distribution with mean zero and variance $\| (I - \Delta_{\eta'}^*)(I - H)\hat{X}_{j}\|_2^2$. Using Lemma 2.2.1 of van der Vaart and Wellner (1996) again, we see that

$$
  \|W_{\eta', j}\|_\psi \lesssim \|(I - \Delta_{\eta'}^*)(I - H)\hat{X}_{j}\|_2 \lesssim \max_{1 \leq i \leq \eta_1} \|\Delta_{\eta', i}^{-1} - \Delta_{\eta_0, i}^{-1}\|_2 \|X_{j}\|_2 \lesssim \|\|X\|_2 \sqrt{\epsilon_n + a_n \epsilon_n^2},
$$

(2.41)

for every $\eta' \in \mathcal{H}_n$, where the last inequality follows from (2.37). Next, to further bound the second term in (2.40), note that for every $\eta_1, \eta_2 \in \mathcal{H}_n$,

$$
  a_n \epsilon_n^2 \geq \sum_{k=1}^2 2a_n d_{\mathcal{H}_n}(\eta_k, \eta_0) \geq a_n d_{\mathcal{H}_n}(\eta_1, \eta_2) \geq \max_{1 \leq i \leq \eta_1} \|\Delta_{\eta_1, i} - \Delta_{\eta_2, i}\|_F^2,
$$

which is further bounded below by

$$
  \min_{1 \leq i \leq \eta_1} \rho_{\min}^2(\Delta_{\eta_2, i}) \max_{1 \leq i \leq \eta_1} \sum_{k=1}^{m_i} \left\{ 1 - 1/\rho_k(\Delta_{\eta_2, i}^{1/2} \Delta_{\eta_1, i}^{-1} \Delta_{\eta_2, i}^{1/2}) \right\}^2,
$$

using (i) of Lemma 2.6. In the last display, we see that $\min_i \rho_{\min}(\Delta_{\eta_2, i})$ is bounded away from zero since

$$
  \max_{1 \leq i \leq \eta_1} \|\Delta_{\eta_2, i}^{-1}\|_{sp} \leq \max_{1 \leq i \leq \eta_1} \|\Delta_{\eta_2, i}^{-1} - \Delta_{\eta_0, i}^{-1}\|_{sp} + \max_{1 \leq i \leq \eta_1} \|\Delta_{\eta_0, i}^{-1}\|_{sp} \lesssim \sqrt{\epsilon_n + a_n \epsilon_n^2} + 1,
$$

and hence every eigenvalue $\rho_k(\Delta_{\eta_2, i}^{1/2} \Delta_{\eta_1, i}^{-1} \Delta_{\eta_2, i}^{1/2})$ is bounded below and above by a multiple of its reciprocal, as $a_n \epsilon_n^2 \to 0$. This implies that $a_n \epsilon_n^2$ is further bounded below by a constant multiple...
Then, using (2.39), (2.40), and (2.41), and by the substitution for every \( \eta \)

Lemma 2.4. Suppose that the Dirac prior at \( \theta \)

Section 2.3.1, with the mixture of Laplace priors and point mass priors replaced by the Dirac measure at \( \theta_0 \) for a prior distribution of \( \theta \). The denominator of the posterior distribution with the Dirac prior at \( \theta_0 \) is bounded as in Lemma 2.1, which can be shown using (2.19) for the
prior concentration condition (C2) and the expressions for the Kullback-Leibler divergence $K(p_{0,i}, p_{0,η,i})$ and variation $\mathcal{V}(p_{0,i}, p_{0,η,i})$ with the true value $θ_0$. For a local test relative to the average Rényi divergence, Lemma 2.2 applied with $F_{1,n}$, modified so that it can be involved only with a given $η_1$ such that $R_n(p_0, p_{θ_0,η_1}) ≥ ϵ_n^2$, implies that a small piece of the alternative is tested with exponentially small errors. Hence, by (C5*), we obtain the contraction rate $ε_n$ relative to $R_n(p_0, p_{θ_0,η})$ for $Π_{θ_0}(\cdot | Y^{(n)})$, as in the proof of Theorem 2.2. The lemma is then obtained by recovering the contraction rate of $η$ with respect to $d_n$ using the approach in the proof of Theorem 2.3. In fact, the contraction rate in the lemma has room for improvement as $θ$ is fixed to $θ_0$ in the prior, but the given one is enough for our purpose.

Proof of Theorem 2.4. We use the fact that for any probability measure $Q$ and its renormalized restriction $Q_A(\cdot) = Q(\cdot \cap A)/Q(A)$ to a set $A$, we have $\|Q - Q_A\|_{TV} ≤ 2Q(A^c)$. First, using a sufficiently large constant $M_3$ that is smaller than $M_3$, define $\mathcal{H}'_n$ as $\mathcal{H}_n$ in (2.11) such that $\mathcal{H}'_n \subset \mathcal{H}_n$. Let $\Pi((θ, η) ∈ \cdot)$ be the prior distribution restricted and renormalized on $\bar{Θ}$ and $\Pi((θ, η) ∈ \cdot | Y^{(n)})$ be the corresponding posterior distribution. Also, $\Pi_{∞}(θ ∈ \cdot | Y^{(n)})$ is the restricted and renormalized version of $Π_{∞}(θ ∈ \cdot | Y^{(n)})$ to the set $\bar{Θ}$. Then the left hand side of the theorem is bounded above by

$$
\left\| \Pi(θ ∈ \cdot | Y^{(n)}) - \Pi(θ ∈ \cdot | Y^{(n)}) \right\|_{TV} + \left\| \Pi(θ ∈ \cdot | Y^{(n)}) - \Pi_{∞}(θ ∈ \cdot | Y^{(n)}) \right\|_{TV},
$$

where the first summand goes to zero in $P_0$-probability since $Π((θ, η) ∈ \bar{Θ} × \mathcal{H}_n | Y(n)) → 1$ in $P_0$-probability by Theorem 2.1 and Theorem 2.3.

To show that the second summand goes to zero in $P_0$-probability, note that for every measurable $B \subset \mathbb{R}^p$, we obtain

$$
\Pi(θ ∈ B | Y^{(n)}) ∝ \int_{B∩\bar{Θ}} \int_{\mathcal{H}'_n} p_{θ,η}(Y^{(n)}) e^{-λ||θ||_1}dΠ(η) dV(θ)
$$

$$
= \int_{B∩\bar{Θ}} \Lambda^*_n(θ, η) e^{-λ||θ||_1} p_{θ_0,\bar{Θ},η}(θ, η)(Y^{(n)}) dΠ(η) dV(θ),
$$

$$
\Pi_{∞}(θ ∈ B | Y^{(n)}) ∝ \int_{B∩\bar{Θ}} \Lambda^*_n(θ) dV(θ)
$$

$$
∝ \int_{B∩\bar{Θ}} \Lambda^*_n(θ) e^{-λ||θ||_1} \int_{\mathcal{H}} p_{θ,η}(Y^{(n)}) dΠ(η) dV(θ),
$$

where $dV(θ) = \sum_{S:s ≤ k_1} k_s \tau_p(s)(p^p)^{-1}(\lambda/2)^{s}d\{κ(θ_S) ⊗ δ_0(θ_{S^c})\}$. In the last line, the factor $e^{-λ||θ||_1} \int_{\mathcal{H}} p_{θ,η}(Y^{(n)}) dΠ(η)$ cancels out in the normalizing constant, but is inserted for the sake of comparison. For any sequences of measures $\{μ_S\}$ and $\{ν_S\}$, if $ν_S$ is absolutely contin-
uous with respect to $\mu_S$ with the Radon-Nikodym derivative $d\nu_S/d\mu_S$, then it can be easily verified that
\[
\left\| \frac{\sum_S \mu_S}{\|\sum_S \mu_S\|_{TV}} - \frac{\sum_S \nu_S}{\|\sum_S \nu_S\|_{TV}} \right\|_{TV} \leq 2 \sup_S \left\| \frac{\mu_S - \nu_S}{\|\sum_S \mu_S\|_{TV}} \right\|_{TV} \leq 2 \sup_S \left\| 1 - \frac{d\nu_S}{d\mu_S} \right\|_{\infty}.
\]
Hence, for $C_n = \int_\mathcal{H} p_{\theta_0,\eta}(Y^{(n)}) d\Pi(\eta)$, we see that the second summand of (2.43) is bounded by
\[
2 \sup_{\theta \in \Theta_n} \left| 1 - \frac{1}{C_n} \int_{\mathcal{H}_n} \Lambda_n^*(\theta, \eta) e^{-\lambda\|\theta\|_1} \Lambda_n^*(\theta) e^{-\lambda\|\theta_0\|_1} p_{\theta_0,\tilde{\eta}_n}(\theta, \eta) (Y^{(n)}) d\Pi(\eta) \right|.
\]
Using the fact that $|\lambda(\|\theta\|_1 - \|\theta_0\|_1)| \leq \lambda\|\theta - \theta_0\|_1 \lesssim \lambda_\epsilon_n \sqrt{n s^*_X}/\|X\|_* \to 0$ on $\tilde{\Theta}_n$ and that $\sup\{|1 - \Lambda_n^*(\theta, \eta)/\Lambda_n^*(\theta)| : \theta \in \Theta_n, \eta \in \mathcal{H}_n'\}$ goes to zero in $\mathbb{P}_0$-probability by Lemma 2.3, the last display is further bounded by
\[
2 \sup_{\theta \in \Theta_n} \left| 1 - \{1 + o(1) + o_\mathbb{P}_0(1)\} \frac{1}{C_n} \int_{\mathcal{H}_n} p_{\theta_0,\tilde{\eta}_n}(\theta, \eta) (Y^{(n)}) d\Pi(\eta) \right|.
\] (2.44)
Now, note that the map $\eta \mapsto \tilde{\eta}_n(\theta, \eta)$ is bijective for every fixed $\theta \in \tilde{\Theta}_n$. Thus for the set defined by $\tilde{\eta}_n(\theta, \tilde{\mathcal{H}}_n) = \{\tilde{\eta}_n(\theta, \eta) : \eta \in \tilde{\mathcal{H}}_n\}$ with given $\theta \in \tilde{\Theta}_n$, we see that
\[
\int_{\tilde{\mathcal{H}}_n} p_{\theta_0,\tilde{\eta}_n}(\theta, \eta) (Y^{(n)}) d\Pi(\eta) = \int_{\tilde{\eta}_n(\theta, \tilde{\mathcal{H}}_n)} p_{\theta_0,\eta}(\eta) (Y^{(n)}) d\Pi_{\theta_0,\eta}(\eta),
\] (2.45)
by the substitution in the integral. Observe that
\[
\tilde{\eta}_n(\theta, \tilde{\mathcal{H}}_n) = \{\eta \in \mathcal{H} : d_{A,n}^2(\eta, \tilde{\eta}_n(\theta, \eta_0)) \leq M_3' s^*_n \epsilon_n, d_{B,n}^2(\eta, \eta_0) \leq M_3^2 \epsilon_n^2\},
\]
and
\[
\|\tilde{\xi}_n - \tilde{\xi}_{\theta_0} - H \tilde{X}(\theta - \theta_0)\|_2 \times \sqrt{n d_{A,n}(\eta, \tilde{\eta}_n(\theta, \eta_0))}.
\]
Hence, we see that $M_3$ can be chosen sufficiently large such that $\tilde{\eta}_n(\theta, \tilde{\mathcal{H}}_n) \subset \tilde{\mathcal{H}}_n$ for every $\theta \in \tilde{\Theta}_n$ as we have $\sqrt{n d_{A,n}(\eta, \eta_0)} \lesssim \|\tilde{\xi}_n - \tilde{\xi}_{\theta_0} - H \tilde{X}(\theta - \theta_0)\|_2 + \|X\|_*\|\theta - \theta_0\|_1$. Therefore, (2.45) is equal to
\[
\int_{\tilde{\eta}_n(\theta, \tilde{\mathcal{H}}_n)} p_{\theta_0,\eta}(\eta) (Y^{(n)}) d\Pi(\eta) + o(1),
\]
by (C9), since $d\Pi(\eta) = d\Pi_{\theta_0,\eta}(\eta)$. Furthermore, $M_3'$ can be chosen sufficiently large such that $\{\eta \in \mathcal{H} : d_{A,n}^2(\eta, \eta_0) \leq K_0 \epsilon_n^2\} \subset \eta_n(\theta, \tilde{\mathcal{H}}_n)$ for every $\theta \in \tilde{\Theta}_n$. This is the case because we also have the inequality of the other direction $\|\tilde{\xi}_n - \tilde{\xi}_{\theta_0} - H \tilde{X}(\theta - \theta_0)\|_2 \lesssim \sqrt{n d_{A,n}(\eta, \eta_0)} + \|X\|_*\|\theta - \theta_0\|_1$.
Hence, with appropriately chosen constants, we obtain

\[
\inf_{\theta \in \Theta_n} \frac{\int_{\tilde{\Theta}_n} (\theta, \tilde{\Theta}_n) p_{\theta_0, \eta} (Y^{(n)}) d\Pi(\eta)}{p_{\eta} \int_{\tilde{\Theta}_n} p_{\theta_0, \eta} (Y^{(n)}) d\Pi(\eta)} = \inf_{\theta \in \Theta_n} \Pi^{\theta_0} \left( \eta \in \tilde{\Theta}_n (\theta, \tilde{\Theta}_n) | Y^{(n)} \right) \\
\geq \Pi^{\theta_0} \left( d_n (\eta, \eta_0) \leq K_0 \epsilon_n \big| Y^{(n)} \right).
\]

The rightmost term goes to one with probability tending to one by Lemma 2.4. This implies that (2.44) goes to zero in \( P_0 \)-probability, completing the proof for the second part of (2.43).

Next, we show that \( \Pi^\infty (\theta \in \Theta_n | Y^{(n)}) \) goes to one in \( P_0 \)-probability to verify that the last summand in (2.43) goes to zero in \( P_0 \)-probability. Observe that \( \Pi^\infty (\theta \in \Theta_n | Y^{(n)}) \) is equal to

\[
\frac{\exp \left\{ -\frac{1}{2} \| (I - H) \tilde{X} (\theta - \theta_0) \|^2_2 + U^T (I - H) \tilde{X} (\theta - \theta_0) \right\}}{\int_{\mathbb{R}^p} \exp \left\{ -\frac{1}{2} \| (I - H) \tilde{X} (\theta - \theta_0) \|^2_2 + U^T (I - H) \tilde{X} (\theta - \theta_0) \right\} d\theta}.
\]

Clearly, the denominator is bounded below by

\[
\pi_p (s_0) \left( \frac{\lambda}{2} \right)^{s_0} \int_{\mathbb{R}^n} \exp \left\{ -\frac{1}{2} \| (I - H) \tilde{X}_{S_0} (\theta_{S_0} - \theta_{0, S_0}) \|^2_2 \right\} d\theta_{S_0}.
\]

Since the measure \( Q \) defined by \( Q (d\theta_{S_0}) = \exp \{ -1/2 \| (I - H) \tilde{X}_{S_0} (\theta_{S_0} - \theta_{0, S_0}) \|^2_2 \} \) is symmetric about \( \theta_{0, S_0} \), the mean of \( (\theta_{S_0} - \theta_{0, S_0}) \) with respect to the normalized probability measure \( \tilde{Q} = Q / Q (\mathbb{R}^n) \) is zero. Note also that \( \Gamma_S = \tilde{X}^T_S (I - H) \tilde{X}_S \) is nonsingular for every \( S \) such that \( s \leq K_1 s_* \) by (C10). Thus, by Jensen’s inequality, (2.47) is bounded below by

\[
\pi_p (s_0) \left( \frac{\lambda}{2} \right)^{s_0} \int_{\mathbb{R}^n} \exp \left\{ -\frac{1}{2} \| (I - H) \tilde{X}_{S_0} (\theta_{S_0} - \theta_{0, S_0}) \|^2_2 \right\} d\theta_{S_0} = \pi_p (s_0) \left( \frac{\lambda}{2} \right)^{s_0} \frac{(2\pi)^{s_0}/2}{\det (\Gamma_{S_0})^{1/2}}.
\]

Applying the arithmetic-geometric mean inequality to the eigenvalues, we obtain \( \det (\Gamma_{S_0}) \leq (\text{tr} (\Gamma_{S_0}) / s_0) s_0 \leq \| (I - H) \tilde{X}_{S_0} \|_*^2 s_0 \leq \| \tilde{X}_{S_0} \|_*^2 s_0 \leq \| \tilde{X}_{S_0} \|_*^2 s_0 \leq \| \tilde{X}_{S_0} \|_*^2 s_0 \leq \| \tilde{X}_{S_0} \|_*^2 s_0 \) by (2.4). Furthermore, we have \( \pi_p (s_0) \gtrsim A^p_1 \| p - A \| s_0 \) by (2.3) and \( \left( \frac{\lambda}{2} \right)^{s_0} \leq p^s_0 \). Hence, the preceding display is further bounded below by a constant multiple of

\[
p^{-(1 + L_2 + A_2) s_0} \left( \frac{A^p_1 \sqrt{L_0}}{L_1} \right)^{s_0}.
\]

To bound the numerator of (2.46), note first that \( \tilde{X}^T_J (I - H) U \) has a normal distribution with
mean zero and variance \( \|(I - H)\tilde{X}_j\|^2 \), and hence we have

\[
\mathbb{P}_0 \left( \|X^T(I - H)U\|_\infty > t \max_{1 \leq j \leq p} \|(I - H)\tilde{X}_j\|_2 \right) \leq 2pe^{-t^2/2}, \quad t > 0,
\]

by the tail probabilities of normal distributions. By choosing \( t = 2\sqrt{\log p} \) and using the inequality \( \|(I - H)\tilde{X}_j\|_2 \leq \|\tilde{X}_j\|_2 \leq 2^{-1/2}\|X\|_* \) for every \( j \leq p \), we thus obtain

\[
\mathbb{P}_0 \left( \|X^T(I - H)U\|_\infty > 2\rho_0^{-1/2}\sqrt{\log p\|X\|_*} \right) \leq \frac{2}{p}. \tag{2.49}
\]

Let \( D_n = 2\rho_0^{-1/2}\sqrt{\log p\|X\|_*} \) and \( U_n = \{ \|X^T(I - H)U\|_\infty \leq D_n \} \). Then it suffices to show that (2.46) goes to zero in \( \mathbb{P}_0 \)-probability on the set \( U_n \) as (2.49) goes to zero. Note that on the set \( U_n \) we have

\[
U^T(I - H)\tilde{X}(\theta - \theta_0) \leq D_n\|\theta - \theta_0\|_1 \leq D_n \frac{2\sqrt{\rho_0}\|X\|_*\|X^T(I - H)\tilde{X}(\theta - \theta_0)\|_2}{\|X\|_*\|\tilde{X}(\theta - \theta_0)\|_2} - D_n\|\theta - \theta_0\|_1.
\]

Using that \( \|u\|_2 \lesssim \|(I - H)u\|_2 \) for every \( u \in \text{span}(\tilde{X}_S) \) with \( s \leq K_1s_* \) by (C10), the preceding display is, for some constant \( C_1 > 0 \), further bounded above by

\[
D_n \frac{2\sqrt{\rho_0}C_1\|(I - H)\tilde{X}(\theta - \theta_0)\|_2}{\|X\|_*\|\tilde{X}(\theta - \theta_0)\|_2} - D_n\|\theta - \theta_0\|_1 \leq \frac{1}{2}\|(I - H)\tilde{X}(\theta - \theta_0)\|_2^2 + 2\sqrt{\rho_0}C_1^2D_n^2\|\tilde{X}(\theta - \theta_0)\|_2^2 - D_n\|\theta - \theta_0\|_1,
\]

by the Cauchy-Schwarz inequality. We have \( s_{\theta - \theta_0} \leq K_1s_* + s_0 \) on the support of the measure \( V \). Hence, on the event \( U_n \), the numerator of (2.46) is bounded above by

\[
\exp \left\{ \frac{2\sqrt{\rho_0}C_1^2D_n^2(K_1s_* + s_0)}{\|X\|^2_*\phi_1(K_1s_* + s_0)^2} - \frac{M_2D_n\epsilon_n\sqrt{s_*}n}{2\|X\|_*} \right\} \sum_{s:s \leq K_1s_*} \pi_p(s) \left( \frac{\lambda}{2} \right)^s e^{-\left(D_n/2\right)\|\bar{\theta}_S - \theta_0\|_1}\,d\theta_S
\]

\[
\leq \exp \left\{ \frac{2\sqrt{\rho_0}C_1^2(K_1 + 1)s_* \log p}{\rho_0\phi_1(K_1s_* + s_0)^2} - \frac{M_2s_* \log p}{\sqrt{\rho_0}} \right\} \sum_{s=0}^p \pi_p(s) \left( L_3\sqrt{\frac{\rho_0}{n}} \right)^s,
\]

since \( D_n \geq \lambda\sqrt{n}/(L_3\sqrt{\rho_0}) \) and \( n\epsilon_n \geq s_* \log p \). Note that we have \( \sum_{s=0}^p \pi_p(s)(L_3\sqrt{\rho_0/n})^s \lesssim \sum_{s=0}(A_2L_3p^{-A_4}(\rho_0/n)^{1/2})^s \lesssim \sum_{s=0}B_2L_3p^{-A_4} = 1 \) by (2.3) and that \( \phi_1(K_1s_* + s_0) \) in the denominators is bounded away from zero by the assumption. Thus, the last display combined with (2.48) shows that (2.46) goes to zero on the event \( U_n \), provided that \( M_2 \) is chosen sufficiently large.

Finally we conclude that (2.43) goes to zero in \( \mathbb{P}_0 \)-probability. Since the total variation metric is bounded by 2, the convergence in mean holds as well. \[\square\]
Proof of Theorem 2.5. Since $\mathbb{E}_0 \| \Pi(\theta \in \cdot | Y^{(n)}) - \Pi^\infty(\theta \in \cdot | Y^{(n)}) \|_{TV}$ tends to zero by Theorem 2.4, it suffices to show that $\mathbb{E}_0 \Pi^\infty(\theta : S_0 \in S_n | Y^{(n)}) \to 0$ for $S_n = \{ S : s \leq K_1 s_* , S \supset S_0, S \neq S_0 \}$. For the orthogonal projection defined by $\tilde{H}_S = (I - H) \tilde{X}_S \Gamma_S^{-1} \tilde{X}_S^T (I - H)$ with $\Gamma_S = \tilde{X}_S^T (I - H) \tilde{X}_S$, we obtain

$$
\Pi^\infty(\theta : S_0 \in S_n | Y^{(n)}) \leq K_1 s_* \sum_{s = s_0 + 1}^s \frac{\pi_p(s)}{\pi_p(s_0)} \left( \frac{\lambda \sqrt{\pi}}{2} \right)^{s - s_0} \max_{S \in S_n : |S| = s} \left[ \frac{\det(\Gamma_S)^{1/2}}{\det(\Gamma_S)^{s - s_0}} \| (\tilde{H}_S - \tilde{H}_{S_0}) u \|_2^2 \right],
$$

by (2.12), since $(\tilde{H}_S - \tilde{H}_{S_0}) \tilde{X}_0 = (\tilde{H}_S - \tilde{H}_{S_0})(I - H) \tilde{X}_S \theta_{0, S_0} = 0$ for every $S \in S_n$ due to $S_0 \subset S$ on $S_n$. Note that $\rho_k(\Gamma_S) \leq \rho_k(\Gamma_S)$ for $k = 1, \ldots, s_0$, because $\Gamma_S$ is a principal submatrix of $\Gamma_S$. Hence, for some $C_1 > 0$,

$$
\det(\Gamma_S) = \prod_{k = 1}^{s_0} \rho_k(\Gamma_S) \leq \prod_{k = 1}^{s_0} \rho_k(\Gamma_S) \leq \frac{\det(\Gamma_S)}{\rho_{\min}(\Gamma_S)^{s - s_0}} \leq \frac{\det(\Gamma_S)}{(C_1 + \rho_0^{-1/2} \phi_2(s))^{s - s_0}}. \tag{2.50}
$$

The last inequality holds since by (C10), there exist a constant $C_1 > 0$ such that $C_1^2 \| v \|_2^2 \leq \| (I - H) v \|_2^2$ for every $v \in \text{span}(\tilde{X}_S)$ with $s \leq K_1 s_*$, and hence we have that by the definition of $\phi_2$,

$$
\rho_{\min}(\Gamma_S) = \inf_{u \in \mathbb{R}^s, u \neq 0} \frac{\| (I - H) \tilde{X}_S u \|_2^2}{\| u \|_2^2} \geq \frac{C_1^2 \phi_2(s)^2 \| X \|_2^2}{\rho_0}.
$$

Now, we shall show that for any fixed $b > 2$,

$$
\mathbb{P}_0 \left( \| (\tilde{H}_S - \tilde{H}_{S_0}) U \|_2^2 \leq b(s - s_0) \log p, \text{ for every } S \in S_n \right) \to 1. \tag{2.51}
$$

Note that $\| (\tilde{H}_S - \tilde{H}_{S_0}) U \|_2^2$ has a chi-squared distribution with degree of freedom $s - s_0$. Therefore, by Lemma 5 of Castillo et al. (2015), there exists a constant $C_2$ such that for every $b > 2$ and given $s \geq s_0 + 1$,

$$
\mathbb{P}_0 \left( \max_{S \in S_n : |S| = s} \| (\tilde{H}_S - \tilde{H}_{S_0}) U \|_2^2 > b \log N_s \right) \leq \left( \frac{1}{N_s} \right)^{(b-2)/4} e^{C_2(s - s_0)},
$$

where $N_s = \binom{p - s_0}{s - s_0}$ is the cardinality of the set $\{ S \in S_n : |S| = s \}$. Since $N_s \leq (p - s_0)^{s - s_0} \leq p^{s - s_0}$, for $T_n$ the event in the relation (2.51), it follows that

$$
\mathbb{P}_0(T_n^c) \leq \sum_{s = s_0 + 1}^{K_1 s_*} \left( \frac{1}{N_s} \right)^{(b-2)/4} e^{C_2(s - s_0)}.
$$
This goes to zero as $p \to \infty$, since for $s \leq K_1 s_*$,

\[ N_s \geq \frac{(p - s)^{s-s_0}}{(s - s_0)!} \geq \frac{(p - K_1 s_*)^{s-s_0}}{(s - s_0)^{s-s_0}} \geq \left( \frac{p - K_1 s_*}{K_1 s_*} \right)^{s-s_0}, \]

and $s_*/p = o(1)$. To complete the proof, it remains to show that $\Pi^\infty(\theta : S_0 \in S_n | Y^{(n)})$ goes to zero on the set $\mathcal{T}_n$. Combining (2.50) and (2.51), we see that $\Pi^\infty(\theta : S_0 \in S_n | Y^{(n)}) \mathcal{T}_n$ is bounded by

\[
\sum_{s=s_0+1}^{K_1 s_*} \pi_p(s) \left( \frac{p}{s_0} \right)^{p-s_0} \left( \frac{\lambda \sqrt{\pi}}{\sqrt{2}} \right)^{s-s_0} \left( \frac{\sqrt{p_0 p_b}}{C_1 \phi_2(s) \| X \|_s} \right)^{s-s_0} \leq K_1 s_* \left( \frac{A_2}{p A_4} \right)^{s-s_0} \left( \frac{s}{s_0} \right) \left( \frac{L_3}{C_1 \phi_1(K_1 s_*) \sqrt{K_1 s_* \pi p_0 p_b}} \right)^{s-s_0},
\]

which holds by the inequalities $\pi_p(s)/\pi_p(s_0) \leq (A_2 p^{-A_4})^{s-s_0}$ and $(p/s_0)^{(p-s_0)/(s_0)} = (s/s_0)$. Note that for $s \leq K_1 s_*$, we have that $(s/s_0) \leq (K_1 s_*)^{s-s_0} \leq (K_1 C_2^2)^{s-s_0}$ for some $C_2 > 0$. Hence, the preceding display goes to zero provided that $a - A_4 + b/2 < 0$ since $s_* = o(n)$. This condition can be translated to $a < A_4 - 1$ by choosing $b$ arbitrarily close to 2.

2.5.2 Proofs for the applications

Proof of Theorem 2.6, (a)–(b). We verify the conditions for the optimal posterior contraction in Corollary 2.1.

- Verification of (C1): Since $\Delta_n,i = \Sigma$ for every $i \leq n$ and $\Sigma_0$ belongs to the support of the prior, $a_n = 1$ and $e_n = 0$ satisfy (C1).

- Verification of (C2): Note that for every $\eta_1, \eta_2 \in \mathcal{H}$,

\[ d_n^2(\eta_1, \eta_2) \leq \frac{q \| Z \|_{sp}^2}{n} \| \beta_1 - \beta_2 \|_\infty^2 + \| \Sigma_1 - \Sigma_2 \|_F^2. \]  

(2.52)

We thus have

\[ \log \Pi(\eta \in \mathcal{H} : d_n(\eta, \eta_0) \leq \epsilon_n) \geq \log \Pi \left( \beta : \| \beta - \beta_0 \|_\infty \leq \frac{\sqrt{n} \epsilon_n}{\sqrt{2q \| Z \|_{sp}}} \right) \]

\[ + \log \Pi \left( \Sigma : \| \Sigma - \Sigma_0 \|_F \leq \frac{\epsilon_n}{\sqrt{2}} \right). \]

Since $\epsilon_n^2 > n^{-1}$, $\| \beta_0 \|_\infty \leq 1$, and $1 \leq \rho_{\min}(\Sigma_0) \leq \rho_{\max}(\Sigma_0) \leq 1$, the first term on the right hand side is bounded below by a multiple of $-q \log n$ due to $q < n$ and $\log \| Z \|_{sp} \leq \log n$, while we bound the second term below by a multiple of $-\log n$ as $\bar{m}$ is fixed. Hence we choose
\[ \epsilon_n = \sqrt{(q \log n)/n}. \]

- **Verification of (C3):** The assumption \( \lambda \| \theta_0 \|_1 \lesssim s_\star \log p \) given in the theorem directly satisfies (C3).

- **Verification of (C4):** This is also a direct consequence of the assumptions on \( \beta_0 \) and \( \Sigma_0 \).

- **Verification of (C5\dagger):** For a sufficiently large \( M \) and \( s_\star = s_0 \vee (q \log n / \log p) \), define \( \mathcal{H}_n = \{ \beta : \| \beta \|_\infty \leq n^M \} \times \{ \Sigma : n^{-M} \leq \rho_{\min}(\Sigma) \leq \rho_{\max}(\Sigma) \leq e^{Ms_\star \log p} \} \), so that the minimum eigenvalue condition (2.6) can be satisfied with \( \log \gamma_n \asymp \log n \). To find a sequence \( \epsilon_n \) satisfying the entropy condition (2.7), note that using (2.52),

\[
\log N \left( \frac{1}{6Mn^{M+3/2}}, \mathcal{H}_n, d_n \right) \leq \log N \left( \frac{1}{6M\sqrt{q} \| Z \|_{sp} n^{M+1}}, \left\{ \beta : \| \beta \|_\infty \leq n^M \right\}, \| \cdot \|_\infty \right) + \log N \left( \frac{1}{6Mn^{M+3/2}}, \left\{ \Sigma : \| \Sigma \|_F \leq \sqrt{me^{Ms_\star \log p}} \right\}, \| \cdot \|_F \right)
\]

and hence we choose \( \epsilon_n = \sqrt{(s_\star \log p)/n} \) so that the rightmost side of the last display is bounded by a multiple of \( n \epsilon_n^2 = s_\star \log p \). To verify the sieve condition (2.8), note that for some positive constants \( b_1, b_2, b_3, b_4 \) and \( b_5 \), an inverse Wishart distribution satisfies

\[
\Pi(\Sigma : \rho_{\min}(\Sigma) < n^{-M}) \leq b_1 e^{-b_2 n^{M}}, \\
\Pi(\Sigma : \rho_{\max}(\Sigma) > e^{Ms_\star \log p}) \leq b_4 e^{-b_5 Ms_\star \log p};
\]  

see, for example, Lemma 9.16 of Ghosal and van der Vaart (2017). Moreover, since the tail bounds of normal distributions give \( \Pi(\beta : \| \beta \|_\infty > n^M) \leq q \exp(-b_6 n^{2M}) \) for some constant \( b_6 > 0 \), the condition (2.8) is met if \( M \) is sufficiently large.

- **Verification of (C6\star):** Since \( \beta_0 \) is in the support of the prior, (C6\star) is satisfied by Remark 2.1.

Therefore by Corollary 2.1, we have the posterior contraction properties of \( \theta \) with \( s_\star = s_0 \vee (q \log n / \log p) \). The corollary also shows that the posterior distributions of \( n^{-1/2} \| Z(\beta - \beta_0) \|_2 \) and \( \| \Sigma - \Sigma_0 \|_F \) contract at the rate \( \sqrt{(s_0 \log p \vee q \log n)/n} \). The rate for \( \beta \) with respect to the \( \ell_2 \)-norm is then easily obtained with the minimum eigenvalue of \( Z^T Z \).

**Proof of Theorem 2.6, (c)–(e).** We shall verify the additional conditions in Theorems 2.4–2.5 and Corollaries 2.2–2.3. Specifically, we check the conditions (C8)–(C12) below, while the remaining conditions will be mentioned in stating the assertions.

- **Verification of (C8):** Note that \( \| (I - H) \tilde{Z}(\beta - \beta_0) \|_2 = 0 \) for \( H \) defined in the theorem. Hence (C8) is trivially met.
• Verification of (C9): We use the standard normal distribution for a prior of each entry of \( \beta \) without loss of generality, but the implications can be easily extended to a multivariate normal prior with arbitrary parameters. Since the priors for \( \Sigma \) cancels out in the Radon-Nikodym derivative due to invariance, we observe that

\[
\left| \log \frac{d\Pi_{n,\theta}}{d\Pi_{n,\theta_0}} (\eta) \right| \lesssim \left\| \beta \right\|_2^2 - \left\| \beta + (\tilde{Z}^T \tilde{Z})^{-1} \tilde{Z}^T \hat{X}(\theta - \theta_0) \right\|_2^2 \\
\leq 2\left\| \beta \right\|_2 \left\| (\tilde{Z}^T \tilde{Z})^{-1} \tilde{Z}^T \hat{X}(\theta - \theta_0) \right\|_2 + \left\| (\tilde{Z}^T \tilde{Z})^{-1} \tilde{Z}^T \hat{X}(\theta - \theta_0) \right\|_2^2.
\]

We see that

\[
\sup_{\theta \in \bar{\Theta}_n} \left\| (\tilde{Z}^T \tilde{Z})^{-1} \tilde{Z}^T \hat{X}(\theta - \theta_0) \right\|_2 \lesssim \|Z\|_{sp} s_* \sqrt{\log p / \rho_{\min}(Z^T Z)},
\]

and

\[
\sup_{\eta \in \bar{H}_n} \|\beta\|_2 \leq \sup_{\eta \in \bar{H}_n} \|\beta - \beta_0\|_2 + \|\beta_0\|_2 \\
\lesssim s_* \sqrt{(\log p) / \rho_{\min}(Z^T Z)} + 1 \\
\lesssim \|Z\|_{sp} s_* \sqrt{\log p / \rho_{\min}(Z^T Z)} + 1,
\]

since the condition number \( \|Z\|_{sp} \sqrt{\rho_{\min}(Z^T Z)} \) of \( Z \) is always greater than or equal to 1. Hence, the condition (C9) is satisfied by the assumption \( \|Z\|_{sp} s_* \sqrt{\log p / \rho_{\min}(Z^T Z)} = o(1) \).

• Verification of (C10): This is directly satisfied by Remark 2.2.

• Verification of (C11): This is trivially satisfied as the parameter space of \( \Sigma \) is a Euclidean space and \( d_{B,n} \) is the Frobenius norm.

• Verification of (C12): Note that the entropy of interest is bounded above by a constant multiple of

\[
\log N \left( \delta, \left\{ \Sigma : \|\Sigma - \Sigma_0\|_F^2 \leq M_3 \epsilon_n^2 \right\}, \|\cdot\|_F \right) \lesssim 0 \vee \log \left( \frac{3 \sqrt{M_3} \epsilon_n}{\delta} \right),
\]

for all \( \delta > 0 \). Thus the second term of (C12) is bounded by \( (s_3^3 \log^3 p / n)^{1/2} \) by Remark 2.3, while it can be easily checked that the first term is bounded by \( (s_5^5 \log^5 p / n)^{1/2} \). Therefore, (C12) is satisfied by the assumption that \( s_5^5 \log^5 p = o(n) \).

Therefore, under (C7), we have the distributional approximation in (2.14) by Theorem 2.4. Under (C7) and (C13), Theorem 2.5 implies that the no-superset result in (2.15) holds. The stronger assertions in (2.16) and (2.17) are explicitly derived from Corollary 2.2 and Corollary 2.3 if the beta-min condition (C14) is also met. \(\square\)
Proof of Theorem 2.7, (a)–(b). We verify the conditions for the optimal posterior contraction in Corollary 2.1.

- **Verification of (C1):** Note that for \( \hat{\sigma}_{jk} \) the \((j,k)\)th element of \( \Sigma - \Sigma_0 \), we obtain

\[
d_n^2(\Sigma, \Sigma_0) = \frac{1}{n} \sum_{i=1}^{n} \| E_i^T (\Sigma - \Sigma_0) E_i \|_F^2 = \frac{1}{n} \sum_{j=1}^{m} \sum_{k=1}^{m} \left[ \hat{\sigma}_{jk}^2 + \sum_{i=1}^{n} e_{ij} e_{ik} \right] \geq \frac{1}{c_n} \| \Sigma - \Sigma_0 \|_F^2.
\]

Hence, we see that \( c_n \) has the same role as \( a_n \) and the condition \( a_n c_n^2 \) is satisfied by the assumption \( c_n s_* \log p = o(n) \). We also have \( \epsilon_n = 0 \) as the true \( \Sigma_0 \) belongs to the support of the prior.

- **Verification of (C2):** Note that

\[
d_n^2(\Sigma_1, \Sigma_2) = \frac{1}{n} \sum_{i=1}^{n} \| E_i^T (\Sigma_1 - \Sigma_2) E_i \|_F^2 \leq \| \Sigma_1 - \Sigma_2 \|_F^2,
\]

for every \( \Sigma_1, \Sigma_2 \in \mathcal{H} \). Hence we obtain that for every \( \epsilon_n > n^{-1/2} \),

\[
\log \Pi(d_n(\Sigma, \Sigma_0) \leq \epsilon_n) \geq \log \Pi(\| \Sigma - \Sigma_0 \|_F \leq \epsilon_n) \gtrsim \log \epsilon_n \gtrsim - \log n,
\]

since \( 1 \lesssim \rho_{\min}(\Sigma_0) \lesssim \rho_{\max}(\Sigma_0) \lesssim 1 \). This leads us to choose \( \epsilon_n = \sqrt{(\log n)/n} \) for (C2) to be satisfied.

- **Verification of (C3):** The assumption \( \lambda \| \theta_0 \|_1 \lesssim s_* \log p \) given in the theorem directly satisfies (C3).

- **Verification of (C4):** Observe that we have the inequalities the assumption and the inequalities \( \rho_{\min}(\Sigma_0) \leq \rho_{\min}(E_i^T \Sigma_0 E_i) \leq \rho_{\max}(E_i^T \Sigma_0 E_i) \leq \rho_{\max}(\Sigma_0) \) for every \( i \leq n \) as \( E_i^T \Sigma_0 E_i \) is a principal submatrix of \( \Sigma_0 \). Hence (C4) is directly satisfied by the assumption on \( \Sigma_0 \).

- **Verification of (C5):** For a sufficiently large \( M > 0 \) and \( s_* = s_0 \vee (\log n/\log p) \), choose \( \mathcal{H}_n = \{ \Sigma : n^{-M} \leq \rho_{\min}(\Sigma) \leq \rho_{\max}(\Sigma) \leq e^{M s_* \log p} \} \). Since \( E_i^T \Sigma E_i \) is a principal submatrix of \( \Sigma \), we have \( \rho_{\min}(E_i^T \Sigma E_i) \geq \rho_{\min}(\Sigma) \geq n^{-M} \) for every \( i \leq n \) and \( \Sigma \in \mathcal{H}_n \). Hence the minimum eigenvalue condition (2.6) is satisfied with \( \log \gamma_n \asymp \log n \). Also, the entropy relative to \( d_n \) is given by

\[
\log N \left( \frac{1}{6m^M M + 3/2}, \mathcal{H}_n, d_n \right) \lesssim \log N \left( \frac{1}{6m^M M + 3/2}, \left\{ \Sigma : \| \Sigma \|_F \leq \sqrt{\pi} e^{M s_* \log p} \right\}, \| \cdot \|_F \right) \lesssim \log n + s_* \log p.
\]

The entropy condition (2.7) is thus satisfied if we choose \( \epsilon_n = \sqrt{(s_* \log p)/n} \). Lastly, using the tail bounds for an inverse Wishart distribution in (2.53), we meet the sieve condition.
(2.8) provided that $M$ is chosen sufficiently large.

- **Verification of (C6*)**: The separability condition is trivially satisfied in this example as there is no nuisance mean part.

Therefore, the contraction properties in Corollary 2.1 are obtained with $s_\ast = s_0 \sqrt{(\log n / \log p)}$. The contraction rate for $\Sigma$ with respect to the Frobenius norm follows from (2.54)

**Proof of Theorem 2.7, (c)–(e).** We verify the conditions (C8)–(C12) to apply Theorems 2.4–2.5 and Corollaries 2.2–2.3.

- **Verification of (C8)–(C10)**: These conditions are trivially satisfied with the zero matrix $H$ as there is no nuisance mean part.
- **Verification of (C11)**: Note that $d_{B,n}(\Sigma_1, \Sigma_2) \leq \|\Sigma_1 - \Sigma_2\|_F$ for every $\Sigma_1, \Sigma_2$ by (2.55), and hence it suffices to show that $H$ is a separable metric space with the Frobenius norm. Since the support of the prior for $\Sigma$ is Euclidean, separability with the Frobenius norm is trivial.
- **Verification of (C12)**: Since the entropy in (C12) is bounded above by a constant multiple of $\log N(\delta, \{\Sigma : \|\Sigma - \Sigma_0\|_F \leq M_3 c_n \epsilon_n^2\}, \|\cdot\|_F) \leq 0$ using (2.54) and (2.55), the term in (C12) is bounded by a multiple of $(s_\ast \sqrt{\log c_n} \sqrt{c_n(s_\ast \log p)^3/n})$ by Remark 2.3.

Hence, under (C7), Theorem 2.4 can be applied to obtain the distributional approximation in (2.14) with the zero matrix $H$. Under (C7) and (C13), Theorem 2.5 implies the no-superset result in (2.15). If the beta-min condition (C14) is also met, the strong results in Corollary 2.2 and Corollary 2.3 hold.

**Proof of Theorem 2.8, (a)–(b).** We verify the conditions for the optimal posterior contraction in Corollary 2.1.

- **Verification of (C1)**: Since $\Delta_{\eta_1}$ is the same for every $i \leq n$ and the true parameters belong to the support of the prior, we see that $a_n = 1$ and $e_n = 0$ satisfy (C1).
- **Verification of (C2)**: Note that for every $\eta_1, \eta_2 \in \mathcal{H}$,

$$
\|\xi_{\eta_1} - \xi_{\eta_2}\|_2^2 = |(\alpha_1 - \alpha_2) + (\mu_1^T \beta_1 - \mu_2^T \beta_2)|^2 + \|\mu_1 - \mu_2\|_2^2 \\
\quad \lesssim |\alpha_1 - \alpha_2|^2 + \|\mu_1\|_2^2 \|\beta_1 - \beta_2\|_2^2 + \|\beta_2\|_2^2 \|\mu_1 - \mu_2\|_2^2,
$$

$$
\|\Delta_{\eta_1} - \Delta_{\eta_2}\|_F^2 = |(\beta_1^T \Sigma_1 \beta_1 - \beta_2^T \Sigma_2 \beta_2) + (\sigma_1^2 - \sigma_2^2)|^2 \\
\quad + 2\|\Sigma_1 \beta_1 - \Sigma_2 \beta_2\|_2^2 + \|\Sigma_1 - \Sigma_2\|_F^2 \\
\quad \lesssim (\|\beta_1\|_2^2 + 1)^2 \|\Sigma_1 - \Sigma_2\|_F^2 + |\sigma_1^2 - \sigma_2^2|^2 \\
\quad + (\|\beta_1\|_2^2 + \|\beta_2\|_2^2 + 1)\|\Sigma_2\|_F^2 \|\beta_1 - \beta_2\|_2^2.
$$

(2.56)
Since $\|\beta_0\|_2$, $|\sigma_0^2|$, and $\|\Sigma_0\|_F$ are bounded, it follows from the last display that there exists a constant $C_1$ such that $|\alpha - \alpha_0| + \|\beta - \beta_0\|_2 + \|\mu - \mu_0\|_2 + |\sigma^2 - \sigma_0^2| + \|\Sigma - \Sigma_0\|_F \leq C_1 \epsilon_n$ implies $d_n(\eta_1, \eta_2) \leq \epsilon_n$ for any small $\epsilon_n$. This shows that (C2) is satisfied as long as we choose $\epsilon_n = \sqrt{\log n/n}$, as we have $|\alpha_0| \vee \|\beta_0\|_\infty \vee \|\mu_0\|_\infty \leq 1$, $\sigma_0^2 \asymp 1$, and $1 \geq \rho_{\min}(\Sigma_0) \leq \rho_{\max}(\Sigma_0) \leq 1$.

- **Verification of (C3):** The assumption $\lambda\|\theta_0\|_1 \leq s_* \log p$ given in the theorem directly satisfies (C3).

- **Verification of (C4):** Since $\Delta_{\eta_0}$ can be written as the sum of two positive definite matrices as
  \[
  \Delta_{\eta} = \begin{pmatrix} \beta^T \Sigma \beta & \beta^T \Sigma \\ \Sigma \beta & \Sigma \end{pmatrix} + \begin{pmatrix} \sigma^2 & 0 \\ 0 & \Psi \end{pmatrix},
  \]
  the condition (C4) is satisfied as we obtain $\sigma_0^2 \wedge \rho_{\min}(\Psi) \leq \rho_{\min}(\Delta_{\eta_0}) \leq \rho_{\max}(\Delta_{\eta_0}) \leq \|\Delta_{\eta_0}\|_F$ by Weyl’s inequality.

- **Verification of (C5'):** For a sufficiently large $M$ and $s_* = s_0 \vee (\log n/\log p)$, choose a sieve as
  \[
  \mathcal{H}_n = \{(\alpha, \beta, \mu) : |\alpha|^2 + \|\beta\|^2 + \|\mu\|^2 \leq n^2M \} \times \{\sigma : n^{-M} \leq \sigma^2 \leq e^{Ms_*, \log p}\}
  \times \{\Sigma : n^{-M} \leq \rho_{\min}(\Sigma) \leq \rho_{\max}(\Sigma) \leq e^{Ms_*, \log p}\}.
  \]
  Then we have $\rho_{\min}(\Delta_{\eta_0}) \geq \sigma_0^2 \wedge \rho_{\min}(\Psi) \geq n^{-M}$ for large $n$, and hence the minimum eigenvalue condition (2.6) is directly met with $\log \gamma_n \asymp \log n$ by the definition of the sieve. To see the entropy condition, observe from (2.56) that for every $\eta_1, \eta_2 \in \mathcal{H}_n$,
  \[
  d_n^2(\eta_1, \eta_2) \leq n^{4M} e^{2Ms_* \log p} \left(|\alpha - \alpha_0|^2 + \|\beta_1 - \beta_2\|^2 + \|\mu_1 - \mu_2\|^2 + \|\Sigma_1 - \Sigma_2\|^2 + |\sigma_1^2 - \sigma_2^2|\right).
  \]
  Therefore, for $\delta_n = 1/(6\min(3M+3/2) e^{Ms_* \log p})$, the entropy relative to $d_n$ is bounded above by
  \[
  \log N \left(\delta_n, \{(\alpha, \beta, \mu) : |\alpha|^2 + \|\beta\|^2 + \|\mu\|^2 \leq n^{2M}\}, \|\cdot\|_2\right)
  + \log N \left(\delta_n, \{\sigma : \sigma^2 \leq e^{Ms_* \log p}\}, \|\cdot\|_2\right)
  + \log N \left(\delta_n, \{\Sigma : \|\Sigma\|_F \leq \sqrt{q} e^{Ms_* \log p}\}, \|\cdot\|_F\right),
  \]
  each summand of which is bounded by a multiple of $\log n + s_* \log p$. This shows that the choice $\epsilon_n = \sqrt{(s_* \log p)/n}$ explicitly satisfies the entropy condition (2.7). Further, it is easy to see that the condition (2.8) holds using the tail bounds for normal and inverse Wishart distributions as in (2.53).

- **Verification of (C6*):** Note that the mean of $Y$ is expressed as $X\theta + Z\xi$ for $Z = 1_n \otimes I_{q+1}$. Since the condition $\rho_{\min}([X_S^T, 1_n]^T[X_S^T, 1_n]) \geq 1$ explicitly implies $\rho_{\min}([X, Z]^T[X, Z]) \geq 1$, the condition (C6*) is satisfied by Remark 2.1.

Therefore we obtain the contraction properties of the posterior distribution as in (2.10) by
Corollary 2.1. The rates for \( \eta \) with respect to more concrete metrics than \( d_n \) can now be obtained. Note that for small \( \delta > 0 \), \( d_n(\eta, \eta_0) \leq \delta \) directly implies \( \| \mu - \mu_0 \|_2 \leq \delta \) and \( \| \Sigma - \Sigma_0 \|_F \leq \delta \) by the definition of \( d_n \). For \( \beta \), observe that

\[
\| \beta - \beta_0 \|_2 \leq \| \Sigma^{-1} \|_{sp} \| \Sigma (\beta - \beta_0) \|_2 \leq \| \Sigma^{-1} \|_{sp} (\| \Sigma \beta - \Sigma_0 \beta_0 \|_2 + \| \Sigma - \Sigma_0 \|_F \| \beta_0 \|_2) \lesssim \| \Sigma^{-1} \|_{sp} \delta.
\]

Since \( \| \Sigma^{-1} \|_{sp} \) is bounded as \( \| \Sigma - \Sigma_0 \|_F \leq \delta \), the preceding display implies \( \| \beta - \beta_0 \|_2 \lesssim \delta \).

Moreover, we have

\[
|\alpha - \alpha_0| \leq |\mu^T \beta - \mu_0^T \beta_0| + \delta \\
\lesssim \| \mu \|_2 \| \beta - \beta_0 \|_2 + \| \beta_0 \|_2 \| \mu - \mu_0 \|_2 + \delta \\
\lesssim (\| \mu \|_2 + 1) \delta,
\]

\[
|\sigma^2 - \sigma^2_0| \leq |\beta^T \Sigma \beta - \beta_0^T \Sigma_0 \beta_0| + |(\beta^T \Sigma \beta + \sigma^2) - (\beta_0^T \Sigma_0 \beta_0 + \sigma_0^2)| \\
\leq \| \beta \|_2 \| \Sigma \beta - \Sigma_0 \beta_0 \|_2 + \| \beta_0 \|_2 \| \Sigma_0 \|_{sp} \| \beta - \beta_0 \|_2 + \delta \\
\lesssim (\| \beta \|_2 + 1) \delta,
\]

which shows that \( |\alpha - \alpha_0| + |\sigma^2 - \sigma^2_0| \lesssim \delta \) as \( \| \mu \|_2 \) and \( \| \beta \|_2 \) are bounded. We finally conclude that \( |\alpha - \alpha_0| + \| \beta - \beta_0 \|_2 + \| \mu - \mu_0 \|_2 + |\sigma^2 - \sigma_0^2| + \| \Sigma - \Sigma_0 \|_F \) contracts at the same rate of \( d_n \).

\[\square\]

**Proof of Theorem 2.8, (c)–(e).** We verify the conditions (C8)–(C12) to apply Theorems 2.4–2.5 and Corollaries 2.2–2.3. The orthogonal projection defined by \( H = \tilde{Z}(\tilde{Z}^T \tilde{Z})^{-1} \tilde{Z}^T \) with \( \tilde{Z} = 1_n \otimes \Delta_{\eta_0}^{-1/2} \) is used to check the conditions.

- **Verification of (C8):** For \( H \) defined above, it is easy to see that (C8) is satisfied.
- **Verification of (C9):** Choose a map \( (\alpha, \beta, \mu, \sigma^2, \Sigma) \mapsto (\alpha + n^{-1} i_n X^*(\theta - \theta_0), \beta, \mu, \sigma^2, \Sigma) \) for \( \eta \mapsto \tilde{\eta}_n(\theta, \eta) \). To check (C9), we shall verify that this map induces \( \Phi(\tilde{\eta}_n(\theta, \eta)) = (\tilde{\xi}_n + H \tilde{X}(\theta - \theta_0), \tilde{\Delta}_n) \) as follows. Using the properties of the Kronecker product that \( (R_1 \otimes R_2)(R_3 \otimes R_4) = (R_1 R_2 \otimes R_3 R_4) \) and \( (R_5 \otimes R_6)^{-1} = R_5^{-1} \otimes R_6^{-1} \) for matrices for which such operations are possible, we see that \( H \) satisfies

\[
H = (1_n \otimes \Delta_{\eta_0}^{-1/2})(1_n^T 1_n \otimes \Delta_{\eta_0}^{-1})(1_n \otimes \Delta_{\eta_0}^{-1/2})^T \\
= \frac{1}{n} (1_n \otimes \Delta_{\eta_0}^{-1/2}) \Delta_{\eta_0} (1_n \otimes \Delta_{\eta_0}^{-1/2})^T \\
= \frac{1}{n} (1_n \otimes I_{q+1}) (1_n^T \otimes I_{q+1}) \\
= \frac{1}{n} (1_n 1_n^T \otimes I_{q+1}).
\]
Hence,
\[
Z(\tilde{Z}^T \tilde{Z})^{-1} \tilde{Z}^T \tilde{X}(\theta - \theta_0) = (I_n \otimes \Delta_n^{1/2}) H (I_n \otimes \Delta_n^{-1/2}) X(\theta - \theta_0) \\
= H X(\theta - \theta_0) = 1_n \otimes \left( n^{-1} 1_n^T X^*(\theta - \theta_0) \right)_{0 \times 1},
\]
which implies that the shift only for \( \alpha \) as in the given map provides \( \Phi(\tilde{n}_n(\theta, \eta)) = (\tilde{\xi}_n + H \tilde{X}(\theta - \theta_0), \tilde{\Delta}_n) \). Now, using the standard normal prior for \( \alpha \) without loss of generality, observe that
\[
\left| \log \frac{d\Pi_{n, \theta}}{d\Pi_{n, \theta_0}} (\eta) \right| \lesssim |\alpha^2 - (\alpha + n^{-1} 1_n^T X^*(\theta - \theta_0))^2| \\
\leq 2|\alpha||n^{-1} 1_n^T X^*(\theta - \theta_0)| + (n^{-1} 1_n^T X^*(\theta - \theta_0))^2,
\]
since the priors for the other parameters cancel out due to invariance. Note that
\[
\sup_{\eta \in \bar{H}_n} |\alpha| \lesssim s_* \sqrt{(\log p)/n} + |\alpha_0| \lesssim 1,
\]
and
\[
\frac{1}{\sqrt{n}} \sup_{\theta \in \bar{\Theta}_n} \|X(\theta - \theta_0)\|_2 \lesssim s_* \sqrt{(\log p)/n},
\]
and thus the condition (C9) is satisfied.

- **Verification of (C10):** This is directly satisfied by Remark 2.2.
- **Verification of (C11):** Note that by (2.56), \( d_{B,n}(\eta_1, \eta_2) \lesssim \|\Sigma_1 - \Sigma_2\|_F + |\sigma_1^2 - \sigma_2^2| + \|\beta_1 - \beta_2\|_2 \)
for every \( \eta_1, \eta_2 \in \bar{H}_n \). Since each of the parameter spaces of \( \Sigma, \sigma^2, \) and \( \beta \) is a separable metric space with each of these norms, (C11) is satisfied.
- **Verification of (C12):** Note again that \( d_{B,n}(\eta, \eta_0) \lesssim \|\Sigma - \Sigma_0\|_F + |\sigma^2 - \sigma_0^2| + \|\beta - \beta_0\|_2 \)
for every \( \eta \in \bar{H}_n \). The inequality also holds for the other direction for every \( \eta \in \bar{H}_n \), by the same argument used for the recovery in the proof of Theorem 2.8, (a)–(b). Hence, for some constants \( C_1, C_2 > 0 \), the entropy in (C12) is bonded above by
\[
\log N \left( C_1 \delta, \left\{ \beta : \|\beta - \beta_0\|_2 \leq C_2 M_3 \epsilon_n^2 \right\}, \|\cdot\|_2 \right) + \log N \left( C_1 \delta, \left\{ \sigma^2 : |\sigma^2 - \sigma_0^2| \leq C_2 M_3 \epsilon_n^2 \right\}, |\cdot| \right) \\
+ \log N \left( C_1 \delta, \left\{ \Sigma : \|\Sigma - \Sigma_0\|_F \leq C_2 M_3 \epsilon_n^2 \right\}, \|\cdot\|_F \right).
\]
Since all nuisance parameters are of fixed dimensions, the last display is bonded by a multiple of \( 0 \vee \log(3C_2 M_3 \epsilon_n/C_1 \delta) \) for every \( \delta > 0 \), so that (C12) is bounded by \( (s_5^3 \log^3 p/n)^{1/2} = o(1) \) by Remark 2.3.
Therefore, under (C7), Theorem 2.4 implies that the distributional approximation in (2.14) holds. Under (C7) and (C13), we obtain the no-superset result in (2.15). The remaining assertions in the theorem are direct consequences of Corollary 2.2 and Corollary 2.3 if the beta-min condition (C14) is also satisfied.

We complete the proof by showing that the covariance matrix of the nonzero part can be written as in the theorem. For given $S$, we obtain

$$
\tilde{X}^T S (I_{n(q+1)} - H) \tilde{X} = X^s_S^T \left( I_n \otimes \{ \Delta^{-1/2}_{\eta_0} \}_{1,1} \right) \left( I_n \otimes I_{q+1} - H \right) \left( I_n \otimes \{ \Delta^{-1/2}_{\eta_0} \}_{1,1} \right) X^*_S
$$

$$
= \{ \Delta^{-1/2}_{\eta_0} \}_{1,1} X^s_S^T H^* X^*_S,
$$

where $\{ \Delta^{-1/2}_{\eta_0} \}_{1,1}$ is the first column of $\Delta^{-1/2}_{\eta_0}$. Note that $\{ \Delta^{-1/2}_{\eta_0} \}_{1,1} = \{ \Delta^{-1}_{\eta_0} \}_{1,1}$, where $\{ \Delta^{-1}_{\eta_0} \}_{1,1}$ is the top-left element of $\Delta^{-1}_{\eta_0}$, which is equal to $(\beta^T \Sigma_0 \beta_0 + \sigma^2 - \beta^T \Sigma_0 (\Sigma_0 + \Psi)^{-1} \Sigma_0 \beta_0)^{-1}$ by direct calculations. For the mean $\hat{\theta}_S$, observe that

$$
\tilde{X}^T S (I_{n(q+1)} - H)(U + \tilde{X} \theta_0) = X^s_S^T \left( I_n \otimes \{ \Delta^{-1/2}_{\eta_0} \}_{1,1} \right) \left( I_n \otimes I_{q+1} - \frac{1}{n} I_n^T \otimes I_{q+1} \right)
$$

$$
\times \left\{ \left( I_n \otimes \{ \Delta^{-1/2}_{\eta_0} \}_{1,1} \right) \left( Y^* - (\alpha_0 + \mu_0^T \beta_0) I_n \right) \right. \right.
$$

$$
\left. \left. + \left( I_n \otimes \{ \Delta^{-1/2}_{\eta_0} \}_{1,(-1)} \right) \left( \text{vec}(W^T) - I_n \otimes \mu_0 \right) \right\},
$$

where $\{ \Delta^{-1/2}_{\eta_0} \}_{1,(-1)}$ is the submatrix of $\Delta^{-1/2}_{\eta_0}$ consisting of columns except for $\{ \Delta^{-1/2}_{\eta_0} \}_{1,1}$ the first column. Since $\{ \Delta^{-1/2}_{\eta_0} \}_{1,(-1)} \{ \Delta^{-1/2}_{\eta_0} \}_{1,(-1)} = \{ \Delta^{-1}_{\eta_0} \}_{1,(-1)}$, where $\{ \Delta^{-1}_{\eta_0} \}_{1,(-1)}$ is the first row of $\Delta^{-1}_{\eta_0}$ without the top-left entry, the last display is equal to

$$
X^s_S^T \left\{ H^* \left[ \{ \Delta^{-1}_{\eta_0} \}_{1,1} \left( Y^* - (\alpha_0 + \mu_0^T \beta_0) I_n \right) + \left( I_n \otimes \{ \Delta^{-1}_{\eta_0} \}_{1,(-1)} \right) \left( \text{vec}(W^T) - I_n \otimes \mu_0 \right) \right] \right\}.
$$

As we have $\{ \Delta^{-1}_{\eta_0} \}_{1,(-1)} = -\{ \Delta^{-1}_{\eta_0} \}_{1,1} \beta^T \Sigma_0 (\Sigma_0 + \Psi)^{-1}$ by direct calculations, it follows that

$$
\hat{\theta}_S = (X^s_S^T H^* X^*_S)^{-1} X^s_S^T \left\{ H^* \left[ (Y^* - (\alpha_0 + \mu_0^T \beta_0) I_n) 
\right.ight.
$$

$$
\left. \left. - \left( I_n \otimes (\beta^T \Sigma_0 (\Sigma_0 + \Psi)^{-1}) \right) \left( \text{vec}(W^T) - I_n \otimes \mu_0 \right) \right] \right\}.
$$

This completes the proof. □

Proof of Theorem 2.9, (a)–(b). We shall verify the conditions for the optimal posterior contraction in Corollary 2.1. First we give the bounds for the eigenvalues of each correlation matrix.
It can be shown that
\[
1 - \alpha = \rho_{\min}(G^\text{CS}_i(\alpha)) \leq \rho_{\max}(G^\text{CS}_i(\alpha)) = 1 + (m_i - 1)\alpha, \tag{2.57}
\]
\[
1 - \frac{\alpha^2}{(1 + |\alpha|)^2} \leq \rho_{\min}(G^\text{AR}_i(\alpha)) \leq \rho_{\max}(G^\text{AR}_i(\alpha)) \leq \frac{1 - \alpha^2}{(1 - |\alpha|)^2}, \tag{2.58}
\]
\[
1 - 2|\alpha| \leq \rho_{\min}(G^\text{MA}_i(\alpha)) \leq \rho_{\max}(G^\text{MA}_i(\alpha)) \leq 1 + 2|\alpha|. \tag{2.59}
\]

The first assertion (2.57) follows directly from the identity \(\rho_k(G^\text{CS}_i(\alpha)) = \rho_k(\alpha_1 m_i \beta_{1 m_i}) + 1 - \alpha\) for every \(k \leq m_i\). For (2.58), see Theorem 2.1 and Theorem 3.5 of Fikioris (2018). The assertion (2.59) is due to Theorem 2.2 of Kulkarni et al. (1999).

- **Verification of (C1):** For the autoregressive correlation matrix, note that

\[
\max_{1 \leq i \leq n} \left\| \sigma^2 G^\text{AR}_i(\alpha) - \sigma_0^2 G^\text{AR}_i(\alpha_0) \right\|^2_F = \bar{m}(\sigma^2 - \sigma_0^2)^2 + 2 \sum_{k=1}^{m-1} (\bar{m} - k)(\sigma^2 \alpha_k^2 - \sigma_0^2 \alpha_0^2)^2.
\]

Using \(\bar{m} \approx n\), we have that

\[
\sum_{k=1}^{m-1} (\bar{m} - k)(\sigma^2 \alpha_k^2 - \sigma_0^2 \alpha_0^2)^2 \leq \frac{1}{n} \sum_{k=1}^{m-1} (\sigma^2 \alpha_k^2 - \sigma_0^2 \alpha_0^2)^2 \sum_{i=1}^{n} (m_i - k) \geq 0 \}
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{m_i-1} (m_i - k)(\sigma^2 \alpha_k^2 - \sigma_0^2 \alpha_0^2)^2,
\]

and hence

\[
\max_{1 \leq i \leq n} \left\| \sigma^2 G^\text{AR}_i(\alpha) - \sigma_0^2 G^\text{AR}_i(\alpha_0) \right\|^2_F \lesssim \frac{1}{n} \sum_{i=1}^{n} \left\| \sigma^2 G^\text{AR}_i(\alpha) - \sigma_0^2 G^\text{AR}_i(\alpha_0) \right\|^2_F.
\]

This gives us \(a_n \approx 1\) for the autoregressive matrices. Similarly, we can also show that \(a_n \approx 1\) satisfies (C1) for the compound-symmetric and the moving average correlation matrices. Also, we have \(e_n = 0\) for (C1) as the true parameter values \(\alpha_0\) and \(\sigma_0^2\) are in the support of the prior.

- **Verification of (C2):** Since the nuisance parameters are of fixed dimensions, the condition (C2) is satisfied with \(e_n = \sqrt{\log n}/n\) due to the restricted range of the true parameters, \(\sigma_0^2 \approx 1\) and \(\alpha_0 \in [b_1 + \varepsilon, b_2 - \varepsilon]\) for some fixed \(\varepsilon > 0\).

- **Verification of (C3):** The assumption \(\lambda \|\theta_0\|_1 \lesssim s_\ast \log p\) given in the theorem directly satisfies (C3).

- **Verification of (C4):** Using (2.57)–(2.59), we see that for the compound-symmetric correlation matrix, the condition (C4) is satisfied with the bounded range of the true parameters provided
that $\overline{m}$ is bounded. For the other correlation matrices, the condition (C4) is satisfied even with increasing $\overline{m}$.

- Verification of (C5*): For a sufficiently large $M > 0$ and $s_* = s_0 \vee (\log n / \log p)$, choose a sieve $\mathcal{H}_n = \{\sigma^2 : n^{-M} \leq \sigma^2 \leq e^{M s_* \log p}\} \times \{\alpha : b_1 + n^{-M} \leq \alpha \leq b_2 - n^{-M}\}$. Then using (2.57)–(2.59), it is easy to see that the minimum eigenvalue of each correlation matrix is bounded below by a polynomial in $n$, which implies that the condition (2.6) is satisfied with $\log \gamma_n \propto \log n$. For the entropy calculation, note that for every type of correlation matrix,

$$d_n^2(\eta_1, \eta_2) = \frac{1}{n} \sum_{i=1}^{n} \|\sigma^2_i G_i(\alpha_1) - \sigma^2_i G_i(\alpha_2)\|_F^2$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \left\{ (\sigma^2_1 - \sigma^2_2)^2 \|G_i(\alpha_1)\|_F^2 + \sigma^2_i \|G_i(\alpha_1) - G_i(\alpha_2)\|_F^2 \right\}.$$  (2.60)

From the identity $\alpha^k - \alpha^k_2 = (\alpha_1 - \alpha_2) \sum_{j=0}^{k-1} \alpha_j \alpha_2^{k-1-j}$ for every integer $k \geq 1$, we have that $|\alpha_1^k - \alpha_2^k| \lesssim k|\alpha_1 - \alpha_2|$ for every $\alpha_1, \alpha_2 \in (b_1, b_2)$. By this inequality we obtain $\|G_i(\alpha_1) - G_i(\alpha_2)\|_F^2 \lesssim \overline{m}^4 |\alpha_1 - \alpha_2|^2$ for every correlation matrix. Then, the last display is bounded by a multiple of $\overline{m}^2(\sigma^2_1 - \sigma^2_2)^2 + e^{2b_* \log p} \overline{m}^4 |\alpha_1 - \alpha_2|^2$ for every $\eta_1, \eta_2 \in \mathcal{H}_n$. The entropy in (2.7) is thus bounded by

$$\log N(\delta_n, \{\sigma^2 : 0 < \sigma^2 \leq e^{M s_* \log p}\}, | \cdot |) + \log N(\delta_n, \{\alpha : 0 < \alpha < 1\}, | \cdot |),$$

for $\delta_n = (6\overline{m}^3 n^{3/2} + C_1 e^{M s_* \log p})^{-1}$ with some constant $C_1 > 0$. It can be easily checked that each term in the last display is bounded by a multiple of $s_* \log p$, by which the entropy condition (2.7) is satisfied with $\epsilon_n = \sqrt{s_* \log p} / \overline{m}$. Using the tail bounds of inverse gamma distributions and properties of the density $\Pi(da)$ near the boundaries, the condition (2.8) is satisfied as long as $M$ is chosen sufficiently large.

- Verification of (C6*): The separation condition is trivially satisfied as there is no nuisance mean part.

Therefore, we obtain the posterior contraction properties of $\theta$ with $s_* = s_0 \vee (\log n / \log p)$ by Corollary 2.1. Since we have $m_i(\sigma^2 - \sigma^2_0)^2 \leq \|\sigma^2 G_i(\alpha) - \sigma^2 G_i(\alpha_0)\|_F^2$ by the diagonal entries of each matrix, the posterior contraction rate $\sqrt{(s_* \log p) / \overline{m} n}$ for $\sigma^2$ with respect to the $\ell_2$-norm is obtained for every correlation matrix, as $\overline{m} n \asymp n_*$. In particular, for the compound-symmetric correlation matrix, this rate is reduced to $\sqrt{(s_* \log p) / n}$ since $\overline{m}$ is bounded in that case. We also have $m_i(\sigma^2 \alpha - \sigma^2_0 \alpha_0)^2 \lesssim \|\sigma^2 G_i(\alpha) - \sigma^2 G_i(\alpha_0)\|_F^2$ for every correlation matrix, as there are more than $m_i$ entries that is equal to $\sigma^2 \alpha - \sigma^2_0 \alpha_0$. Hence, by the relation $|\alpha - \alpha_0| \leq |\sigma^2 \alpha - \sigma^2_0 \alpha_0| + |\alpha| \|\sigma^2 - \sigma^2_0\|_F$, the same rate is also obtained for $\alpha$ relative to the $\ell_2$-norm. □
Proof of Theorem 2.9, (c)–(e). We verify the conditions (C8)–(C12) to apply Theorems 2.4–2.5 and Corollaries 2.2–2.3.

- **Verification of (C8)–(C10):** These conditions are trivially satisfied with the zero matrix $H$ since there is no nuisance mean part.
- **Verification of (C11):** Using (2.60), we have $d_{B,n}(\eta_1, \eta_2) \lesssim m|\sigma_1^2 - \sigma_2^2| + m^2|\alpha_1 - \alpha_2|$ for every $\eta_1, \eta_2 \in \mathcal{H}_n$. Since the parameter spaces of $\alpha$ and $\sigma^2$ are Euclidean and hence separable under the $\ell_2$-metric, the condition (C11) is satisfied.
- **Verification of (C12):** Using the results of contraction rates of $\sigma^2$ and $\alpha$, note that there exists a constant $C_1 > 0$ such that

$$\{\eta \in \mathcal{H}: d_{B,n}(\eta, \eta_0) \leq M_3\epsilon_n/\sqrt{m}\} \subset \{\sigma^2: |\sigma^2 - \sigma_0^2| \leq C_1\epsilon_n/\sqrt{m}\} \times \{\alpha: |\alpha - \alpha_0| \leq C_1\epsilon_n/\sqrt{m}\}.$$  

By Remark 2.3, (C12) is bounded by a multiple of $\{(s_5^5 \log^3 p)/n\}^{1/2}$, which goes to zero by the assumption.

Therefore, under (C7), the distributional approximation in (2.14) holds with the zero matrix $H$ by Theorem 2.4. Under (C7) and (C13), Theorem 2.5 implies that the no-superset result in (2.15) holds. The strong results in Corollary 2.2 and Corollary 2.3 follow explicitly from the beta-min condition (C14).

Proof of Theorem 2.10, (a)–(b). We verify the conditions for the optimal posterior contraction in Corollary 2.1.

- **Verification of (C1):** Using the assumption $\max_i \|Z_i\|_{sp} \lesssim 1$, note that

$$\max_{1 \leq i \leq n} \|Z_i(\Psi - \Psi_0)Z_i^T\|_F^2 \leq \|\Psi - \Psi_0\|_F^2 \max_{1 \leq i \leq n} \|Z_i\|_{sp}^4 \lesssim \frac{1}{n} \sum_{i=1}^n \|Z_i(\Psi - \Psi_0)Z_i^T\|_F^2 \|Z_i^T Z_i\|_F^{-1} \|Z_i^T Z_i\|_{sp}^4 \lesssim \frac{1}{n} \sum_{i=1}^n \|Z_i(\Psi - \Psi_0)Z_i^T\|_F^2,$$

where the last inequality holds since $\min_i \{\rho_{\min}(Z_i^T Z_i) : m_i \geq q\} \gtrsim 1$ and $\sum_{i=1}^n 1(m_i \geq q) \asymp n$. Thus we have $a_n \asymp 1$ and $\epsilon_n = 0$.

- **Verification of (C2):** The condition is satisfied with $\bar{c}_n = \sqrt{(\log n)/n}$ as $\Psi$ is fixed dimensional and we have $1 \lesssim \rho_{\min}(\Psi_0) \leq \rho_{\max}(\Psi_0) \lesssim 1$.

- **Verification of (C3):** The assumption $\lambda \|\theta_0\|_1 \lesssim s_* \log p$ given in the theorem directly satisfies (C3).
• **Verification of (C4):** By Weyl’s inequality, we obtain that

\[
\begin{align*}
\min_{1 \leq i \leq n} \rho_{\min}(\sigma^2 I_m + Z_i \Psi_0 Z_i^T) & \geq \sigma^2 + \min_{1 \leq i \leq n} \rho_{\min}(Z_i \Psi_0 Z_i^T), \\
\max_{1 \leq i \leq n} \rho_{\max}(\sigma^2 I_m + Z_i \Psi_0 Z_i^T) & \leq \sigma^2 + \rho_{\max}(\Psi_0) \max_{1 \leq i \leq n} \|Z_i\|_{sp}^2.
\end{align*}
\] (2.62) (2.63)

Since \(Z_i \Psi_0 Z_i^T\) is nonnegative definite, the right hand side of (2.62) further bounded below by \(\sigma^2\), while the right hand side of (2.63) is bounded. The condition (C4) is thus satisfied.

• **Verification of (C5†):** For a sufficiently large \(M\) and \(s_\star = s_0 \vee (\log n/\log p)\), define a sieve as \(\mathcal{H}_n = \{\Psi : n^{-M} \leq \rho_{\min}(\Sigma) \leq \rho_{\max}(\Sigma) \leq e^{Ms_\star \log p}\}\), so that the minimum eigenvalue condition (2.6) can be satisfied with \(\log \gamma_n \approx \log n\). Similar to the proof of Theorem 2.6, (a)–(b), it can be easily shown that the conditions (2.7) and (2.8) are satisfied with \(\epsilon_n = \sqrt{(s_\star \log p)/n}\).

• **Verification of (C6∗):** The separation condition is trivially satisfied as there is no nuisance mean part.

Therefore, the posterior contraction rates for \(\theta\) are given by Corollary 2.1. The contraction rate for \(\Sigma\) relative to the Frobenius norm is a direct consequence of (2.61).

**Proof of Theorem 2.10, (c)–(e).** We verify the conditions (C8)–(C12) to apply Theorems 2.4–2.5 and Corollaries 2.2–2.3.

• **Verification of (C8)–(C10):** These conditions are trivially satisfied with the zero matrix \(H\) since there is no nuisance mean part.

• **Verification of (C11):** It is easy to see that \(d_{B,n}(\eta, \eta_0) \lesssim \|\Psi - \Psi_0\|_F\) since \(\max_i \|Z_i\|_{sp} \lesssim 1\). The separability of the space is thus trivial.

• **Verification of (C12):** For some \(C_1 > 0\), the entropy in (C12) is bounded above by a multiple of \(\log N(\delta, \{\Sigma : \|\Sigma - \Sigma_0\|_F^2 \leq M_3 C_1 \epsilon_n^2\}, \|\cdot\|_F) \lesssim 0 \vee \log(3\sqrt{M_3 C_1 \epsilon_n/\delta})\) by (2.61). The expression in (C12) is thus bounded by a constant multiple of \(s_\star^5 \log^3 p = o(1)\) by Remark 2.3.

Hence, under (C7), Theorem 2.4 can be applied to obtain the distributional approximation in (2.14) with the zero matrix \(H\). Under (C7) and (C13), we obtain the no-superset result in (2.15) by Theorem 2.5. The strong results in Corollary 2.2 and Corollary 2.3 follow explicitly from the beta-min condition (C14).

**Proof of Theorem 2.11, (a)–(b).** We verify the conditions for the optimal posterior contraction in Corollary 2.1.

• **Verification of (C1):** Since \(\Delta_{n,i} = \Omega^{-1}\) for every \(i \leq n\) and \(\Omega_0 \in M_0^+(cL)\) for some \(0 < c < 1\), \(a_n = 1\) and \(\epsilon_n = 0\) satisfy (C1).
• **Verification of (C2):** Using (i) of Lemma 2.6 and the relation \(1-x \times 1-x^{-1}\) as \(x \to 1\), observe that \(\|\Omega^{-1} - \Omega_0^{-1}\|_F \lesssim \|\Omega - \Omega_0\|_F \lesssim \epsilon_n\) if the right hand side is small enough. Thus, there exists a constant \(C_1 > 0\) such that \(\{\Omega : \|\Omega^{-1} - \Omega_0^{-1}\|_F \leq \epsilon_n\} \supseteq \{\Omega : \|\Omega - \Omega_0\|_F \leq C_1\epsilon_n\}\). Furthermore, although the components of \(\Omega\) are not a priori independent as the prior is truncated to \(\mathcal{M}_0^+ (L)\), the truncation can only increase prior concentration since \(\Omega_0 \in \mathcal{M}_0^+ (cL)\) for some \(0 < c < 1\). Hence, for some \(C_2 > 0\),

\[
\Pi \left(\|\Omega^{-1} - \Omega_0^{-1}\|_F \leq \epsilon_n\right) \geq \Pi \left(\|\Omega - \Omega_0\|_\infty \leq C_2\epsilon_n/m\right) \gtrsim \left(\frac{C_2\epsilon_n}{m}\right)^{m+d},
\]

which justifies to choose \(\epsilon_n \sim \sqrt{(m+d)(\log n)/n}\) for (C2).

• **Verification of (C3):** The assumption \(\lambda_0 \|\theta_0\|_1 \lesssim s_* \log p\) given in the theorem directly satisfies (C3).

• **Verification of (C4):** This is trivially met as \(\Omega_0 \in \mathcal{M}_0^+ (cL)\) for some \(0 < c < 1\).

• **Verification of (C5\textsuperscript{:}):** Note that the minimum eigenvalue condition (2.6) is trivially satisfied with \(\gamma_n = 1\) since the prior is put on \(\mathcal{M}_0^+ (L)\). Now, for \(\tilde{r}_n = M s_* \log p / \log n\) with \(s_* = s_0 \vee (m^2_n / \log p)\) and sufficiently large \(M\), choose a sieve as \(\mathcal{H}_n = \{\Omega \in \mathcal{M}_0^+ (L) : \sum_{j,k} 1\{\omega_{jk} \neq 0\} \leq \tilde{r}_n\}\), that is, the maximum number of edges of \(\Omega\) does not exceed \(\tilde{r}_n\). Then, for \(\delta_n = 1/6\tilde{r}_n m^{3/2}\), the entropy in (2.7) is bounded by

\[
\log N(\delta_n / \tilde{r}_n, \mathcal{H}_n, \|\cdot\|_\infty) \leq \log \left\{\left(\frac{mL}{\delta_n}\right)^{m+\tilde{r}_n} \left(\frac{m}{\tilde{r}_n}\right)\right\} \leq (m + \tilde{r}_n) \log (mL/\delta_n) + 2\tilde{r}_n \log \tilde{r}_n,
\]

where in the second term, the factor \((mL/\delta_n)^m\) comes from the diagonal elements of \(\Omega\), while the rest is from the off-diagonal entries. It is easy to see that the last display is bounded by a multiple of \(s_* \log p\) with chosen \(\tilde{r}_n\), and hence the entropy condition (2.7) is satisfied. Lastly, note that for some \(C_3 > 0\),

\[
\log \Pi(\mathcal{H} \setminus \mathcal{H}_n) = \log \Pi(|\mathcal{Y}| > \tilde{r}_n) \lesssim -\tilde{r}_n \log \tilde{r}_n \leq -C_3 M s_* \log p.
\]

Therefore the condition (2.8) is satisfied with sufficiently large \(M\).

• **Verification of (C6\textsuperscript{+}):** The separation condition is trivially met as there is no nuisance mean part.

Hence, we obtain the posterior contraction properties for \(\theta\) by Corollary 2.1. The corollary also implies that the posterior distribution of \(\Omega^{-1}\) contracts to \(\Omega_0^{-1}\) at the rate \(\epsilon_n = \sqrt{(s_0 \log p \vee (m+d) \log n)/n}\) with respect to the Frobenius norm. This is also translated as
convergence of $\Omega$ to $\Omega_0$ at the same rate, since we obtain

$$
\|\Omega - \Omega_0\|_F^2 \lesssim \|\Omega^{-1} - \Omega_0^{-1}\|_F^2 \lesssim \epsilon_n^2, \quad (2.64)
$$

by (i) of Lemma 2.6 and the inequality $1 - x \asymp 1 - x^{-1}$ as $x \to 1$.

**Proof of Theorem 2.11, (c)–(e).** We verify the conditions (C8)–(C12) to apply Theorems 2.4–2.5 and Corollaries 2.2–2.3.

- **Verification of (C8)–(C10):** These conditions are trivially satisfied with the zero matrix $H$ since there is no nuisance mean part.
- **Verification of (C11):** For every $\Omega_1, \Omega_2 \in \hat{\mathcal{H}}_n$, note that
  
  $$
  \|\Omega_1^{-1} - \Omega_2^{-1}\|_F \lesssim \|\Omega_1 - \Omega_2\|_F \lesssim \|\Omega_1^{-1} - \Omega_2^{-1}\|_F \lesssim \epsilon_n, \quad (2.65)
  $$

  using (i) of Lemma 2.6 and the inequality $1 - x \asymp 1 - x^{-1}$ as $x \to 1$ again. By the first inequality, it suffices to show that $\mathcal{H}$ is separable metric space with the Frobenius norm. This is trivial as the parameter space is Euclidean.

- **Verification of (C12):** Note that by (2.64), there exists a constant $C_1 > 0$ such that the entropy in (C12) is bounded by $\log N(\delta, \{\Omega : \|\Omega - \Omega_0\|_F \leq C_1 \epsilon_n\}, d_{B,n})$ for every $\delta > 0$. Using (2.65), the entropy is further bounded by $\log N(C_2 \delta, \{\Omega : \|\Omega - \Omega_0\|_F \leq C_1 \epsilon_n\}, \|\cdot\|_F)$ for some $C_2 > 0$. This is clearly bounded by a multiple of $0 \lor m^2 \log(3C_1 \epsilon_n/C_2 \delta)$, and hence using Remark 2.3 we bound (C12) by a multiple of $(s \lor m) \sqrt{(s \lor \log p)^3/n}$ which goes to zero by assumption.

Therefore, under (C7), we obtain the distributional approximation in (2.14) with the zero matrix $H$ by Theorem 2.4. Under (C7) and (C13), the no-superset result in (2.15) holds by Theorem 2.5. Lastly, we obtain the strong results in Corollary 2.2 and Corollary 2.3 if the beta-min condition (C14) is also met.

**Proof of Theorem 2.12, (a)–(b).** We verify the conditions for the optimal posterior contraction in Corollary 2.1.

- **Verification of (C1):** Since $\Delta_{n,i} = \sigma^2$ for every $i \leq n$ and $\sigma_0^2$ belongs to the support of the prior, we have $a_n = 1$ and $e_n = 0$.

- **Verification of (C2):** Note that we write $d_n^2(\eta, \eta_0) = |\sigma^2 - \sigma_0^2|^2 + \|g_\beta - g_0\|_{2,n}^2$. To verify the prior concentration condition, observe that

$$
\log \Pi(\eta \in \mathcal{H} : d_n(\eta, \eta_0) \leq \epsilon_n) \\
\geq \log \Pi\left(\beta : \|g_\beta - g_0\|_{2,n} \leq \frac{\epsilon_n}{\sqrt{2}}\right) + \log \Pi\left(\sigma : |\sigma^2 - \sigma_0^2| \leq \frac{\epsilon_n}{\sqrt{2}}\right),
$$
where the second term on the right hand side is trivially bounded below by a constant multiple of $\log \beta_n$. To bound the first term, note that there exists $\beta_0 \in \mathbb{R}^J$ with $\|\beta_0\|_\infty < \|g_0\|_\infty$ such that

$$
\|\beta^*_T B - g_0\|_\infty \lesssim J^{-\alpha}\|g_0\|_\infty, \quad (2.66)
$$

by the well-known approximation theory of B-splines (De Boor, 1978, page 170), which gives

$$
\|g_\beta - g_0\|_{2,n} \leq \|g_\beta - g_0\|_\infty \lesssim J^{-\alpha} + \|g_\beta - g_\beta^\ast\|_\infty, \quad (2.67)
$$

since $\|\beta_\ast\|_\infty$ is bounded. Furthermore, we use the following properties of the B-splines: for every $\beta \in \mathbb{R}^J$,

$$
\|\beta\|_\infty \lesssim \|g_\beta\|_\infty \lesssim \|\beta\|_\infty, \quad \|\beta\|_2 \lesssim \sqrt{J}\|g_\beta\|_{2,n} \lesssim \|\beta\|_2, \quad (2.68)
$$

where the second assertion requires the regressors $z_1, \ldots, z_n$ to be sufficiently distributed on the interval $[0, 1]$. For the proof of the first assertion, see Lemma E.6 of Ghosal and van der Vaart (2017). Hence, if $J^{-\alpha} \lesssim \bar{\epsilon}_n$, then for some $C_1, C_2 > 0$,

$$
\log \Pi(\beta : \|g_\beta - g_0\|_{2,n} \leq C_1\bar{\epsilon}_n) \geq \log \Pi(\beta : \|\beta - \beta_\ast\|_\infty \leq C_2\bar{\epsilon}_n) \gtrsim J\log \bar{\epsilon}_n,
$$

by the first relation of (2.68). By matching $J^{-\alpha} \lesssim \bar{\epsilon}_n$ and $J\log n \lesssim n\epsilon_n^2$ to satisfy (C2), we choose $J = J_n \asymp (n/\log n)^{1/(2\alpha+1)}$, and hence $\bar{\epsilon}_n = n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)}$.

- **Verification of (C3):** The assumption $\lambda\|\theta_0\|_1 \lesssim s_\ast \log p$ given in the theorem directly satisfies (C3).
- **Verification of (C4):** This is directly satisfied by $\sigma_0^2 \asymp 1$.
- **Verification of (C5)\textsuperscript{1}:** For a sufficiently large constant $M$ and $s_\ast = s_0 \vee (J\log n/\log p)$, choose $\mathcal{H}_n = \{g_\beta : \|\beta\|_\infty \leq nM\} \times \{\sigma : n^{-M} \leq \sigma^2 \leq e^{Ms_\ast \log p}\}$, from which the minimum eigenvalue condition (2.6) is directly satisfied with $\log \gamma_n = \log n$. To see the entropy condition (2.7), note that for every $\eta_1, \eta_2 \in \mathcal{H}_n$, we have $d_n^2(\eta_1, \eta_2) \lesssim \|\beta_1 - \beta_2\|_\infty^2 + |\sigma_1^2 - \sigma_2^2|^2$ by (2.68). Hence, for some $C_3 > 0$, the entropy in (2.7) is bounded above by a multiple of

$$
\log N \left( \frac{1}{C_3mn^{M+3/2}}, \{\beta : \|\beta\|_\infty \leq nM, \|\cdot\|_\infty\} \right) + \log N \left( \frac{1}{C_3mn^{M+3/2}}, \{\sigma : \sigma^2 \leq e^{Ms_\ast \log p}, \|\cdot\|\} \right).
$$

The display is further bounded by a multiple of $J\log n + s_\ast \log p$, and hence (2.7) is satisfied with $\epsilon_n = \sqrt{(s_\ast \log p)/n}$. Using the tail bounds of normal and inverse gamma distributions,
the condition (2.8) is also satisfied.

- **Verification of (C6*)**: The separation condition holds by Remark 2.1 as we have \( d_{A,n}(\eta_*, \eta_0) = \|g_{\beta_*} - g_0\|_{2,n} \lesssim \bar{\epsilon}_n \) for \( \eta_* = (g_{\beta_*}, \sigma_0^2) \) in view of (2.66).

Therefore, the contraction rates for \( \theta \) are given by Corollary 2.1. The rate for \( g \) is also obtained by the corollary.

\[ \text{Proof of Theorem 2.12, (c)--(e).} \] We first show why \( J \) cannot be used for \( J^\star \). Observe that the left side of (C8) is equal to

\[ s_*^2 \log p \| \tilde{\xi}_{r_0} - H \tilde{\xi}_{r_0} \|_2^2 = \frac{ns_*^2 \log p}{\sigma_0^2} \| g_0 - \tilde{\beta}^T B J \|_{2,n}^2, \]

where \( \tilde{\beta} = (W_*^TW_*)^{-1}W_*(g_0(z_1), \ldots, g_0(z_n))^T \) is the least squares solution. Since \( \tilde{\beta} \) is the solution minimizing \( \| g_0 - \tilde{\beta}^T B J \|_{2,n}^2 \), for some \( \beta_* \in \mathbb{R}^J \), the last display is bounded above by

\[ \frac{ns_*^2 \log p}{\sigma_0^2} \| g_0 - \beta_*^T B J \|_{2,n}^2 \lesssim \frac{ns_*^2 \log p}{J^{2\alpha}}, \]

by (2.66). If \( J_* \) is chosen equal to \( J \) used for the optimal contraction, then we have \( s_* = s_0 \vee (J \log n / \log p) \) and (C8) is not satisfied as the rightmost side of the last display does not tend to zero. Thus a suboptimal contraction rate is unavoidable for the assertion to hold.

Note that we should verify the conditions (C2), (C5†), and (C6*) again, while showing that \( J_* \) chosen in the theorem satisfies (C8). The conditions (C9)--(C12) are also verified below.

- **Verification of (C2)**: By examining the proof for (a)--(b), it can be seen that the prior concentration condition (C2) requires \( J_* \log n \lesssim n\bar{\epsilon}_n^2 \) and \( J_*^{-\alpha} \lesssim \bar{\epsilon}_n \). For the former inequality to hold, choose

\[ \bar{\epsilon}_n = (\log n)^{1/2} \left( \frac{M_n s_0^2 \log p}{n^{2\alpha - 1}} \right)^{1/(4\alpha)} \vee (\log n)^{(\alpha/2)(\alpha - 1)} \left( \frac{M_n}{n^{2\alpha - 3} \log p} \right)^{1/(4(\alpha - 1))}. \]

The latter inequality also holds as \( J_*^{-\alpha} \lesssim (M_n s_0^2 \log p)^{-1/2} < n^{-1/2} \).

- **Verification of (C5†)**: Choose a sieve \( \mathcal{H}_n \) as in the proof for (a)--(b) with

\[ s_* = s_0 \vee (n\bar{\epsilon}_n^2 / \log p) = s_0 \vee (\log n) \left( \frac{M_n s_0^2 \log p}{(\log p)^{2\alpha - 1}} \right)^{1/2\alpha} \vee \left( \frac{(\log n)^{2\alpha} M_n n}{(\log p)^{2\alpha - 1}} \right)^{1/(2(\alpha - 1))}. \]

Then the minimum eigenvalue condition (2.6) is directly satisfied with \( \log \gamma_n \asymp \log n \) as before. Since the entropy is bounded by \( J_* \log n + s_* \log p \), the condition (2.7) is also satisfied. The sieve condition (2.8) is also satisfied as \( \log J_* = o(n) \).
• **Verification of (C6’):** This is trivially satisfied by Remark 2.1 as in the proof for (a)–(b).

• **Verification of (C8):** It is easy to see that \( J_* \) chosen in the theorem makes the right hand side of (2.69) tend to zero.

• **Verification of (C9):** For a given \( \theta \), let \( \tilde{\eta}_n(\theta, \eta) = (g_\beta(\cdot) + B_{J_*}^t(\cdot)(W_*^T W_*)^{-1}W_* X(\theta - \theta_0), \sigma^2) \), where \( \eta = (g_\beta(\cdot), \sigma^2) \), which directly satisfies \( \Phi(\tilde{\eta}_n(\theta, \eta)) = (\tilde{\xi}_n + HX(\theta - \theta_0), \tilde{\Delta}_n) \). Since each entry of \( \beta \) has the standard normal prior, \( g_\beta(\cdot) \) is a zero mean Gaussian process with the covariance kernel \( K(t_1, t_2) = B_{J_*}(t_1)^T B_{J_*}(t_2) \), and thus its reproducing kernel Hilbert space (RKHS) \( \mathcal{K} \) is the set of all functions of the form \( \sum_k \zeta_k B_{J_*}(t_k)^T B_{J_*}(\cdot) \) with coefficients \( \zeta_k, k \in \{1, 2, \ldots\} \). It is easy to see that the shift \( (\theta - \theta_0)^T X^T W_*(W_*^T W_*)^{-1}B_{J_*}(\cdot) \) is in the RKHS \( \mathcal{K} \) since it is expressed as \( (\theta - \theta_0)^T X^T W_*(W_*^T W_*)^{-1}W_* B_{J_*}(\cdot) \) using an invertible matrix \( W_* \in \mathbb{R}^{J_* \times J_*} \) with rows \( B_{J_*}(t_k) \) evaluated by some \( t_k, k = 1, \ldots, J_* \). Hence, by the Cameron-Martin theorem, for \( \nu = (\nu_1, \ldots, \nu_{J_*})^T = W_*^{-1}(W_*^T W_*)^{-1}W_* X(\theta - \theta_0) \) and \( \|\cdot\|_\mathcal{K} \) the RKHS norm, we see that

\[
\log \frac{d\Pi_{n, \theta}}{d\Pi_{n, \theta_0}}(\eta) = \sum_{k=1}^{J_*} \nu_k g_\beta(t_k) - \frac{1}{2}\|\nu^T \tilde{W}_* B_{J_*}\|_\mathcal{K}^2
= \nu^T \tilde{W}_* \beta - \frac{1}{2}\|\tilde{W}_* \nu\|_2^2,
\]

almost surely. This gives that

\[
\left|\log \frac{d\Pi_{n, \theta}}{d\Pi_{n, \theta_0}}(\eta)\right| \lesssim \|\beta\|_2\|\nu^T (W_*^T W_*)^{-1}W_* X(\theta - \theta_0)\|_2 + \|\nu^T (W_*^T W_*)^{-1}W_* X(\theta - \theta_0)\|_2.
\]

Note that we have

\[
\sup_{\eta \in \mathcal{H}_n} \|\beta\|_2 \leq \sup_{\eta \in \mathcal{H}_n} \|\beta - \beta_*\|_2 + \|\beta_*\|_2 \lesssim \sqrt{J_*} \sup_{\eta \in \mathcal{H}_n} \|g_\beta - g_{\beta_*}\|_{2, n} + 1 \lesssim s_* \sqrt{\log p / n} + 1,
\]

and

\[
\sup_{\theta \in \Theta_n} \|\nu^T (W_* X(\theta - \theta_0))\|_2 \lesssim \frac{\|W_*\|_{sp} \sup_{\theta \in \Theta_n} \|X(\theta - \theta_0)\|_2}{\rho_{\min}(W_*^T W_*)} \lesssim s_* \sqrt{\log p / J_* n},
\]

and thus (C9) is satisfied since \( s_* \sqrt{\log p / n} \rightarrow 0 \).

• **Verification of (C10):** This condition is met by Remark 2.2.

• **Verification of (C11):** Since we have \( d_{B,n}(\eta_1, \eta_2) = |\sigma_1^2 - \sigma_2^2| \) for every \( \sigma_1^2, \sigma_2^2 \in (0, \infty) \) and the parameter space of \( \sigma^2 \) is Euclidean, the condition is trivially satisfied.

• **Verification of (C12):** Since the entropy in (C12) is bounded above by a multiple of \( 0 \lor \log(3\sqrt{M_1} \epsilon_n / \delta) \) for every \( \delta > 0 \), (C12) is bounded by \( (s_*^5 \log^3 p / n)^{1/2} \).

Therefore, under (C7), we have the distributional approximation in (2.14) by Theorem 2.4. Under (C7) and (C13), Theorem 2.5 implies that the no-superset result in (2.15) holds. The
stronger assertions in (2.16) and (2.17) are explicitly derived from Corollary 2.2 and Corollary 2.3 if the beta-min condition (C14) is also met.

Here we elaborate on how \( J_* \) is chosen. To make the right hand side of (2.69) tend to zero, for an arbitrary \( M_n \to \infty \), we need \( J_*^{2\alpha} \gtrsim M_n s_n^2 \log p \geq M_n (n s_n^2 \log p / n^3 \epsilon_\alpha^4 / \log p) \), from which we obtain \( J_* \gtrsim (M_n n s_0^2 \log p)^{1/2\alpha} \). By examining the proof for (a)–(b), it can be seen that the prior concentration condition (C2) requires \( J_* \log n \lesssim n \epsilon_n^2 \) and \( J_*^{\alpha} \lesssim \bar{\epsilon}_n \). By matching \( J_* \log n \lesssim n \epsilon_n^2 \) and \( M_n n^3 \epsilon_\alpha^4 / \log p \lesssim J_*^{2\alpha} \), we also obtain \( J_* \gtrsim \{M_n n (\log n)^2 / \log p\}^{1/2(\alpha-1)} \). Thus, we obtain \( J_* \) and \( \bar{\epsilon}_n \) above. It is easy to see that this also satisfies \( J_*^{\alpha} \lesssim \bar{\epsilon}_n \) since \( J_*^{\alpha} \lesssim (M_n n s_0^2 \log p)^{-1/2} < n^{-1/2} \).

**Proof of Theorem 2.13, (a)–(b).** We verify the conditions for the optimal posterior contraction in Corollary 2.1.

- **Verification of (C1):** If \( v_0 \) is strictly positive on \([0, 1]\), then \( v_0 \) satisfy the same approximation rule in (2.66) for some \( \beta_* \in (0, \infty)^J \) with \( \|\beta_*\|_\infty < \|v_0\|_{\mathcal{C}^\alpha} \) (see Lemma E.5 of Ghosal and van der Vaart (2017)). Therefore the approximation in (2.67) also holds for \( v_0 \) even if \( \beta \) is restricted to have positive entries only, and thus by (2.66) and (2.68),

\[
\|v_{\beta_*} - v_0\|_\infty \lesssim J^{-\alpha}, \quad \text{for some } \beta_* \in (0, \infty)^J, \\
\|v_{\beta_1} - v_{\beta_2}\|_\infty \lesssim \sqrt{J}\|v_{\beta_1} - v_{\beta_2}\|_{2, n}, \quad \beta_1, \beta_2 \in (0, \infty)^J,
\]

which tells us that we have \( a_n \asymp J \) and \( e_n \asymp J^{1-2\alpha} \) for (C1).

- **Verification of (C2):** Note that if \( J^{-\alpha} \lesssim \bar{\epsilon}_n \), it follows that for some \( C_1 > 0 \),

\[
\log \Pi(\beta : \|v_{\beta} - v_0\|_{2, n} \leq \epsilon_n) \geq \log \Pi(\beta : \|\beta - \beta_*\|_\infty \leq C_1 \epsilon_n) \gtrsim J \log \epsilon_n.
\]

This implies that the condition (C2) is satisfied with \( \epsilon_n = \sqrt{(J \log n)/n} \).

- **Verification of (C3):** The assumption \( \lambda \|\theta_0\|_1 \lesssim s_* \log p \) given in the theorem directly satisfies (C3).

- **Verification of (C4):** Since \( v_0 \) is strictly positive on \([0, 1]\) and belongs to a fixed multiple of the unit ball of \( \mathcal{C}^\alpha[0, 1] \), we have that

\[
1 \lesssim \inf_{z \in [0, 1]} v_0(z) \leq \sup_{z \in [0, 1]} v_0(z) \lesssim 1.
\]

The condition (C4) is thus satisfied.

- **Verification of (C5†):** For a sufficiently large \( M \), choose a sieve as \( \mathcal{H}_n = \prod_{j=1}^J \{\beta_j : n^{-M} \leq \beta_j \leq n^M\} \). Then the minimum eigenvalue condition (2.6) is satisfied with \( \log \gamma_n \asymp \log n \).
because for every $i \leq n$, 

$$\inf_{\beta \in \mathcal{H}_n} v_\beta(z_i) = \inf_{\beta \in \mathcal{H}_n} \sum_{j=1}^J B_{j,j}(z_i) \beta_j \geq \inf_{\beta \in \mathcal{H}_n, 1 \leq j \leq J} \beta_j \sum_{j=1}^J B_{j,j}(z_i) \geq n^{-M},$$

where $B_{j,j}$ and $\beta_j$ denote the $j$th components of $B_j$ and $\beta$, respectively. Next, it is easy to see that the entropy condition (2.7) is met with $\epsilon_n = \sqrt{(s_* \log p)/n}$, similar to the case of the partial linear models. Moreover, the condition (2.8) holds since an inverse Gaussian prior on each $\beta_j$ produces $\Pi(H_\mathcal{H}_n) \lesssim J \epsilon_n$ for some constant $C_2$, by its exponentially small bounds for tail probabilities on both sides. By matching $J^{-\alpha} \asymp \bar{\epsilon}_n$ and $n \bar{\epsilon}_n \asymp J \log n$, we obtain $J \asymp (n/\log n)^{1/(2\alpha+1)}$ and $\bar{\epsilon}_n = (\log n/n)^{\alpha/(2\alpha+1)}$. Note that the conditions $a_n \epsilon_n^2 \to 0$ and $e_n \to 0$ hold only if $\alpha > 1/2$.

- **Verification of (C6’):** The separation condition holds as there is no additional mean part.

Hence, we obtain the posterior contraction properties for $\theta$ by Corollary 2.1. The contraction rate for $v$ is also obtained by the same corollary.

**Proof of Theorem 2.13, (c)–(e).** We verify the conditions (C8)–(C12) to apply Theorems 2.4–2.5 and Corollaries 2.2–2.3.

- **Verification of (C8)–(C10):** These conditions are trivially satisfied as there is no nuisance mean part.

- **Verification of (C11):** For every $v_{\beta_1}, v_{\beta_2} \in \mathcal{H}_n$, note that $d_{B,n}(\eta_1, \eta_2) = \|v_{\beta_1} - v_{\beta_2}\|_{2,n} \lesssim \|\beta_1 - \beta_2\|_2$ by (2.68). Since we put a prior for $v$ using the B-splines through a Euclidean parameter $\beta$, the separability is trivially satisfied.

- **Verification of (C12):** Note that by the inequality $\|v_\beta - v_0\|_{2,n}^2 \lesssim \|v_\beta - v_{\beta_*}\|_{2,n}^2 + \epsilon_n^2$, the entropy in the integrand is bounded by

$$\log N \left( \delta \sqrt{J}, \{\beta : \|\beta - \beta_*\|_2^2 \leq C_1 J \epsilon_n^2\}, \|\cdot\|_2 \right) \lesssim 0 \vee J \log \left( \frac{3\sqrt{C_1 \epsilon_n}}{\delta} \right),$$

for some $C_1 > 0$. Thus, (C12) is bounded above by a multiple of $\{(s_*^2 \vee J)J(s_* \log p)^3/n\}^{1/2}$ by Remark 2.3, and hence goes to zero by the assumption. The condition $\alpha > 2$ is seen to be necessary by the inequality $(s_*^2 \vee J)J(s_* \log p)^3/n \geq J^2 n^2 \epsilon_n^6 = n^{2(-\alpha+2)/(2\alpha+1)} \log n^{2(3\alpha-1)/(2\alpha+1)}$.

Under (C7), the distributional approximation in (2.14) holds with the zero matrix $H$ by Theorem 2.4. Under (C7) and (C13), the no-superset result in (2.15) holds by Theorem 2.5. We also obtain the strong results in Corollary 2.2 and Corollary 2.3 if the beta-min condition (C14) is also met.
2.5.3 Auxiliary results

Here we provide some auxiliary results used to prove the main results in this chapter.

**Lemma 2.5.** Let \( p_k \) be the density of \( N_r(\mu_k, \Sigma_k) \) for \( k = 1, 2 \). Then,

\[
K(p_1, p_2) = \frac{1}{2} \left\{ \log \frac{\det(\Sigma_2)}{\det(\Sigma_1)} + \text{tr}(\Sigma_1 \Sigma_2^{-1}) - r + \|\Sigma_2^{-1/2} (\mu_1 - \mu_2)\|^2 \right\},
\]

\[
V(p_1, p_2) = \frac{1}{2} \left\{ \text{tr}(\Sigma_1 \Sigma_2^{-1} \Sigma_1 \Sigma_2^{-1}) - 2 \text{tr}(\Sigma_1 \Sigma_2^{-1}) + r \right\} + \|\Sigma_1^{1/2} \Sigma_2^{-1} (\mu_1 - \mu_2)\|^2.
\]

**Proof.** Let \( Z = \Sigma_1^{-1/2}(X - \mu_1) \sim N_r(0, I) \) for \( X \sim p_1 \) and \( A = \Sigma_1^{1/2} \Sigma_2^{-1} \Sigma_1^{1/2} \). Then by direct calculations, we have

\[
K(p_1, p_2) = \mathbb{E}_{p_1} \left\{ \log \frac{p_1}{p_2}(X) \right\} = \frac{1}{2} \left\{ \log \frac{\det \Sigma_2}{\det \Sigma_1} + \mathbb{E}_{p_1} Z^T AZ - r + (\mu_1 - \mu_2)^T \Sigma_2^{-1} (\mu_1 - \mu_2) \right\},
\]

which verifies the first assertion. After some algebra, we also obtain

\[
V(p_1, p_2) = \mathbb{E}_{p_1} \left\{ \left( \log \frac{p_1}{p_2}(X) - K(p_1, p_2) \right)^2 \right\} = \frac{1}{4} \mathbb{E}_{p_1} \left\{ -Z^T Z + Z^T AZ + 2(\mu_1 - \mu_2)^T \Sigma_2^{-1} \Sigma_1^{1/2} Z - \text{tr}(A) + r \right\}^2.
\]

The rightmost side involves forms of \( \mathbb{E}_{p_1} (Z^T Q_1 Z Z^T Q_2 Z) \) for two positive definite matrices \( Q_1 \) and \( Q_2 \). It can be shown that this equals \( 2 \text{tr}(Q_1 Q_2) + \text{tr}(Q_1) \text{tr}(Q_2) \); for example, see Lemma 6.2 of Magnus (1978). Plugging in this for the expected values of the products of quadratic forms, it is easy (but tedious) to verify the second assertion. \qed

**Lemma 2.6.** For \( r \times r \) positive definite matrices \( \Sigma_1 \) and \( \Sigma_2 \), let \( d_1, \ldots, d_r \) be the eigenvalues of \( \Sigma_1^{1/2} \Sigma_1^{-1} \Sigma_2^{1/2} \). Then the following assertions hold:

(i) \( \rho_{\max}^2(\Sigma_2) \|\Sigma_1 - \Sigma_2\|_F^2 \leq \sum_{k=1}^r (d^{-1}_k - 1)^2 \leq \rho_{\min}^{-2}(\Sigma_2) \|\Sigma_1 - \Sigma_2\|_F^2 \),

(ii) \( \max_k |d_k - 1| \) can be made arbitrarily small if \( g^2(\Sigma_1, \Sigma_2) \) is chosen sufficiently small, where \( g \) is defined in (2.30).

**Proof.** Let \( A = \Sigma_2^{-1/2} \Sigma_1 \Sigma_2^{-1/2} \). Since the eigenvalues of \( A - I_r \) are \( d_1^{-1} - 1, \ldots, d_r^{-1} - 1 \), we have that

\[
\|\Sigma_1 - \Sigma_2\|_F^2 = \|\Sigma_2^{-1/2} (A - I_r) \Sigma_2^{-1/2}\|_F^2 \leq \rho_{\max}^2(\Sigma_2) \|A - I_r\|_F^2 = \rho_{\max}^2(\Sigma_2) \sum_{k=1}^r (d^{-1}_k - 1)^2.
\]

67
Conversely, using the sub-multiplicative property of the Frobenius norm \( \|BC\|_F \leq \|B\|_{sp} \|C\|_F \),
\[
\sum_{k=1}^{r} (d_k^{-1} - 1)^2 = \|A - I_r\|_F^2 = \|\Sigma_1^{-1/2}(\Sigma_1 - \Sigma_2)\Sigma_2^{-1/2}\|_F^2 \leq \rho_{\max}^2(\Sigma_2^{-1})\|\Sigma_1 - \Sigma_2\|_F^2.
\]
These verify (i). Now, note that by direct calculations,
\[
\frac{(\det \Sigma_1)^{1/4}(\det \Sigma_2)^{1/4}}{\det((\Sigma_1 + \Sigma_2)/2)^{1/2}} = \left\{ \frac{1}{2^r} \det(A^{1/2} + A^{-1/2}) \right\}^{-1/2} = \left\{ \prod_{k=1}^{r} \frac{1}{2} (d_k^{1/2} + d_k^{-1/2}) \right\}^{-1/2}.
\]
Hence, \( g^2(\Sigma_1, \Sigma_2) < \delta \) for a sufficiently small \( \delta > 0 \) implies that
\[
\prod_{k=1}^{r} \frac{1}{2} (d_k^{1/2} + d_k^{-1/2}) < (1 - \delta^2/2)^{-2}.
\]
Since every term in the product of the last display is greater than or equal to 1, we have \( (d_k^{1/2} + d_k^{-1/2})/2 < (1 - \delta^2/2)^{-2} \) for every \( k \). As a function of \( d_k \), \( (d_k^{1/2} + d_k^{-1/2})/2 \) has the global minimum at \( d_k = 1 \), and hence \( \delta \) can be chosen sufficiently small to make \( |d_k - 1| \) small for every \( k = 1, \ldots, r \), which establishes (ii). \( \square \)
Chapter 3

Bayesian sparse regression for categorical responses

3.1 Introduction

High-dimensional sparse regression has received a lot of attention and its theoretical investigation has recently been one of the most interesting topics in the Bayesian community. Most theoretical studies on posterior distributions for high-dimensional regression models are directed to continuous response variables (e.g., Castillo et al., 2015; Gao et al., 2015; Martin et al., 2017; Song and Liang, 2017; Belitser and Ghosal, 2019; Ročková, 2018). While these are the archetypes, models for discrete response variables are also of great interest and very useful in many fields of applications. Multi-categorical (nominal) response variables are especially noteworthy in that they are needed to account for outcomes with more than two attributes, which are very common in many areas such as classification and decision sciences. Examples of models for such response variables include multinomial logit models, multinomial probit models, multi-class support vector machines, etc. However, despite their usefulness in the high-dimensional setting, theoretical properties of Bayesian methods have not been studied in any of these models.

Amongst others, the logit models are the most popular as they enjoy immediate interpretations of log odds. In this chapter, we investigate high-dimensional logit models for categorical response variables $Z_i$, $i = 1, \ldots, n$, with $m$-categories, under the Bayesian framework. Let the response variable be coded as $Z_i \in \{0, 1, \ldots, m - 1\}$. For the total number $d^*$ of parameters, let $X_i \in \mathbb{R}^{(m-1) \times d^*}$ be a design matrix of the $i$th observation and $\beta \in \mathbb{R}^{d^*}$ be a vector of high-dimensional parameters. Then a general logit model for the categorical response $Z_i$ can be
written as

$$\log \left( \frac{\mathbb{P}_{\beta}(Z_i = \ell)}{\mathbb{P}_{\beta}(Z_i = 0)} \right) = X_i^T \beta, \quad \ell = 1, \ldots, m - 1, \quad i = 1, \ldots, n, \quad (3.1)$$

where $X_i(\ell)$ is the $\ell$th row of $X_i$ and $\mathbb{P}_{\beta}$ is the probability operator with $\beta$. The covariate $X_i(\ell)$ quantifies the characteristics of the category ‘$\ell$’ against the reference category ‘0’. It is obvious that the model involves a high-dimensional logistic regression model for binary response variables. The form (3.1) is general in that the covariate can vary with choices, but it is often assumed that the covariate is not category-specific in many settings. We present the following two examples to elaborate more on this.

Example 3.1 (Variable selection in multinomial logit models). The standard multinomial logit models take a simpler form for the right hand side of (3.1), say $\tilde{X}_i^T \tilde{\beta}_\ell$, for some covariates $\tilde{X}_i \in \mathbb{R}^p$ and parameters $\tilde{\beta}_\ell \in \mathbb{R}^p$ with $p > 0$. That is, the covariate of the $i$th observation is not choice-specific but common for all the categories. The model still belongs to (3.1) by writing $X_i = I_{m - 1} \otimes \tilde{X}_i^T$ and $\beta = (\tilde{\beta}_1^T, \ldots, \tilde{\beta}_{m - 1}^T)^T$. We write $d^* = p(m - 1)$ to match the expression with the form (3.1). In this situation, it makes more sense for the parameters linked to the same covariate to be included or excluded together, which should be handled by group-level sparsity.

Example 3.2 (Conditional logit models). The logit models sometimes take the general form itself in (3.1). This model is often called the conditional logit model (McFadden, 1973), which is particularly useful in many observational studies and decision making sciences where choice-specific data are available. We refer to Hoffman and Duncan (1988) for more discussion. For this framework, individual-level sparsity is a natural treatment for inference in the high-dimensional settings, but group-level sparsity may still be of interest depending on data and research questions.

In view of Example 3.1, group sparse modeling is extremely useful for the form (3.1) and is actually necessary from a practical point of view. We study frequentist asymptotic properties of the posterior distribution for the regression model in (3.1) under group sparsity. With a lasso penalty, the idea of group sparse estimation was considered for linear models in Yuan and Lin (2006) and extended to logistic regression in Meier et al. (2008). Even from the frequentist perspective, theoretical studies on high-dimensional group sparse estimation are mostly directed to the linear models (Nardi and Rinaldo, 2008; Huang and Zhang, 2010; Lounici et al., 2011), and extensions are rather sparse (see Blazère et al. (2014) for some breakthroughs in the generalized linear model setting). In the Bayesian framework, the topic has only recently been studied even for linear regression models; see Ning et al. (2018) and Bai et al. (2019) for some brand-new developments. Note that a generalized linear model for scalar response variables may not be
enough for our purpose. Theoretical properties of the posterior distribution for the model (3.1) have not been thoroughly examined and this study is the first investigation. We are primarily interested in the high-dimensional setting that $p > n$, where $p$ is the number of groups. Note that $p \leq d^*$. If $p = d^*$, the model is reduced to an individual sparse regression model. Our main goal is to examine nearly optimal posterior contraction rates and model selection consistency. Guaranteed frequentist coverage of credible sets will also be considered.

A thorough search of the literature reveals that Atchadé (2017) and Wei and Ghosal (2019) are the only available Bayesian works on posterior contraction for high-dimensional logistic regression for binary responses. Atchadé (2017) obtained a posterior contraction rate with respect to the $\ell_2$-metric with a Laplace prior distribution, by constructing a test relative to the corresponding metric under a more general setting for certain class of nonlinear regression models. Wei and Ghosal (2019) also used the testing approach developed in Atchadé (2017) to obtain a comparable contraction rate using continuous shrinkage priors. However, posterior contraction in high-dimensional logit models for multi-categorical responses has not been studied anywhere. Moreover, even for simple logistic regression with binary responses, there is no available results on further theoretical properties of the posterior distribution beyond contraction rates, e.g., selection consistency. This chapter goes far beyond Atchadé (2017) and Wei and Ghosal (2019) in this sense. We mention that the technique developed in Atchadé (2017) cannot be directly used to establish selection consistency through an approximation to the posterior distribution due to the spinode of the prior distribution at zero. By putting a mild restriction on the magnitude of the true coefficients, we loosen the condition on the rate parameter for the prior distribution so that the sharpness of the prior can be asymptotically ignorable. Then model selection consistency can be verified through a shape approximation. The approximation also enables to quantify the remaining uncertainty on the parameter via the posterior distribution. To the best of our knowledge, this study is the first examination of selection consistency and a correct quantification of uncertainty in high-dimensional logit models even for binary responses, from the Bayesian point of view.

To sum up, we generalize the available results in the literature in several directions as follows. None of these extensions appear to be straightforward.

- Multi-categorical response variables are considered as well as binary responses. This generalization is very useful in many regression and classification problems, and should gain a great attention from both theoretical and practical points of view.

- Group-level sparsity is adopted, which clearly incorporates the individual-level sparse modeling. As mentioned above, group sparse modeling is extremely useful in our setting for categorical response variables. The size of each group can be different across groups, so our modeling structure is more flexible than Example 3.1. The size of each group can also vary
with the sample size \( n \), by which the final rate is affected by its order.

- Under a mild restriction on the true value of the linear predictor, we obtain faster posterior contraction rates than those in Atchadé (2017) and Wei and Ghosal (2019) at the level of compatibility coefficients. This can be done by deriving a contraction rate with respect to the root-average-squared Hellinger metric with an exponentially powerful test.

- Posterior contraction rates without any assumption on the true linear predictor are also available, though they are slower. In contrast to Atchadé (2017) and Wei and Ghosal (2019) who focused on the \( \ell_2 \)-contraction rate, we also study an \( \ell_{2,1} \)-rate through a construction of a suitable test function, which is translated into an \( \ell_1 \)-rate for the ungrouped situation. Unlike Remark 5 of Atchadé (2017), our \( \ell_1 \)-rate depends on an \( \ell_1 \)-type compatibility condition, rather than the \( \ell_2 \)-compatibility number used in Atchadé (2017) and Wei and Ghosal (2019). In this sense, our \( \ell_1 \)-rate improves these findings with a weaker compatibility condition.

- Model selection consistency is also verified through a shape approximation to the posterior distribution. As posterior consistency does not necessarily mean that the support of the true model is correctly recovered, selection consistency is an interesting object in the context of high-dimensional analysis.

- The shape approximation and selection consistency give rise to a Bernstein-von Mises theorem. The result enables to correctly quantify remaining uncertainty on the parameter through the posterior distribution.

The rest of this chapter is organized as follows. Section 3.2 introduces the notations used in the chapter and the prior distribution for the high-dimensional regression coefficients. Section 3.3 provides the main results on posterior contraction of the high-dimensional logit model for categorical responses under group sparsity. In Section 3.4, a shape approximation to the posterior is characterized, and selection consistency and uncertainty quantification are studied based on it. All the technical details are deferred to Section 3.5.

### 3.2 Setup and prior specification

#### 3.2.1 Notation

Here we describe the notations used throughout the chapter. We first let \( X = (X_1^T, \ldots, X_n^T)^T \in \mathbb{R}^{n(m-1) \times d^*} \) denote the design matrix formed by stacking up all \( X_i, i = 1, \ldots, n \). We assume that \( G_1, \ldots, G_p \) are \( p \) disjoint subsets of \( \{1, \ldots, d^*\} \) representing \( p \) groups such that \( \bigcup_{j=1}^p G_j = \{1, \ldots, d^*\} \). We write \( q_j \) for the cardinality of each subset \( G_j \), i.e., \( q_j = |G_j| \), and write \( \bar{q} := \max_j q_j \). For given \( G_1, \ldots, G_p \), let \( \beta_j \in \mathbb{R}^{q_j}, j = 1, \ldots, p \), be disjoint subvectors of \( \beta \in \mathbb{R}^{d^*} \).
elements of which are chosen by \( G_j \). Similarly, we also define \( X_j \in \mathbb{R}^{n(m-1) \times q_j} \), \( j = 1, \ldots, p \), disjoint submatrices of \( X \in \mathbb{R}^{n(m-1) \times d^*} \), where columns of each \( X_j \) are chosen by \( G_j \).

For a vector \( \beta \) and a set \( S \subset \{1, \ldots, p\} \) of group indices, we write \( \beta_S = \{ \beta_j, j \in S \} \) and \( \beta_{S^c} = \{ \beta_j, j \notin S \} \) to separate \( \beta \) into zero and nonzero coefficients using \( S \). We also write \( S_\beta = \{ j : \beta_j \neq 0_q \} \subset \{1, \ldots, p\} \) for the effective group index determined by \( \beta \). The cardinalities of \( S \) and \( S_\beta \) are denoted by \( s \) := \( |S| \) and \( s_\beta := |S_\beta| \). In particular, the group index of the true parameter \( \beta_0 \) and its cardinality are written as \( S_0 \) and \( s_0 \), respectively. We also let \( d_S := \sum_{j \in S} q_j \) denote the dimension of \( \beta_S \), and write \( d_0 := d_{S_0} \). Similar to \( \beta_S \), we define \( X_S \in \mathbb{R}^{n(m-1) \times d_S} \) whose columns are chosen by all \( G_j, j \in S \).

Let \( \| \cdot \|_q \) denote the \( \ell_q \)-norm and \( \| \cdot \|_\infty \) denote the max-norm of vectors. For a \( d^* \)-dimensional vector \( \beta \), we write \( \| \beta \|_{2,1} = \sum_{j=1}^p \| \beta_j \|_2 \) for the \( \ell_{2,1} \)-norm, which is typically used in the context of group sparsity. For any matrix, let \( \| \cdot \|_{sp} \) denote the spectral norm. The notations \( \rho_{\min}(\cdot) \) and \( \rho_{\max}(\cdot) \) denote the minimum and maximum eigenvalues of a square matrix, respectively. Also for a matrix \( X \) with \( d^* \) columns, we define a matrix norm: \( \| X \|_0 = \max_{1 \leq j \leq p} \| X_j \|_{sp} \). This is a natural generalization of the norm \( \| X \|_* \), the square root of the maximum diagonal entry of \( X^T X \), used for the individual-level sparse inference in the literature (e.g., Castillo et al., 2015; Belitser and Ghosal, 2019), taking into account for the group variable selection scheme in this study (note that our \( \| \cdot \|_0 \)-norm becomes identical to \( \| \cdot \|_* \) if \( \bar{q} = 1 \)).

We define the multinomial response variable \( Y_{it} := 1 \) \((Z_i = t), i = 1, \ldots, n, \ell = 1, \ldots, m-1, \) such that for any \( i \), \( \sum_{t=1}^{m-1} Y_{it} = 1 \) if \( Z_i > 0 \) and \( \sum_{t=1}^{m-1} Y_{it} = 0 \) otherwise. Hereafter, we work with the response variable \( Y(n) := (Y_1, \ldots, Y_n)^T \) in place of \( Z_i \), \( i = 1, \ldots, n, \) where \( Y_i = (Y_{i1}, \ldots, Y_{im-1})^T \). Let \( f_{\beta,i} \) and \( f_{0,i} \) be densities of \( Y_i \) with parameters \( \beta \) and \( \beta_0 \), respectively. We write joint densities as \( f_{\beta} = \prod_{i=1}^n f_{\beta,i} \) and \( f_0 = \prod_{i=1}^n f_{0,i} \), and the likelihood ratio as \( \Lambda_n(\beta) = (f_{\beta}/f_0)(Y(n)) \). The root-average-squared Hellinger metric between joint densities \( f_{\beta} \) and \( f_0 \) is denoted by \( H_n(\beta, \beta_0) = \left\{ (1 - \int S_{n\beta}^{0i} f_{\beta,i} f_{0,i})^{1/2} \right\}^{1/2} \) with the Hellinger distance \( H(f_{\beta,i}, f_{0,i}) = \{ (\sqrt{f_{\beta,i}} - \sqrt{f_{0,i}})^2 d\nu \}^{1/2} \) for a dominating counting measure \( \nu \). The notations \( \mathbb{P}_\beta \) and \( \mathbb{E}_\beta \) denote the probability and expectation operators with a parameter \( \beta \). In particular, those with the true parameter \( \beta_0 \) are abbreviated by \( \mathbb{P}_0 \) and \( \mathbb{E}_0 \). The subscripts are often suppressed in general settings.

The total variation metric between two probability measures \( P \) and \( Q \) is written as \( \| P - Q \|_{TV} \). For a semimetric space \((\mathcal{F}, d)\), let \( N(\epsilon, \mathcal{F}, d) \) denote the \( \epsilon \)-covering number, the minimal number of \( \epsilon \)-balls that is needed to cover \( \mathcal{F} \). For sequences \( a_n \) and \( b_n \), \( a_n \lesssim b_n \) (or \( b_n \gtrsim a_n \)) stands for \( a_n \leq C b_n \) for some constant \( C > 0 \) independent of \( n \). This inequality is also used for relations involving constant sequences.

Define the function \( h : \mathbb{R}^{m-1} \mapsto (0, \infty)^{m-1} \) such that \( h(x) = (e^{x_1}, \ldots, e^{x_{m-1}})^T \) for any \( x = (x_1, \ldots, x_{m-1})^T \in \mathbb{R}^{m-1} \). For the expected value \( \mu_i := \mathbb{E}_0(Y_i) = (1 + h(X_i \beta_0)^T \mathbb{I}_{m-1})^{-1} h(X_i \beta_0) \) of \( Y_i \) under the true model, let \( W_i = \mathrm{diag}(\mu_i) - \mu_i \mu_i^T \) be the covariance matrix of \( Y_i \). The
covariance of the whole observation vector $Y^{(n)}$ is then written as

$$W = \begin{pmatrix}
W_1 & \cdots & 0_{(m-1)\times(m-1)} \\
\vdots & \ddots & \vdots \\
0_{(m-1)\times(m-1)} & \cdots & W_n
\end{pmatrix}. $$

Using this, we define the following two compatibility numbers:

$$\psi_{2,1}(s) = \inf_{\beta: 1 \leq |S_\beta| \leq s} \frac{\|W^{1/2}X\beta\|_2 S_\beta|^{1/2}}{\|X\|_0 \|\beta\|_{2,1}}, \quad \psi_2(s) = \inf_{\beta: 1 \leq |S_\beta| \leq s} \frac{\|W^{1/2}X\beta\|_2}{\|X\|_0 \|\beta\|_2}, $$

which will be used for the basic $\ell_{2,1}$- and $\ell_2$-contraction rates. Stronger versions of these compatibility numbers are also considered; the uniform compatibility number and the smallest scaled singular value are defined by

$$\phi_{2,1}(s) = \inf_{\beta: 1 \leq |S_\beta| \leq s} \frac{\|X\beta\|_2 S_\beta|^{1/2}}{\|X\|_0 \|\beta\|_{2,1}}, \quad \phi_2(s) = \inf_{\beta: 1 \leq |S_\beta| \leq s} \frac{\|X\beta\|_2}{\|X\|_0 \|\beta\|_2}, $$

respectively. These stronger compatibility conditions are particularly useful for interpretation when $\|X\beta_0\|_\infty$ is bounded, in which case we have $\phi_{2,1}(s) \lesssim \psi_{2,1}(s)$ and $\phi_2(s) \lesssim \psi_2(s)$ for every $s \geq 1$, as the minimum eigenvalue of $W$ is bonded away from zero in such a case (see Lemma 3.6).

Note that unlike Castillo et al. (2015) for individual sparse linear regression, where the compatibility numbers are scaled by $\|X\|_s$, our definitions are scaled by $\|X\|_0$. This is designed to incorporate our group sparse modeling for compatibility. (Recall that the norm $\|X\|_0$ is identical to $\|X\|_s$ if the individual-level sparsity is imposed.) Observe that we have $\|X\|_0 \leq \sqrt{n}\|X\|_s$. The term $\|X\|_s$ is expected to increase at the order $\sqrt{n}$ in typical regression settings, and hence $\|X\|_0$ is also expected to do so at most at the order $\sqrt{qn}$. Therefore, under common assumptions on the design matrix $X$, the compatibility numbers $\phi_{2,1}$ and $\phi_2$ are likely to be bounded away from zero for models of moderate dimension unless $q$ increases extremely fast (we refer to Example 7 of Castillo et al. (2015) for a more detailed discussion).

### 3.2.2 Prior specification

A prior distribution should be carefully designed to obtain a good posterior contraction rate. Similar to Castillo et al. (2015) for individual sparse linear regression, we first select a group dimension $s$ from a prior distribution $\pi_p(s)$, and then a group index set $S \subset \{1, \ldots, p\}$ is randomly chosen for given $s$. A vector of nonzero coefficients $\beta_S$ is then selected from a continuous prior $g_S$ on $R^{ds}$ while $\beta_{Sc}$ is set to zero. The resulting prior distribution for $(S, \beta)$ is summarized
as

$$(S, \beta) \mapsto \frac{\pi_p(s)}{\binom{p}{s}} g_S(\beta_S)\delta_0(\beta_{S^c}),$$  \hspace{1cm} (3.2)$$

where $\delta_0$ is the Dirac measure at zero on $\mathbb{R}^{d^* - d_S}$ (with suppressed dimensionality). It remains to specify $\pi_p$ and $g_S$. For the prior $\pi_p$ on a group size, we consider a prior distribution such that for some constants $A_1, A_2, A_3, A_4 > 0$,

$$\frac{A_1}{(p \vee n\overline{q}) A_3} \leq \frac{\pi_p(s)}{\pi_p(s - 1)} \leq \frac{A_2}{(p \vee n\overline{q}) A_4}, \hspace{1cm} s = 1, \ldots, p.$$  \hspace{1cm} (3.3)$$

This prior distribution is modified from Castillo et al. (2015) and designed to be tailored for optimal asymptotic properties in group sparse estimation. If $\overline{q} = 1$, that is, if the sparsity is imposed only at the individual level, then the prior distribution in (3.3) becomes identical to the one used in the literature (Castillo et al., 2015; Martin et al., 2017; Belitser and Ghosal, 2019). For a prior $g_S$ on the nonzero coefficients, we consider the following $s$-fold product of continuous distributions,

$$g_S(\beta_S) = \prod_{j \in S} \left( \frac{\lambda_j}{a_j} \right)^{q_j} \exp(-\lambda_j \parallel \beta_j \parallel_2), \hspace{1cm} a_j = \sqrt{\pi} \left( \frac{\Gamma(q_j + 1)}{\Gamma(q_j/2 + 1)} \right)^{1/q_j},$$  \hspace{1cm} (3.4)$$

for regularization hyperparameters $\lambda_j$, $j = 1, \ldots, p$. See Lemma 2.1 of Ning et al. (2018) for the derivation of the expression for $a_j$. The prior distribution in (3.4), just as the Laplace prior corresponds to the lasso, mimics the group lasso penalty through the $\ell_{2,1}$-norm. However, one should note that sparsity in our modeling is directly handled by the point mass part in (3.2). The regularization parameters $\lambda_j$, $j = 1, \ldots, p$, can differ across groups, so that they can effectively account for unequal group size as the group lasso. We allow $\lambda_j$ to vary with the sample size $n$ such that for constants $L_1, L_2, L_3 > 0$,

$$\frac{\parallel X \parallel_o}{L_1(p^{1/\overline{q}} \sqrt{n})^{L_2}} \leq \lambda \leq \frac{\parallel X \parallel_o}{L_3 \sqrt{n}},$$  \hspace{1cm} (3.5)$$

where $\underline{\lambda} = \min_j \lambda_j$ and $\overline{\lambda} = \max_j \lambda_j$. A particularly interesting case is that $\lambda_j$ is set to the order of the lower bound for every $j$. Coupled with the condition (C1) below, this makes the bound requirement on the true signal very mild. We shall see that the upper bound also gives a reasonable restriction. See (C1) below and the following paragraph for more details.
3.3 Posterior contraction rates

With the prior distribution specified in Section 3.2.2, the posterior distribution $\Pi(\cdot | Y^{(n)})$ of $\beta$ is defined by Bayes’ rule as

$$
\Pi(\beta \in B | Y^{(n)}) = \frac{\int_B \Lambda_n(\beta) d\Pi(\beta)}{\int_{\mathbb{R}^d} \Lambda_n(\beta) d\Pi(\beta)}, \text{ for any measurable } B \subset \mathbb{R}^d.
$$

In this section, we study contraction properties of the posterior distribution under appropriate assumptions on the true coefficients $\beta_0$ and the design matrix $X$. For simplicity, all the results in this chapter are provided in asymptotic forms. As mentioned in Section 3.1, we first assume that the true parameter belongs to certain norm-bounded subset of the parameter space as follows.

(C1) The true parameter $\beta_0$ satisfies

$$
\lambda \| \beta_0 \|_2, 1 \lesssim s_0 (\log p \lor q \log n).
$$

Note that if the sparsity is imposed at the individual level, i.e., $q = 1$, the condition is reduced to $\lambda \| \beta_0 \|_1 \lesssim s_0 \log p$ using the $\ell_1$-norm of $\beta_0$. A sufficient condition for (C1) which gives a clearer interpretation for the magnitude of $\beta_0$ is

$$
\| \beta_0 \|_\infty \lesssim \frac{\log p \lor q \log n}{\lambda \sqrt{q}}. \tag{3.6}
$$

In particular, the lowest allowed value of $\lambda$ in (3.5) makes this restriction extremely mild with a suitably chosen $L_2$. Even with the highest allowed value of $\lambda$, the condition is not too restrictive if $\|X\|_0$ increases at a moderate order. For example, it is likely that $\|X\|_0 \lesssim \sqrt{qn}$ in typical regression settings, which makes the right hand side of (3.6) at least $\log n$, whereby the condition is still reasonable as it grows with $n$.

We should mention that unlike our condition in (C1), Atchadé (2017) does not require any size restriction on $\beta_0$. We believe that a similar approach may be employed for our multinomial logit model with group sparsity, but as in Atchadé (2017), it should require a ‘large’ value of $\lambda_j$ to restrict the sample space to some good subspace. However, later in this chapter, selection consistency will be verified through a distributional approximation with normal mixtures, which is hampered by the cusp at zero induced by this ‘large’ rate parameter. We will see that $\lambda$ that does not increase too fast is indeed necessary for the distributional approximation. The condition (C1) allows us to consider a ‘small’ $\lambda_j$ in our range in (3.5), which suffices for the cusp to be asymptotically ignorable.
3.3.1 Effective dimensionality

As a preliminary step, we first state a lemma showing that the denominator of the posterior distribution is not too small in terms of our target rate with probability tending to one. The condition (C1) is required to establish the lower bound by verifying the prior concentration, which implies that the prior distribution specified in Section 3.2.2 puts sufficient mass around the truth, where the closeness is measured through the Kullback-Leibler divergence and variation. The lemma plays a basic role in deriving the main results for the effective group dimension and posterior contraction.

**Lemma 3.1 (Evidence lower bound).** Suppose that (C1) is satisfied. Then there exists a constant $K_0 > 0$ such that

$$P_0 \left( \int_{\mathbb{R}^p} \Lambda_n(\beta)d\Pi(\beta) \geq \pi_p(s_0)e^{-K_0s_0(\log p/\sqrt{\log n})} \right) \to 1. \quad (3.7)$$

Using Lemma 3.1, we now examine the effective group dimension, by which we can restrict our attention to models of relatively small sizes. The following theorem shows that the posterior distribution is concentrated on much smaller group dimensions than the full size $p$. No further conditions are required other than (C1) for the theorem to hold.

**Theorem 3.1 (Effective group size).** Suppose that (C1) is satisfied. Then there exists a constant $K_1 > 1$ such that

$$E_0\Pi \left( \beta : s_\beta > K_1s_0 \big| Y^{(n)} \right) \to 0.$$ 

Theorem 3.1 allows us to rule out models of extremely large sizes in deriving posterior contraction properties. It is worth noting that only proper choices of the prior distribution on group dimension $s$ make the threshold in Theorem 3.1 proportional to the true group dimension $s_0$. By reading the proof of Theorem 3.1, one can see that the use of the prior in (3.3), which is viewed as a generalization of the prior in Castillo et al. (2015) for the group sparse modeling, is crucial to obtain such an optimal rate due to the lower bound in (3.7). Our threshold $K_1s_0$ is free of any assumption on the design matrix $X$. This is the main difference with the studies directly dealing with models such as Castillo et al. (2015) and Atchadé (2017). The simplicity comes from the restriction on the true parameter in the condition (C1), which is in this sense a great benefit considering its mildness. Readers can also refer to Section 4 of Castillo et al. (2015) for a relevant discussion.
3.3.2 Posterior contraction with restricted design

We are now ready to examine the posterior contraction rates for the model. In this subsection, we study posterior contraction under some reasonable assumption on the size of the true value of the linear predictor $X\beta_0$. The unrestricted case will be considered in the next subsection, where the contraction rates are actually slower than those given here.

The next theorem shows that the posterior distribution of $\beta$ contracts to the truth $\beta_0$ at the desired rate with respect to the root-average-squared Hellinger metric. To study posterior consistency and the contraction rates, we need a good separation of the true parameter and other values in the parameter space. For the root-average-squared Hellinger metric, there always exists a test with exponentially small errors for the truth and convex sets separated enough from the truth (Ghosal and van der Vaart, 2007). The theorem then follows by appropriately controlling the metric entropy measured by the minimum number of such small pieces that covers some subset of the parameter space, while showing that prior probability of the outside of the subset goes to zero fast enough (see, e.g., Ghosal et al., 2000; Ghosal and van der Vaart, 2007; Giné and Nickl, 2016; Ghosal and van der Vaart, 2017).

**Theorem 3.2** (Contraction rate, Hellinger). Suppose that (C1) is satisfied. Then there exists a constant $K_2 > 0$ such that

$$
\mathbb{E}_0 \Pi \left( \beta : H_n(\beta, \beta_0) > K_2 \sqrt{\frac{s_0(\log p \vee \tilde{q} \log n)}{n}} \middle| Y^{(n)} \right) \to 0.
$$

Theorem 3.2 provides basic results on the posterior contraction, but its implication for the coefficient $\beta$ is somewhat vague as it is relative to the root-average-squared Hellinger metric. The posterior contraction rates with respect to more concrete metrics can be recovered under an additional assumption. We make the following condition for the recovery.

(C2) For a given increasing $M_n$, the true parameter satisfies $\exp(\|\theta_0\|_{\infty}) = o(M_n^2)$ and

$$
\frac{M_n s_0 \sqrt{\log p \vee \tilde{q} \log n} \max_i \|X_i\|_0}{\psi_{2,1}((K_1 + 1)s_0\|X\|_0)} \to 0.
$$

An interesting example for $M_n$ is to choose $n^c$ for some small $c > 0$. Then the restriction on the true linear predictor has the form $\|X\beta_0\|_{\infty} < 2c \log n$ while the boundedness requirement in (C2) remains unrestrictive. Under the additional condition (C2), the next theorem provides the posterior contraction rates for $\beta$ with respect to more concrete metrics than the root-average-squared Hellinger metric.
Theorem 3.3 (Recovery). Suppose that (C1) and (C2) are satisfied. Then there exists a constant $K_3$ such that

\[
E_0 \Pi \left( \beta : \| W^{1/2} X (\beta - \beta_0) \|_2 > K_3 \sqrt{s_0 (\log p \lor q \log n)} \right| Y^{(n)} \right) \to 0,
\]

\[
E_0 \Pi \left( \beta : \| \beta - \beta_0 \|_2 > \frac{K_3 \sqrt{s_0 (\log p \lor q \log n)}}{\psi_2 ((K_1 + 1)s_0) \| X \|_0} \right| Y^{(n)} \right) \to 0,
\]

\[
E_0 \Pi \left( \beta : \| \beta - \beta_0 \|_{2,1} > \frac{K_3 s_0 \sqrt{\log p \lor q \log n}}{\psi_{2,1} ((K_1 + 1)s_0) \| X \|_0} \right| Y^{(n)} \right) \to 0.
\]

We are particularly interested in the second and third assertions of Theorem 3.3. The contraction rates in the results depend on the compatibility numbers $\psi_2$ and $\psi_{2,1}$, which are dependent on the true linear predictor $X \beta_0$ through the matrix $W$. Recall that the theorem allows $\| X \beta_0 \|_{\infty}$ to increase at a moderate order. If $\| X \beta_0 \|_{\infty}$ is known to be bounded, the rates in Theorem 3.3 are reduced to simpler forms in terms of the compatibility coefficients $\phi_2$ and $\phi_{2,1}$. The boundedness assumption on $\| X \beta_0 \|_{\infty}$ is rather easily met if $d_0$ is bounded. Yet, since $d_0$ can be increasing, this typically requires that some components of $X \beta_0$ should effectively cancel out. Comparable assumptions are often made in the literature in the context of high-dimensional generalized linear models (e.g., Ghosal, 1997; Abramovich and Grinshtein, 2016). For convenience, we give a modification of (C2) as follows. The contraction rates are summarized in the following corollary.

(C2*) The true parameter $\beta_0$ satisfies $\| X \beta_0 \|_{\infty} \leq M$ for some sufficiently large $M > 0$ and

\[
\frac{s_0 \sqrt{\log p \lor q \log n} \max_i \| X_i \|_0}{\phi_{2,1} ((K_1 + 1)s_0) \| X \|_0} \to 0.
\]

Corollary 3.1 (Recovery, bounded linear predictor). Suppose that (C1) and (C2*) are satisfied. Then there exists a constant $K_4$ such that

\[
E_0 \Pi \left( \beta : \| X (\beta - \beta_0) \|_2 > K_4 \sqrt{s_0 (\log p \lor q \log n)} \right| Y^{(n)} \right) \to 0,
\]

\[
E_0 \Pi \left( \beta : \| \beta - \beta_0 \|_2 > \frac{K_4 \sqrt{s_0 (\log p \lor q \log n)}}{\phi_2 ((K_1 + 1)s_0) \| X \|_0} \right| Y^{(n)} \right) \to 0,
\]

\[
E_0 \Pi \left( \beta : \| \beta - \beta_0 \|_{2,1} > \frac{K_4 s_0 \sqrt{\log p \lor q \log n}}{\phi_{2,1} ((K_1 + 1)s_0) \| X \|_0} \right| Y^{(n)} \right) \to 0.
\]

Similar bounds were also obtained for the frequentist convergence rates in group sparse
estimation for linear models (Huang and Zhang, 2010; Lounici et al., 2011) and for generalized linear models (Blazère et al., 2014). Readers can also refer to Ning et al. (2018) for comparable posterior contraction rates in linear regression under group sparsity.

The precise interpretations for the contraction rates are still somewhat hampered by vagueness of the compatibility conditions. As mentioned in Section 3.2.1, these compatibility numbers are likely to be bounded away from zero in common regression settings. These are then removed from the rates. In the next section, we will see that \( \phi_{2,1}((K_1 + 1)s_0) \) is indeed required to be bounded away from zero for our distributional approximation.

### 3.3.3 Posterior contraction for general prediction

It is worthwhile to note that Atchadé (2017) did not make any assumptions on the true value of the linear predictor \( X\beta_0 \) for posterior contraction. For a Bernstein-von Mises theorem and selection consistency in the next section, we will need to assume that \( \|X\beta_0\|_\infty \) is bounded, so the results in the previous subsection suffice to proceed. Nonetheless, it is still of interest to relax this restriction if one’s interest lies only on posterior contraction for the model. Rather than recovering contraction rates from the root-average-squared Hellinger metric, this can be established by directly constructing suitable test functions for \( \ell_2 \) and \( \ell_{2,1} \)-metrics. Note that Atchadé (2017) studied only an \( \ell_2 \)-contraction rate through a suitable test for the corresponding metric, which is comparable to ours below in Theorem 3.4 if \( q = 1 \), and \( \ell_q \)-rates for \( q \in (0, 2) \) were obtained with an \( \ell_2 \)-type compatibility condition (see Remark 5 of Atchadé (2017)). Our \( \ell_{2,1} \)-contraction rate differs in the sense that it is also obtained by a directly constructed test for the \( \ell_{2,1} \)-metric, and hence improves that in Remark 5 of Atchadé (2017) since \( \psi_2(s) \leq \psi_{2,1}(s) \) for any \( s \geq 1 \).

To construct suitable test functions, we need the following two modifications of (C2), respectively. The constructed tests are provided in Lemma 3.9.

(C2\(^\dagger\)) The true parameter \( \beta_0 \) satisfies

\[
\frac{s_0 \sqrt{\log p} \sqrt{q} \log n \max_i \|X_i\|_\infty}{\psi^2 \left( (K_1 + 1)s_0 \right) \|X\|_\infty} \to 0.
\]

(C2\(^\ddagger\)) The true parameter \( \beta_0 \) satisfies

\[
\frac{s_0 \sqrt{\log p} \sqrt{q} \log n \max_i \|X_i\|_\infty}{\psi_{2,1}^2 \left( (K_1 + 1)s_0 \right) \|X\|_\infty} \to 0.
\]

Note here that the compatibility numbers \( \psi_2 \) and \( \psi_{2,1} \) are squared in the denominators, which is the main difference from the conditions (C2) and (C2\(*\)). Since \( \psi_2(s) \leq \psi_{2,1}(s) \) for any \( s \geq 1 \),
it is clear that (C2*) is stronger than (C2†). We need only the latter for the $\ell_{2,1}$-contraction rate, but the former is needed for the first assertion of the following theorem.

**Theorem 3.4 (Contraction rates, general prediction).** Suppose that (C1) and (C2*) are satisfied. Then there exists a constant $K_5$ such that

$$
\mathbb{E} \Pi\left( \beta : \|\beta - \beta_0\|_2 > \frac{K_5 \sqrt{s_0 (\log p \vee \bar{q} \log n)}}{\psi_2((K_1 + 1)s_0 \|X\|_0)} y^{(n)} \right) \to 0,
$$

$$
\mathbb{E} \Pi\left( \beta : \|\beta - \beta_0\|_{2,1} > \frac{K_5 s_0 \sqrt{(\log p \vee \bar{q} \log n)}}{\psi_{2,1}((K_1 + 1)s_0 \|X\|_0)} y^{(n)} \right) \to 0.
$$

The second assertion holds even if the condition (C2*) is replaced by (C2†).

Again, we emphasize that if individual-level sparsity is imposed, say $\bar{q} = 1$, our $\ell_{2,1}$-rate in Theorem 3.4 is reduced to an $\ell_1$-rate which is faster (at the compatibility level) than the one given in Remark 5 of Atchadé (2017). Note also that the $\ell_2$-rate here is comparable to that in Atchadé (2017) if $\bar{q} = 1$.

It is worth comparing the contraction rates in Theorem 3.3 and Theorem 3.4. Clearly, the rates in Theorem 3.3 are faster than those in Theorem 3.4 at the compatibility levels as the compatibility numbers in Theorem 3.4 are squared. This is a natural consequence in that the former makes some size restriction of $X\beta_0$. A similar situation can also be found in the literature for linear regression; for example, compare Section 2 and Section 4 of Castillo et al. (2015). If $\|X\beta_0\|_\infty$ is bounded, the compatibility numbers $\psi_2$ and $\psi_{2,1}$ in Theorem 3.4 can be replaced by $\phi_2$ and $\phi_{2,1}$. Still, these rates are slower than the rates in Corollary 3.1 due to the squared compatibility numbers.

### 3.4 Bernstein-von Mises theorem and selection consistency

#### 3.4.1 Shape approximation to the posterior distribution

In the previous section, the posterior contraction rates for the high-dimensional regression coefficients have been studied under group sparsity. Combined with the assumption that the true parameters are sufficiently distinguishable from zero (which we will call the beta-min condition following the convention), these results imply that the posterior distribution of the group index $S$ includes the true support $S_0$ with probability tending to one. However, this does not imply the selection consistency, the property that the posterior distribution of $S$ is concentrated on $S_0$ with probability tending to one.

To verify the selection consistency, we need to deal with the marginal posterior distribution of the group index $S$, whose exact form is obtained by integrating $\beta$ out from the posterior.
distribution. In our logit model, its closed form expression is not available due to the lack of conjugacy of the multinomial likelihood and the prior. We shall thus rely on a shape approximation to the posterior distribution for the selection consistency, which is the most natural solution to the problems in the literature (e.g., Castillo et al., 2015). The techniques used to derive the classical Bernstein-von Mises theorems for finite dimensional models may not work in our high-dimensional sparse setup. In this subsection, we directly characterize a distributional approximation for the posterior distribution.

As mentioned above, we need to assume that $\|X_0\|_\infty$ is bounded and the compatibility $\phi_{2,1}((K_1 + 1)s_0)$ is bounded away from zero throughout this section. These are clarified in the following conditions.

(C3) The true parameter $\beta_0$ satisfies $\|X_0\|_\infty \leq M$ for a sufficiently large $M$.
(C4) The compatibility number $\phi_{2,1}((K_1 + 1)s_0)$ is bounded away from zero.

The condition (C4) is weaker than assuming that $\phi_2$ is bounded away from zero, as we have $\phi_{2,1}(s) \geq \phi_2(s)$ for any $s \geq 1$ by the Cauchy-Schwarz inequality. As we discussed in the last paragraph of Section 3.2.1, this condition is likely to be satisfied under typical settings.

Now, we find an approximate likelihood ratio for a distributional approximation to the posterior distribution. For $\theta_i := X_i\beta$ and $\theta_{0i} := X_i\beta_0$, observe that by Taylor’s theorem, the log-likelihood ratio is given by

$$\log \Lambda_n(\beta) = \sum_{i=1}^{n} (Y_i - \mu_i)^T (\theta_i - \theta_{0i}) - \frac{1}{2} \sum_{i=1}^{n} (\theta_i - \theta_{0i})^T \{\text{diag}(\bar{\mu}_i) - \bar{\mu}_i \bar{\mu}_i^T\} (\theta_i - \theta_{0i}),$$

where $\bar{\mu}_i = (1 + h(\bar{\theta}_i)^T 1_{m-1})^{-1} h(\bar{\theta}_i)$ for $\bar{\theta}_i = t\theta_{0i} + (1 - t)\theta_i$ with some $t \in (0, 1)$. We will use the approximate likelihood ratio $\Lambda_n^*$ defined by replacing $\bar{\mu}_i$ in $\Lambda_n$ by $\mu_i$, ignoring the remainder of the Taylor expansion. The logarithm of $\Lambda_n^*(\beta)$ is then a quadratic form of $\beta - \beta_0$. Whereas the exact quadratic expression for the log-likelihood ratio is available for normal models (Castillo et al., 2015), this is clearly not the case here for our logit model.

Ignoring the remainder, we first show that the model is locally asymptotic normal for $\beta$ in a neighborhood of $\beta_0$. The assertion is made in Lemma 3.2 below, which requires the following stronger boundedness condition.

(C5) The true parameter $\beta_0$ satisfies

$$\frac{s_0^3 (\log p \vee q \log n)^{3/2} \max_i \|X_i\|_o}{\|X\|_o} \to 0.$$ 

The condition is required to make the remainder of the local asymptotic normal expansion tend to zero. It is clear that the condition (C5) (combined with (C3)) is stronger than (C2*),
and hence (C2*) will be overlooked henceforth. The condition (C5) also implies that \( s_0^3 (\log p \vee q_1 \log n)^3 = o(n) \) by the inequality \( \| X \|_\infty \leq \sqrt{n} \max_i \| X_i \|_\infty \). The lemma for locally asymptotic normality is formally stated below.

**Lemma 3.2 (Locally asymptotic normality).** Suppose that (C5) is satisfied and define

\[
A_n = \left\{ \beta \in \mathbb{R}^d^* : s_\beta \leq K_1 s_0, \| \beta - \beta_0 \|_{2,1} \leq \frac{M s_0 \sqrt{\log p \vee q_1 \log n}}{\| X \|_\infty} \right\},
\]

for a sufficiently large \( M > 0 \). Then we have that

\[
\sup_{\beta \in A_n} | \log \Lambda_n(\beta) - \log \Lambda_n^*(\beta) | \to 0.
\]

The lemma implies that the true likelihood ratio \( \Lambda_n \) can be effectively replaced by the approximate one \( \Lambda_n^* \) in the subspace \( A_n \subset \mathbb{R}^d^* \). Since the posterior probability of the set \( A_n \) tends to one in probability by Corollary 3.1 and (C4), the lemma allows us to construct a distributional approximation with normal mixtures in the subset \( A_n \) of the parameter space.

Besides the non-normal likelihood, we also need to make the influence of the prior distribution asymptotically ignorable in the posterior distribution. This is the main reason why a 'large' \( \lambda_j \) in Atchadé (2017) is not good for a distributional approximation, since a large \( \lambda_j \) makes the prior distribution for the corresponding nonzero coefficient shrink heavily towards zero. Similar to Castillo et al. (2015), we need the condition

\[
s_0^2 (\log p \vee q_1 \log n) / \| X \|_\infty \to 0,
\]

which is often called the 'small' lambda regime. This condition is, however, automatically satisfied by our range in (3.5) and the condition (C5). More specifically, the small lambda regime is translated into \( s_0^2 (\log p \vee q_1 \log n) = o(n) \) by inserting the upper bound of \( \lambda_j \), which is clearly weaker than (C5) as (C5) implies \( s_0^3 (\log p \vee q_1 \log n)^3 = o(n) \).

Using the approximate likelihood ratio \( \Lambda_n^* \), the approximate distribution of the posterior can be written as

\[
\Pi^\infty(\beta \in \mathcal{B} | Y^{(n)}) = \frac{\sum_{S : s \leq K_1 s_0} \pi_p(s)^{-1} \left\{ \prod_{j \in S} (\lambda_j / a_j)^{q_j} \right\} \int_{\mathbb{R}^d^*} \Lambda_n^*(\beta) d\{ L(\beta_S) \otimes \delta_0(\beta_{S^c}) \}}{\sum_{S : s \leq K_1 s_0} \pi_p(s)^{-1} \left\{ \prod_{j \in S} (\lambda_j / a_j)^{q_j} \right\} \int_{\mathbb{R}^d^*} \Lambda_n^*(\beta) d\{ L(\beta_S) \otimes \delta_0(\beta_{S^c}) \}},
\]

for every measurable \( \mathcal{B} \subset \mathbb{R}^d^* \), where \( L \) denotes the Lebesgue measure. Similar to the true posterior distribution, \( \Pi^\infty(\cdot | Y^{(n)}) \) is a mixture of products of continuous distributions and point masses at zero, but the continuous parts are multivariate normal due to \( \Lambda_n^* \). For \( \mu = (\mu_1^T, \ldots, \mu_n^T)^T \), let \( U = W^{-1/2}(Y^{(n)} - \mu) \) be the standardized observation vector. Then \( \Pi^\infty(\cdot | Y^{(n)}) \)
can also be expressed as

$$
\Pi^\infty(\cdot | Y^{(n)}) = \sum_{S: s \leq K_1 s_0} \hat{w}_S N(\hat{\beta}_S, (X_S^T W X_S)^{-1}) \otimes \delta_{0,S^c}, \quad (3.8)
$$

where \( \hat{\beta}_S = (X_S^T W X_S)^{-1} X_S^T W^{1/2} (U + W^{1/2} X \beta_0) \) and

$$
\hat{w}_S \propto \frac{\pi_p(s)}{L_S} \left\{ \prod_{j \in S} \left( \frac{\lambda_j}{a_j} \right)^{q_j} \right\} \sqrt{\frac{(2\pi)^d}{\det(X_S^T W X_S)}} \exp \left\{ -\frac{1}{2} \| W^{1/2} X_S \hat{\beta}_S \|^2 \right\} \quad (3.9)
$$

It is easy to see that both expressions are identical. The results are formally summarized in the following theorem.

**Theorem 3.5 (Distributional approximation).** Suppose that (C1) and (C3)–(C5) are satisfied. Then

$$
E_0 \left\| \Pi(\cdot | Y^{(n)}) - \Pi^\infty(\cdot | Y^{(n)}) \right\|_{TV} \to 0.
$$

Theorem 3.5 allows us to use the distribution \( \Pi^\infty(\cdot | Y^{(n)}) \) in place of the posterior distribution in deriving the selection consistency. Combined with the selection consistency, it also guarantees frequentist coverage of credible sets. These will be considered in the following subsections. Besides, the distribution \( \Pi^\infty(\cdot | Y^{(n)}) \) may also be used for the computation of the posterior distribution through approximation. Due to the multinomial likelihood and the Laplace-type prior in (3.4), the closed form expression for the marginal posterior distribution of \( S \) is not directly available. The actual computation of the posterior distribution is thus hampered and one may need to explore different model spaces by jointly updating the posterior for \( (S, \beta) \) using proper Markov chain Monte Carlo (MCMC) methods such as the reversible jump MCMC, which typically exhibits slow mixing and extensive computational burden. This can be avoided by using the approximate distribution. Note that \( \Pi^\infty(\cdot | Y^{(n)}) \) contains the true linear predictor \( X \beta_0 \) in its form. Suppose that we have a ‘good’ estimator for \( \beta_0 \) such that \( X \beta_0 \) is well recovered, which may be implemented through penalized estimation with a well-known optimization strategy. The true parameter \( \beta_0 \) in the approximate distribution can easily be replaced by this estimator. We then have an ‘approximate’ marginal posterior for \( S \), which is the weights \( \hat{w}_S \) with \( \beta_0 \) replaced by the estimator. Due to the dimensionality, posterior sampling of \( S \) may still require to explore the model space in its MCMC iterations, which can be easily conducted by using standard updating schemes such as Gibbs sampling or a Metropolis-Hastings algorithm. Once the ‘approximate’ marginal posterior for \( S \) is obtained, each mixture component is directly computed by a product of a normal distribution and a point mass as in (3.8), with \( \beta_0 \) replaced.
by the estimator.

It may be worth mentioning that the normal approximation may fail to work well in multi-dimensions. The performance can be partly improved by simulating the true conditional posterior of $\beta_S$ given $S$ through MCMC iterations, while still approximating the marginal posterior of $S$. Nevertheless, the whole computation through the approximation may still behave poorly due to the bounding error arising when an estimator replaces the true value. We do not discuss more details as the approximation result in this subsection is primarily used as a technical gadget for further investigations in the following subsections.

### 3.4.2 Selection consistency

The distributional results in Theorem 3.5 enable to use the approximate distribution to derive the selection consistency. To this end, we first show that the posterior distribution is concentrated on subsets of the true group index $S_0$ with probability tending to one. The results will then be combined with the beta-min condition to verify the selection consistency. The theorem requires the following condition on the prior.

(C6) The true parameter $\beta_0$ satisfies $s_0 \lesssim p^a$ for some $a \geq 0$ such that $a + B_* < A_4$, where

$$B_* := \max_{1 \leq s \leq n} \max_{0 \leq j \neq k \leq m-1} \left\{ \frac{1}{\mu_{ij}} + \frac{1}{\mu_{ik}} \right\}, \quad \mu_{i0} = 1 - \sum_{j=1}^{m-1} \mu_{ij}.$$  

The condition gives an explicit lower bound for the constant $A_4$ in the prior distribution. Note that if any of $\mu_{ij}$, $i = 1, \ldots, n$, $j = 0, \ldots, m-1$, is close to zero or one, $A_4$ should be chosen sufficiently large. Indeed, this is almost always the case in a practical sense, which means that we need a prior $\pi_\sigma$ that decays sufficiently fast for the selection consistency. It may be worth noting that $B_*$ is simply reduced to 1 in many normal linear models (e.g., Castillo et al., 2015), and a large $B_*$ here is needed due to the asymmetric likelihood function. Also note that $B_*$ is bounded only if the condition (C3) is satisfied. If $B_*$ is not bounded, $A_4$ should be chosen to be a sequence increasing at the order of $B_*$. This induces a suboptimal contraction rate. We do not consider such a case in this study.

**Theorem 3.6 (Selection, no supersets).** Suppose that (C1) and (C3)–(C6) are satisfied. Then

$$E_0 \Pi \left( \beta: S_\beta \supset S_0, S_\beta \neq S_0 \right| Y^{(n)}) \rightarrow 0.$$  

Note that coefficients that are too close to zero cannot be identified by any statistical method for variable selection. We thus need to assume that the absolute values of the true
nonzero coefficients are bigger than some threshold, which is so-called the beta-min condition in the literature. The threshold is clearly dependent on the contraction rates in Theorem 3.3. The condition is clarified as follows.

(C7) The true parameter $\beta_0$ satisfies

$$
\min_{\beta_{0,j} \neq 0} \|\beta_{0,j}\|_2 > \frac{K_3 \sqrt{s_0 (\log p \vee \log n)}}{\phi_2((K_1 + 1)s_0)\|X\|_0}.
$$

By combining the condition with the contraction rate results in Theorem 3.3, we finally obtain the selection consistency. We close this subsection by writing this formally in the corollary below.

**Corollary 3.2** (Selection consistency). Suppose that (C1) and (C3)–(C7) are satisfied. Then

$$
E_0 \Pi (\beta : S_\beta \neq S_0 | Y^{(n)}) \rightarrow 0.
$$

### 3.4.3 Frequentist coverage of credible sets

Recall that Theorem 3.5 implies that the posterior distribution is approximated by the mixture distribution $\Pi^\infty(\cdot | Y^{(n)})$. It is easy to see that the selection consistency in Corollary 3.2 reduces the approximate distribution to a single mixture component with the true support $S_0$. Then the resulting distribution can be used to quantify coverage of credible sets. The reduced distribution is formulated in the following corollary, which we call a Bernstein-von Mises theorem for our high-dimensional logit model.

**Corollary 3.3** (Bernstein-von Mises). Suppose that (C1) and (C3)–(C7) are satisfied. Then

$$
E_0 \left\| \Pi(\cdot | Y^{(n)}) - \mathcal{N}(\hat{\beta}_{S_0}, (X_S^T W X_S)^{-1}) \otimes \delta_{0,S_0}\right\|_{TV} \rightarrow 0.
$$

Corollary 3.3 implies that the marginal posterior distribution of $\beta_j \in \mathbb{R}^{q_j}$ for the $j$th group, given by a mixture of a continuous component and a point mass, is well approximated by a $q_j$-variate normal distribution if $j \in S_0$, and by a point mass if $j \notin S_0$.

Let $\beta_{j,\ell}$ and $\beta_{0,j,\ell}$ be the $\ell$th entries of $\beta_j$ and $\beta_{0,j}$. For any $\alpha \in (0, 1)$, a reasonable upper 1 $-$ $\alpha$ credible limit for $\beta_{j,\ell}$ can be constructed using the marginal posterior distribution as in Castillo et al. (2015) (see page 2000 of that paper). Denoting this credible limit by $\hat{R}_{j,\ell}^\alpha$, we see that by Corollary 3.3, $P_0(\beta_{0,j,\ell} = 0) \rightarrow 1$ if $j \notin S_0$. For $j \notin S_0$, we have $P_0(\beta_{0,j,\ell} \leq \hat{R}_{j,\ell}^\alpha) \rightarrow 1 - \alpha$
if the corresponding entry of $\hat{\beta}_{S_0}$ has an asymptotically normal distribution with suitable mean and variance. The asymptotic normality can be verified by the Lindeberg-Feller central limit theorem which requires the following condition given below. In the statement, $e_k$ stands for a $d_0$-dimensional unit vector whose only nonzero value is in the kth entry and $X_{S_0,i} \in \mathbb{R}^{(m-1) \times d_0}$ is the submatrix of $X_{S_0}$ for the ith observation.

(C8) The true parameter $\beta_0$ satisfies that, for given $k \leq d_0$,

$$
\max_{1 \leq i \leq n} \frac{e_k^T (X_{S_0}^T W X_{S_0})^{-1} X_{S_0,i}^T W_i X_{S_0,i} (X_{S_0}^T W X_{S_0})^{-1} e_k}{e_k^T (X_{S_0}^T W X_{S_0})^{-1} e_k} \to 0.
$$

The condition (C8) is needed to satisfy Lindeberg’s condition. Note that the denominator is the sum of the numerator over $i$, which means that the contribution of any particular term to the variance is uniformly asymptotically negligible. Under the condition, we have the asymptotic normality of $e_k^T \hat{\beta}_{0,S_0}$, the $k$th entry of $\hat{\beta}_{S_0}$, in terms of weak convergence. We close this section by formally stating the assertion in the following lemma.

**Lemma 3.3** (Asymptotic normality of $\hat{\beta}_{S_0}$). Suppose that (C3) and (C8) are satisfied. Then $e_k^T (\hat{\beta}_{S_0} - \beta_{0,S_0}) / \sqrt{e_k^T (X_{S_0}^T W X_{S_0})^{-1} e_k}$ weakly converges to the standard normal distribution.

### 3.5 Proofs

#### 3.5.1 Preliminary results

As a preliminary step, here we provide three lemmas. The first lemma quantifies the upper bound for exponential moments of independent bounded random vectors. The results are directly applicable to our multinomial response variable. The second lemma gives a stochastic bound of $\max_{1 \leq j \leq p} \|X_j^T (Y^{(n)} - \mu)\|_2$ for our logit model. The third one gives bounds for eigenvalues of the covariance matrix of a multinomial random variable.

**Lemma 3.4.** Let $(Z_j \in \mathbb{R}^{r_j})_{j=1}^n$ be a sequence of independent random vectors such that for every $j \leq n$, $\mathbb{E}Z_j = 0$ and $\mathbb{P}\{Z_j \in \text{supp}(Z_j)\} = 1$ for a bounded support $\text{supp}(Z_j)$ of $Z_j$ (note that for every $j \leq n$, the entries in $Z_j$ need not be independent). Let $Z = (Z_1^T, \ldots, Z_n^T)^T \in \mathbb{R}^{\sum_{j=1}^n r_j}$. For any real positive semidefinite matrix $Q$, we have

$$
\mathbb{E} \exp \left\{ t Z^T Q Z \right\} \leq \exp \left\{ \frac{t \max_j b_j^2 \text{tr}(Q)}{1 - 2t \max_j b_j^2 \|Q\|_{sp}} \right\}, \quad 0 < t < \frac{1}{2 \max_j b_j^2 \|Q\|_{sp}},
$$

where $b_j^2 = \mathbb{E} \|Z_j\|_2^2$. 

87
where for every \( j \leq n \),
\[
\bar{b}_j = \max_{\xi_j \in \text{supp}(Z_j)} \|\xi_j\|_2, \quad \tilde{b}_j = \max_{\xi_j, \xi'_j \in \text{supp}(Z_j)} \|\xi_j - \xi'_j\|_2.
\]

**Lemma 3.5.** For the logit model in \((3.1)\), there exists a constant \( C_* > 0 \) such that for any \( \beta_0 \in \mathbb{R}^{d^*} \),
\[
\mathbb{P}_0 \left( \max_{1 \leq j \leq p} \|X^T_j (Y^{(n)} - \mu)\|_2 > C_* \|X\|_\infty \sqrt{\log p \lor q \log n} \right) \to 0. \tag{3.10}
\]

**Lemma 3.6.** Let \( Z = (Z_1, \ldots, Z_m)^T \) be a multinomial random variable with success probabilities \( \pi_k \), \( k = 1, \ldots, m \), such that \( \sum_{k=1}^m Z_k = 1 \) and \( \sum_{k=1}^m \pi_k = 1 \). Let \( \tilde{Z} = (Z_1, \ldots, Z_{m-1})^T \) be the reduced form of \( Z \). Then the covariance matrix \( \Sigma \) of \( \tilde{Z} \) satisfies
\[
\pi_m \min_{1 \leq k \leq m-1} \pi_k \leq \rho_{\min}(\Sigma) \leq \rho_{\max}(\Sigma) \leq \max_{1 \leq k \leq m-1} \pi_k.
\]

### 3.5.2 Proofs for the main results

In this subsection, we provide proofs for Theorems 3.1–3.6.

**Proof of Theorem 3.1.** Define the set \( \mathcal{B} = \{ \beta : s_\beta > \bar{s} \} \) for a given integer \( \bar{s} \geq s_0 \). It follows from \((3.3)\) that
\[
\Pi(\mathcal{B}) \leq \pi_p(s_0) \sum_{s=\bar{s}+1}^p \left\{ \frac{A_2}{(p \lor q)^{A_4}} \right\}^{s-s_0} \leq \pi_p(s_0) \sum_{j=0}^{\bar{s}+1-s_0} \left\{ \frac{A_2}{(p \lor q)^{A_4}} \right\}^j,
\]
where the summation on the rightmost side is bounded. Hence, for \( \mathcal{E}_n \) the event in \((3.7)\), we have that by Lemma 3.1,
\[
\mathbb{E}_0 \Pi(\mathcal{B} | Y^{(n)}) 1_{\mathcal{E}_n} = \mathbb{E}_0 \left[ \int_8 \Lambda_n(\beta) d\Pi(\beta) \int \Lambda_n(\beta) d\Pi(\beta) 1_{\mathcal{E}_n} \right] \leq \pi_p(s_0)^{-1} \exp\{C_1 s_0 (\log p \lor q \log n)\} \Pi(\mathcal{B}) \leq \exp \{ (\bar{s} + 1 - s_0) [\log A_2 - A_4 (\log p \lor q \log n)] + C_1 s_0 (\log p \lor q \log n) \},
\]
for some constant \( C_1 > 0 \) and sufficiently large \( n \). Choosing \( \bar{s} = C_2 s_0 \) for an integer \( C_2 > 1 + 2C_1/A_4 \), the rightmost side of the display goes to zero. The proof is then complete since...
\[ \mathbb{E}_0 \Pi(B \mid Y^{(n)}) \leq \mathbb{E}_0 \Pi(B \mid Y^{(n)}) \mathbb{I}_{E_n} + \mathbb{P}_0(E^*_n) \text{ and } \mathbb{P}_0(E^*_n) \to 0 \text{ by Lemma 3.1.} \]

**Proof of Theorem 3.2.** Let \( A'_n = \{ \beta \in \mathbb{R}^d : s_\beta \leq K_1 s_0 \} \). Then for every \( \epsilon > 0 \),

\[
\mathbb{E}_0 \Pi \left( \beta \in \mathbb{R}^d : H_n(\beta, \beta_0) > \epsilon \mid Y^{(n)} \right) 
\leq \mathbb{E}_0 \Pi \left( \beta \in A'_n : H_n(\beta, \beta_0) > \epsilon \mid Y^{(n)} \right) + \mathbb{E}_0 \Pi \left( A^c_n \mid Y^{(n)} \right).
\]

Since the second term on the right hand side goes to zero by Theorem 3.1, it suffices to show that the first term goes to zero.

By Lemma 2 of Ghosal and van der Vaart (2007), there exists a test \( \varphi_n \) such that for every \( \beta_1 \) with \( H_n(\beta_0, \beta_1) > \epsilon \),

\[
\mathbb{E}_0 \varphi_n \leq \exp(-n\epsilon^2/2), \quad \sup_{\beta : H_n(\beta, \beta_1) \leq \epsilon/18} \mathbb{E}_\beta (1 - \varphi_n) \leq \exp(-n\epsilon^2/2).
\]

To apply Lemma 9 of Ghosal and van der Vaart (2007), we now show that \( \log N(\epsilon_n/36, A^*_n, H_n) \leq n\epsilon_n^2 \), for \( \epsilon_n = \sqrt{s_0 (\log p \vee \log n)/n} \) and \( A^*_n := \{ \beta \in A'_n : \| \beta - \beta_0 \|_\infty \leq D_n \} \) with \( D_n := n\sqrt{q}(p^{1/7} \vee n)^{L_2} \| X \|_0^{-1} \). Since the squared Hellinger distance is bounded by the Kullback-Leibler divergence, using Lemma 3.6 and (3.24), we have that for every \( \beta \) in \( A'_n \),

\[
\sqrt{n} H_n(\beta, \beta_0) \leq \| X(\beta - \beta_0) \|_2 \leq \| X \|_{C^0} \| \beta - \beta_0 \|_{2,1} \leq s_0 \sqrt{q} \| X \|_{C^0} \| \beta - \beta_0 \|_{\infty},
\]

where the last inequality follows by the inequality \( |S_{\beta - \beta_0}| \leq s_\beta + s_0 \leq (K_1 + 1)s_0 \) for such \( \beta \).

We then obtain that for some \( C_1 > 0 \),

\[
N(\epsilon_n/36, A^*_n, H_n) \leq N \left( \frac{C_1 \sqrt{n} \epsilon_n}{s_0 \sqrt{q} \| X \|_{C^0}}, A^*_n, \| \cdot \|_{\infty} \right) \leq \left( \frac{p}{K_{1s_0}} \right) \left( \frac{3s_0 \sqrt{q} D_n \| X \|_{C^0}}{C_1 \sqrt{n} \epsilon_n} \right)^{K_{1s_0} q}.
\]

By plugging in the values of \( D_n \),

\[
\log N(\epsilon_n/36, A^*_n, H_n) \leq s_0 \log p + s_0 \sqrt{q}(\log n + \log s_0 + \log q) \leq n\epsilon_n^2.
\]

Now, Lemma 9 of Ghosal and van der Vaart (2007) implies that for every \( \epsilon > \epsilon_n \), there exists a test \( \tilde{\varphi}_n \) such that for some \( C_2 > 0 \),

\[
\mathbb{E}_0 \tilde{\varphi}_n \leq \exp\left(\frac{C_2 n^2 \epsilon_n - n\epsilon^2/2}{2}\right), \quad \sup_{\beta \in A^*_n : H_n(\beta, \beta_0) > \epsilon} \mathbb{E}_\beta (1 - \tilde{\varphi}_n) \leq \exp(-n\epsilon^2/2).
\]

Note that by (3.3), we have \( - \log \pi_p(\theta_0) \leq s_0 (\log p \vee \log n) \) as \( \pi_p(0) \) is bounded away from
zero. Hence, for $\mathcal{E}_n$ the event in (3.7) and some constant $C_3 > 0$,

\[
\mathbb{E}_0 \Pi \left( \beta \in \mathcal{A}_n' : H_n(\beta, \beta_0) > \epsilon \mid Y^{(n)} \right) \\
\leq \mathbb{E}_0 \Pi \left( \beta \in \mathcal{A}_n' : H_n(\beta, \beta_0) > \epsilon \mid Y^{(n)} \right) \Pi \left( 1 - \varphi_n \right) + \mathbb{E}_0 \varphi_n + \mathbb{P}_0 \mathcal{E}_n^c
\]

\[
\leq \left\{ \beta \in \mathcal{A}_n' : H_n(\beta, \beta_0) > \epsilon \right\} \mathbb{E}_0 \beta \left( 1 - \varphi_n \right) + \Pi \left( \mathcal{A}_n' \setminus \mathcal{A}_n^* \right) e^{C_{3n}\epsilon^2} + \mathbb{E}_0 \varphi_n + \mathbb{P}_0 \mathcal{E}_n^c.
\]

The last term $\mathbb{P}_0 \mathcal{E}_n^c$ on the right most side goes to zero by Lemma 3.1. The remaining terms other than $\Pi \left( \mathcal{A}_n' \setminus \mathcal{A}_n^* \right) e^{C_{3n}\epsilon^2}$ go to zero by choosing $\epsilon = C_4 \epsilon_n$ for a sufficiently large $C_4$. Next, to complete the proof, note that

\[
\Pi \left( \mathcal{A}_n' \setminus \mathcal{A}_n^* \right) = \Pi \{ \beta \in \mathbb{R}^{d_s} : s_\beta \leq K_1 s_0, \| \beta - \beta_0 \|_\infty > D_n \}
= \sum_{S : s \leq K_1 s_0} \frac{\pi_p(s)}{p} \int \left\{ \beta : \| \beta - \beta_0 \|_\infty > D_n \right\} g_s(\beta_S) d\beta_S
\leq \sum_{S : s \leq K_1 s_0} \left[ e^{s_\beta (\| \beta \|_2,1)} \left( \frac{p}{s} \right) - 1 \right] \left( \frac{A_2}{(p \vee n\bar{q})^4} \right) s \\
\times \int \left\{ \beta : \| \beta - \beta_0 \|_\infty > D_n \right\} g_s(\beta_S - \beta_0, S) d\beta_S,
\]

by the inequalities $g_S(\beta_S) \leq \bar{c} (\| \beta \|_2,1) g_S(\beta_S - \beta_0, S)$ and $\pi_p(s) \leq A_2 (p \vee n\bar{q})^{-4} \pi_p(0) \leq A_2 (p \vee n\bar{q})^{-4}$ for every $S$. Letting $\bar{\beta}_S = \beta_S - \beta_0, S$ and $\bar{\beta}_j = \beta_j - \beta_{0,j}$, it is easy to see that for every $\zeta \in \mathbb{R}^{d_s}$,

\[
\int_{\| \bar{\beta}_S \|_\infty > \zeta} g_s(\bar{\beta}_S) d\bar{\beta}_S \leq \int_{\| \bar{\beta}_S \|_\infty > \zeta} \prod_{j \in S} \left( \frac{\lambda_j}{a_j} \right)^{q_j} \exp \left( -\frac{\lambda_j}{a_j} \| \bar{\beta}_j \|_1 \right) d\bar{\beta}_S,
\]

by the trivial inequality $\| \bar{\beta}_j \|_2 \geq \| \bar{\beta}_j \|_1 / \sqrt{q_j} \geq \| \bar{\beta}_j \|_1 / \sqrt{q}$. Using the tail probability of Laplace distributions given by $\int_{|x| > t} 2^{-1} \lambda e^{-\lambda |x|} dx = \exp(\lambda t)$ for every $t > 0$, we see that the right hand side of the last display is further bounded by

\[
\int_{\| \bar{\beta}_S \|_\infty > \zeta} \left( \frac{\lambda}{2\sqrt{q}} \right)^{d_s} \exp \left( -\frac{\lambda}{\sqrt{q}} \| \bar{\beta}_S \|_1 \right) d\bar{\beta}_S \prod_{j \in S} \left( \frac{2\sqrt{q} \lambda_j}{a_j \lambda} \right)^{q_j}
\leq d_s \exp \left( -\frac{\lambda \zeta}{\sqrt{q}} \right) \left( \frac{\lambda}{\zeta} \right)^{d_s} \prod_{j \in S} \left( \frac{2\sqrt{q}}{a_j} \right)^{q_j}.
\]

Note further that $\prod_{j \in S} \left( 2\sqrt{q}/a_j \right)^{q_j} \leq \bar{q}^{d_s/2} \leq e^{(1/2)s\log n}$ as $a_j \geq 2$ for every $j \leq p$ (see (3.4)).
and that
\[
d_S \log(\lambda/\gamma) \leq d_S \log \left\{ L_1 L_3 (p^{1/q} \vee n)^{L_2} n^{-1/2} \right\} \leq n \epsilon_n^2.
\]
Hence, for some \( C_5 > 0 \), we bound the rightmost side of (3.12) by
\[
e^{\gamma_0 \| \beta_0 \|_{2,1}} \sum_{s=1}^{K_1} \left\{ \frac{A_2}{(p \vee n^q)^{A_4}} \right\}^s \exp \left\{ -\gamma \left( \frac{D_n - \| \beta_0 \|_{\infty}}{\sqrt{q}} \right) + C_5 n \epsilon_n^2 \right\}
\]
\[
\lesssim s_0 q \exp \left\{ -\gamma D_n / \sqrt{q} + 2 \gamma_0 \| \beta_0 \|_{2,1} + C_5 n \epsilon_n^2 \right\},
\]
since the summation is bounded and \( \gamma \| \beta_0 \|_{\infty} / \sqrt{q} \leq \gamma_0 \| \beta_0 \|_{2,1} \). Note that \( \gamma \| \beta_0 \|_{2,1} \lesssim n \epsilon_n^2 \) by (C1) and \( \gamma D_n / \sqrt{q} = \gamma n (p^{1/q} \vee n)^{L_2} \| X \|_0^{-1} \gtrsim n \) by (3.5). Then the right hand side of the last display is further bounded by a constant multiple of \( e^{-C_6 n + C_7 n \epsilon_n^2} \) for some \( C_6, C_7 > 0 \), and hence \( \Pi(A_n \setminus A_n^*) e^{C_7 n \epsilon_n^2} \) goes to zero. This completes the proof. \( \square \)

To prove Theorem 3.3, we need to recover the closeness of \( \theta_i \) and \( \theta_{0i} \) from the smallness of \( H(f_{\beta,i}, f_{0,i}) \). Note that \( H(f_{\beta,i}, f_{0,i}) \) has the global minimum zero at \( \theta_i = \theta_{0i} \) and strictly increases as \( \theta_i \) gets farther from \( \theta_{0i} \) along a straight line in any direction on \( \mathbb{R}^{m-1} \) (see the first part of the proof of Lemma 3.7 below). Hence whenever \( \| \theta_{0i} \|_{\infty} \) is bounded, \( H(f_{\beta,i}, f_{0,i}) \to 0 \) implies that \( \| \theta_i - \theta_{0i} \|_{\infty} \to 0 \), regardless of their orders. This, however, does not always hold for increasing \( \| \theta_{0i} \|_{\infty} \). To see this through contraposition, plug in \( \theta_i = n \) and \( \theta_{0i} = 2n \) for the case \( m = 2 \). Then we have \( \| \theta_i - \theta_{0i} \|_{\infty} \to \infty \) but \( H(f_{\beta,i}, f_{0,i}) \to 0 \), which contradicts the assertion (see (3.13) below for the expression for \( H(f_{\beta,i}, f_{0,i}) \)). This happens because \( v(\theta_{0i}) := \lim_{t \to \infty} \inf_{\theta_i : \| \theta_i \|_{\infty} = t} H(f_{\beta,i}, f_{0,i}) \) goes to zero as \( \theta_{0i} \) moves in certain direction on \( \mathbb{R}^{m-1} \) (in the example above for \( m = 2 \), i.e., the univariate case, both \( \theta_{0i} \) and \( \theta_i \) move in the same direction).

By the shape of \( H(f_{\beta,i}, f_{0,i}) \) as a function of \( \theta_i \), we see that if \( H(f_{\beta,i}, f_{0,i}) \) goes to zero faster than \( v(\theta_{0i}) \), then \( H(f_{\beta,i}, f_{0,i}) \to 0 \) still implies \( \| \theta_i - \theta_{0i} \|_{\infty} \to 0 \), which gives an explicit condition on \( \| \theta_{0i} \|_{\infty} \). This is formalized in the following lemma.

**Lemma 3.7.** For any \( \beta_0 \) satisfying \( \exp(\| X \beta_0 \|_{\infty}) = o(M_n^2) \) with a given increasing sequence \( M_n \), \( H(f_{\beta,i}, f_{0,i}) \lesssim M_n^{-1} \) implies that \( \| \theta_i - \theta_{0i} \|_{\infty} \to 0 \).

**Proof of Theorem 3.3.** First, observe that
\[
\zeta_i(\theta_i) := \frac{1}{2} H^2(f_{\beta,i}, f_{0,i}) = 1 - \frac{1 + h((\theta_i + \theta_{0i})/2)^T 1_{m-1}}{\sqrt{1 + h(\theta_i)^T 1_{m-1}} \sqrt{1 + h(\theta_{0i})^T 1_{m-1}}}, \tag{3.13}
\]
for the function \( h \) introduced in Section 3.2.1. By the Taylor expansion,
\[
\zeta_i(\theta_i) = \nabla \zeta_i(\theta_{0i})^T (\theta_i - \theta_{0i}) + \frac{1}{2} (\theta_i - \theta_{0i})^T \nabla^2 \zeta_i(\theta_{0i})(\theta_i - \theta_{0i}) + o(\| \theta_i - \theta_{0i} \|_{2}), \quad \text{as } \theta_i \to \theta_{0i},
\]
91
where $\nabla \zeta_i$ and $\nabla^2 \zeta_i$ are the gradient vector and Hessian matrix of $\zeta_i$, respectively. By differentiating $\log(1 - \zeta_i(x))$ with respect to $x \in \mathbb{R}^{m-1}$ using the chain rule, it can be checked that $\nabla \zeta_i(\theta_0) = 0$ and $\nabla^2 \zeta_i(\theta_0) = W_i/4$ for every $i \leq n$. Since $\zeta_i(\theta_i) \leq M_n^{-1}$ implies $\theta_i \rightarrow \theta_0i$ by Lemma 3.7, plugging in $\nabla \zeta_i(\theta_0)$ and $\nabla^2 \zeta_i(\theta_0)$, the Taylor expansion implies that there exists a constant $C_1 > 0$ such that if $M_n^{-1} \geq C_1 \zeta_i(\theta_i)$,

$$M_n^{-1} \geq C_1 \zeta_i(\theta_i) \geq \frac{C_1}{16} \|W_i^{1/2}(\theta_i - \theta_0)\|^2 = \frac{C_1}{16} \|W_i^{1/2}(\theta_i - \theta_0)\|^2 \land M_n^{-1},$$

while if $C_1 \zeta_i(\theta_i) > M_n^{-1}$,

$$C_1 \zeta_i(\theta_i) \geq M_n^{-1} \geq \frac{C_1}{16} \|W_i^{1/2}(\theta_i - \theta_0)\|^2 \land M_n^{-1}.$$

This means that we have $\zeta_i(\theta_i) \geq 16^{-1} \|W_i^{1/2}(\theta_i - \theta_0)\|^2 \land (C_1 M_n)^{-1}$ for both cases. Define $I_{n,M_n} = \{i \leq n : C_1 \|W_i^{1/2}(\theta_i - \theta_0)\|^2 \geq 16 M_n^{-1}\}$ and let $|I_{n,M_n}|$ be its cardinality. Then,

$$\epsilon_n^2 \gtrsim \frac{H_n^2(\beta, \beta_0) \geq 2 |I_{n,M_n}|}{C_1 M_n n} + \frac{1}{8n} \sum_{i \notin I_{n,M_n}} \|W_i^{1/2}(\theta_i - \theta_0)\|^2, \quad (3.14)$$

which particularly implies that $n \epsilon_n^2 \gtrsim |I_{n,M_n}|/M_n$. Hence, using the definition of $\psi_{2,1}$, we note that for some constant $C_2 > 0$ and every $\beta$ such that $s_\beta \leq K_1 s_0$,

$$\frac{1}{n} \sum_{i \notin I_{n,M_n}} \|W_i^{1/2}(\theta_i - \theta_0)\|^2 \geq \frac{1}{n} \sum_{i = 1}^n \|W_i^{1/2}(\theta_i - \theta_0)\|^2 - \frac{|I_{n,M_n}|}{n} \max_{1 \leq i \leq n} \|W_i^{1/2}(\theta_i - \theta_0)\|^2 \geq \frac{\psi_{2,1}^2((K_1 + 1)s_0)}{(K_1 + 1)s_0 n} \|X\|^2 - \frac{M_n \epsilon_n^2}{C_2} \max_{1 \leq i \leq n} \|X_i\|^2 \|\beta - \beta_0\|^2_{2,1},$$

since $|S_{\beta - \beta_0}| \leq s_\beta + s_0 \leq (K_1 + 1)s_0$ for such $\beta$. The condition (C2) implies that the first term of the rightmost side dominates the second term, so the last display is bounded below by a constant multiple of $\psi_{2,1}^2((K_1 + 1)s_0)) \|X\|^2 \|\beta - \beta_0\|^2_{2,1}/(s_0 n)$. This implies that we have

$$\|\beta - \beta_0\|^2_{2,1} \lesssim \frac{s_0 \sqrt{\log p} \sqrt{\log n}}{\psi_{2,1}((K_1 + 1)s_0)) \|X\|_0}, \quad (3.15)$$

which verifies the third assertion of the theorem. Note also that by (3.14),

$$\epsilon_n^2 \gtrsim H_n(\beta, \beta_0) \geq \frac{1}{8n} \|W^{1/2}X(\beta - \beta_0)\|^2 \geq \frac{1}{8n} \sum_{i \in I_{n,M_n}} \|W_i^{1/2}X_i(\beta - \beta_0)\|^2.$$
Thus, we have that for some $C_3 > 0$,

$$\|W^{1/2}X(\beta - \beta_0)\|_2^2 \lesssim n\epsilon_n^2 + \|I_{n,M_n}\| \max_{1 \leq i \leq n} \|W_i^{1/2}X_i(\beta - \beta_0)\|_2^2$$

$$\leq n\epsilon_n^2 + C_3n\epsilon_n^2 \max_{1 \leq i \leq n} \|X_i\|_2^2 \|\beta - \beta_0\|^{2,1},$$

where the second inequality follows from (3.14). By combining (C2) and (3.15), the last display is bounded by a multiple of $n\epsilon_n^2$, which concludes the first assertion of the theorem. The second assertion then follows from the definition of $\psi_2$.

To proof Theorem 3.4 for posterior contraction with general prediction, we give the following two lemmas. The first one proves the self-concordant property of the multinomial logit model (Bach, 2010), which is used to construct suitable tests in the second lemma. The second lemma directly constructs test functions for the $\ell_2,1$- and $\ell_2$-norms.

**Lemma 3.8.** For any $v = (v_1, \ldots, v_{m-1})^T \in \mathbb{R}^{m-1}$ and $w = (w_1, \ldots, w_{m-1})^T \in \mathbb{R}^{m-1}$, there exists some constant $R > 0$ such that the function $\eta(t) := \log(1 + h(w + tv)T1_{m-1})$ satisfies $|\eta''(t)| \leq R\|v\|_{2}\eta''(t)$ for every $t > 0$.

**Lemma 3.9.** Let $\bar{E}_n$ be the complement of the event in (3.10) and define $q_\beta(y) := 1_{\bar{E}_n}(y)f_\beta(y)$ for any $y$ in the sample space. Also define

$$Q_1 := \left\{ q_\beta : s_\beta \leq K_1s_0, \|\beta - \beta_1\|_2 \leq \frac{\delta \sqrt{n}}{2\psi_2^2((K_1 + 1)s_0)\|X\|_0} \right\},$$

for a given $\beta_1$ such that $\|\beta_1 - \beta_0\|_2 > \delta \sqrt{n}/[\psi_2^2((K_1 + 1)s_0)\|X\|_0]$. Then for every $\delta > 16C* \sqrt{K_1 + 1}\epsilon_n$ and $\beta_0$ satisfying (C2), there exists a test function $\varphi_n^*$ such that

$$E_0\varphi_n^* \leq \exp(-n\delta^2/64), \quad \sup_{q \in Q_1} \int (1 - \varphi_n^*)qd\nu_n \leq \exp(-n\delta^2/64),$$

where $\nu_n$ is a dominating $n$-fold product counting measure. Similarly, define

$$\bar{Q}_1 := \left\{ q_\beta : s_\beta \leq K_1s_0, \|\beta - \beta_1\|_{2,1} \leq \frac{\delta \sqrt{(K_1 + 1)s_0n}}{2\psi_2^{2,1}((K_1 + 1)s_0)\|X\|_0} \right\},$$

for a given $\beta_1$ such that $\|\beta_1 - \beta_0\|_{2,1} > \delta \sqrt{(K_1 + 1)s_0n}/[\psi_2^{2,1}((K_1 + 1)s_0)\|X\|_0]$. Then for every $\delta > 16C* \sqrt{K_1 + 1}\epsilon_n$ and $\beta_0$ satisfying (C2), there also exists a test function $\varphi_n^*$ such that

$$E_0\varphi_n^* \leq \exp(-n\delta^2/64), \quad \sup_{q \in \bar{Q}_1} \int (1 - \varphi_n^*)qd\nu_n \leq \exp(-n\delta^2/64).$$
Proof of Theorem 3.4. The theorem can be proved very similar to Theorem 3.2. We only need to evaluate the estimate of the entropy to obtain globally powerful tests from the local tests in Lemma 3.9. We show more details on the first assertion. The second assertion can be verified similarly.

Let \( A_n' = \{ \beta \in \mathbb{R}^d : s_\beta \leq K_1s_0 \} \) and \( A_n := \{ \beta \in A_n' : \|\beta - \beta_0\|_\infty \leq D_n \} \) for \( D_n := n\sqrt{q}(p^{1/q} \vee n)^{1/2}\|X\|_0^{-1} \). Using (3.11), it is easy to see that for some \( C_1 > 0 \),

\[
\sup_{\delta > 16C_1\sqrt{K_1 + 1}\epsilon_n} N\left( \frac{\delta}{128}, \mathcal{A}_n^*, \| \cdot \|_2 \right) \leq N\left( \frac{C_1\sqrt{K_1 + 1}}{\sqrt{80q}}, \mathcal{A}_n^*, \| \cdot \|_\infty \right) \leq e^{C_1n\epsilon_n^2}.
\]

Therefore, similar to Lemma 9 of Ghosal and van der Vaart (2007) and Theorem 6.2 of Kleijn and van der Vaart (2006), there exists a test \( \bar{\varphi}_n^* \) such that for every \( \delta > 16C_1\sqrt{K_1 + 1}\epsilon_n \),

\[
E_0\bar{\varphi}_n^* \leq \frac{1}{2} \exp(C_1n\epsilon_n^2 - n\delta^2/64), \quad \sup_{q \in \mathcal{Q}_0} \int (1 - \bar{\varphi}_n^*) q d\nu_n \leq \exp(-n\delta^2/64),
\]

for sufficiently large \( n \), where \( \mathcal{Q}_0 = \{ q_\beta : \beta \in \mathcal{A}_n^*, s_\beta \leq K_1s_0, \|\beta - \beta_0\|_2 > \delta\sqrt{n}/|\psi_2^2((K_1 + 1)s_0)\|X\|_0 \} \). Note that we can restrict our attention to the set \( \mathcal{A}_n^* \) due to Theorem 3.1. Now, to complete the proof of the first assertion, observe that \( E_0\Pi(\beta \in A_n' : \|\beta - \beta_0\|_2 > \delta\sqrt{n}/|\psi_2^2((K_1 + 1)s_0)\|X\|_0 | Y^{(n)}) \) is bounded by

\[
E_0\Pi\left( \beta \in A_n' : \|\beta - \beta_0\|_2 > \frac{\delta\sqrt{n}}{|\psi_2^2((K_1 + 1)s_0)\|X\|_0} \bigg| Y^{(n)} \right) \mathbb{1}_{\mathcal{E}_n} + \mathbb{P}_0\bar{\mathcal{E}}_n^c,
\]

where the second term goes to zero by Lemma 3.5. The first term is bounded by

\[
E_0\Pi\left( \beta \in A_n' : \|\beta - \beta_0\|_2 > \frac{\delta\sqrt{n}}{|\psi_2^2((K_1 + 1)s_0)\|X\|_0} \bigg| Y^{(n)} \right) \mathbb{1}_{\mathcal{E}_n} \mathbb{1}_{\mathcal{E}_n}(1 - \bar{\varphi}_n^*) + E_0\bar{\varphi}_n^* + \mathbb{P}_0\mathcal{E}_n^c
\]

\[
\leq \left\{ \sup_{q \in \mathcal{Q}_0} \int (1 - \bar{\varphi}_n^*) q d\nu_n + \Pi(A_n' \setminus A_n^*) \right\} e^{C_2n\epsilon_n^2} + E_0\bar{\varphi}_n^* + \mathbb{P}_0\mathcal{E}_n^c,
\]

for some \( C_2 > 0 \). Choosing \( \delta = M\epsilon_n \) for a sufficiently large \( M \), the display tends to zero, which verifies the first assertion.

For the second assertion, observe that there exists a test \( \bar{\varphi}_n^* \) such that for every \( \delta > 16C_1\sqrt{K_1 + 1}\epsilon_n \),

\[
E_0\bar{\varphi}_n^* \leq \frac{1}{2} \exp(C_1n\epsilon_n^2 - n\delta^2/64), \quad \sup_{q \in \mathcal{Q}_0} \int (1 - \bar{\varphi}_n^*) q d\nu_n \leq \exp(-n\delta^2/64),
\]

through a similar entropy calculation, where \( \bar{\mathcal{Q}}_0 = \{ q_\beta : \beta \in \mathcal{A}_n^*, s_\beta \leq K_1s_0, \|\beta - \beta_0\|_{2,1} > \delta\sqrt{(K_1 + 1)s_0}/|\psi_{2,1}^2((K_1 + 1)s_0)\|X\|_0 \} \). The rest of the proof is then very similar to that of
the first assertion by choosing $\delta = M\epsilon_n$ for a sufficiently large $M$.

**Proof of Theorem 3.5.** We shall use the fact that for any probability measure $Q$ and its renormalized restriction $Q_\mathcal{F}$ to a set $\mathcal{F}$, we have $\|Q - Q_\mathcal{F}\|_{\text{TV}} \leq 2Q(\mathcal{F})$. Let $\widetilde{\Pi}(\cdot)$ be the prior distribution restricted to $\mathcal{A}_n$ defined in Lemma 3.2 and $\Pi(\cdot | Y^{(n)})$ be the corresponding posterior. Similarly, let $\widetilde{\Pi}_\infty(\cdot | Y^{(n)})$ be the renormalized restriction of $\Pi_\infty(\cdot | Y^{(n)})$ to the set $\mathcal{A}_n$.

Then, the total variation metric in the theorem is bounded above by

$$
\left\| \Pi(\cdot | Y^{(n)}) - \widetilde{\Pi}(\cdot | Y^{(n)}) \right\|_{\text{TV}} + \left\| \widetilde{\Pi}(\cdot | Y^{(n)}) - \widetilde{\Pi}_\infty(\cdot | Y^{(n)}) \right\|_{\text{TV}}
$$

Using the argument given at the beginning of the proof, the expected value of the first summand of the display goes to zero by Theorem 3.3.

To show that the second summand of (3.16) tends to zero, observe that

$$
\widetilde{\Pi}(\beta \in \mathcal{B} | Y^{(n)}) \propto \int_{\mathcal{B} \cap \mathcal{A}_n} \Lambda_n(\beta)e^{-\sum_{j=1}^p \lambda_j \|\beta_j\|^2} dV(\beta)
$$

$$
\widetilde{\Pi}_\infty(\beta \in \mathcal{B} | Y^{(n)}) \propto \int_{\mathcal{B} \cap \mathcal{A}_n} \Lambda_n^*(\beta)e^{-\sum_{j=1}^p \lambda_j \|\beta_{0,j}\|^2} dV(\beta),
$$

where $dV(\beta) = \sum_{S: S \leq K_1 s_0} \pi_p(s)^{(p)}(s)^{-1} \{ \prod_{j \in S} (\lambda_j/a_j)^{q_j} \} d\{ L(\beta_S) \otimes \delta_0(\beta_S) \}$. In the second line, the factor $e^{-\sum_{j=1}^p \lambda_j \|\beta_{0,j}\|^2}$ cancels out in the normalizing constant, but is inserted to compare the two measures. Note that for any sequences of measures $\{\mu_S\}$ and $\{\nu_S\}$,

$$
\left\| \frac{\sum_S \mu_S}{\|\sum_S \mu_S\|_{\text{TV}}} - \frac{\sum_S \nu_S}{\|\sum_S \nu_S\|_{\text{TV}}} \right\|_{\text{TV}} \leq \frac{2 \sup_S \left\| \frac{\mu_S - \nu_S}{\|\sum_S \mu_S\|_{\text{TV}}} \right\|}{\|\sum_S \mu_S\|_{\text{TV}}} \leq 2 \sup_S \left\| 1 - \frac{d\nu_S}{d\mu_S} \right\|_{\infty},
$$

if for every $S$, $\nu_S$ is absolutely continuous with respect to $\mu_S$ with Radon-Nikodym derivative $d\nu_S/d\mu_S$. Therefore, we obtain

$$
\left\| \Pi(\cdot | Y^{(n)}) - \widetilde{\Pi}_\infty(\cdot | Y^{(n)}) \right\|_{\text{TV}} \leq 2 \sup_{\beta \in \mathcal{A}_n} \left| 1 - \exp \left\{ \log \Lambda_n(\beta) - \log \Lambda^*_n(\beta) - \sum_{j=1}^p \lambda_j (\|\beta_j\|^2 - \|\beta_{0,j}\|^2) \right\} \right|.
$$

The right hand side can be further bounded using the inequality $|e^x - 1| \leq e|x| - 1 \leq e|x| |x|$ for every $x$. Hence the display tends to zero by Lemma 3.2 and the relation $\bar{\lambda}s_0 \sqrt{\log p} \sqrt{q / \log n / \|X\|_1} \to 0$, which is satisfied by (3.5) and (C5) (see the paragraph followed by Lemma 3.2).

Now, we shall show that the expected value of the third summand in (3.16) goes to zero.
Observe that $\Pi^{\infty}(\beta \in \mathcal{A}_{s_0}' | Y^{(n)})$ is equal to

$$\frac{\int_{\mathcal{A}_{s_0}'} N_n(\beta) dV(\beta)}{\int_{\mathbb{R}^d} N_n(\beta) dV(\beta)}.$$  \hspace{1cm} (3.17)

It is easy to see that the denominator is bounded below by

$$\frac{\pi_p(s_0)}{\prod_{j \in S_0} a_j^{q_j}} \int_{\mathbb{R}^d_0} \exp \left\{ U^T W^{1/2} X_{s_0}(\beta_{s_0} - \beta_{0,s_0}) - \frac{1}{2} (\beta_{s_0} - \beta_{0,s_0})^T X_{s_0}^T W X_{s_0}(\beta_{s_0} - \beta_{0,s_0}) \right\} d\beta_{s_0}. \hspace{1cm} (3.18)$$

Note that for $\Gamma_{s_0} = X_{s_0}^T W X_{s_0}$, the integral in the last display equals

$$\frac{(2\pi)^{d_0/2}}{\det(\Gamma_{s_0})^{1/2}} \exp \left\{ \frac{1}{2} U^T W^{1/2} X_{s_0} \Gamma_{s_0}^{-1} X_{s_0}^T W^{1/2} U \right\} \geq \frac{(2\pi)^{d_0/2}}{\det(\Gamma_{s_0})^{1/2}}.$$  \hspace{1cm} (3.19)

Applying the arithmetic-geometric mean inequality to the eigenvalues, we see that

$$\det(\Gamma_{s_0}) \leq \left( \frac{\text{tr}(\Gamma_{s_0})}{d_0} \right)^d \leq \|W^{1/2} X_{s_0}\|^2 \leq \rho_{\text{max}}(W)^{d_0} \|X\|^2,$$

where $\|\cdot\|_*$ is the maximum diagonal entry of the Gram matrix as defined in Section 3.2.1. Since $\|X\|_* \leq \|X\|_2$, using (3.5) and Lemma 3.6, the display is further bounded by $(L^2 \lambda^2)^{d_0} (p^{1/\sqrt{d}} \sqrt{d})^{d_0}$. Hence, by putting everything together with the inequalities $\pi_p(s_0) \gtrsim A_1 s_0 / (p \sqrt{d})$ and $(p s_0) \leq p s_0$, (3.18) is further bounded below by a constant multiple of

$$\exp \left\{ s_0 (\log A_1 - \log p) - (A_3 + L_2) s_0 (\log \sqrt{d}) - C_1 d_0 \log \sqrt{d} \right\},$$

for some $C_1 > 0$, since $\sum_{j \in S_0} q_j \log a_j \lesssim d_0 \log d$ as $a_j = O(\sqrt{d})$ by Stirling’s approximation.

Next, we establish an upper bound of the numerator of (3.17). Due to Lemma 3.5, we can restrict our attention to $\mathcal{E}_n$ the complement of the event in (3.10). For $D_n := C_n \|X\|_1 \log p \sqrt{d} \log n$ with $C_n$ in Lemma 3.5, it follows that on the event $\mathcal{E}_n$,

$$U^T W^{1/2} X (\beta - \beta_0) = \sum_{j=1}^p U^T W^{1/2} X_j (\beta_j - \beta_{0,j}) \leq D_n \|\beta - \beta_0\|_{2,1} \leq D_n \frac{2 \|X(\beta - \beta_0)\|_{2,1} |S_{\beta - \beta_0}|^{1/2}}{\|X\|_1 \phi_{2,1}(|S_{\beta - \beta_0}|)} - D_n \|\beta - \beta_0\|_{2,1}.$$  \hspace{1cm} (3.19)

96
Since $\rho_{\min}(W)$ is bounded away from zero by (C3) and Lemma 3.6, for some $C_2 > 0$, the display is further bounded above by

$$\begin{align*}
D_n \frac{2C_2 ||W^{1/2}X(\beta - \beta_0)||_2 |S_{\beta - \beta_0}|^{1/2}}{||X||_0^2 \phi_{2,1}(|S_{\beta - \beta_0}|)} - D_n ||\beta - \beta_0||_{2,1} \\
\leq \frac{1}{2} ||W^{1/2}X(\beta - \beta_0)||_2^2 + \frac{2C_2^2 D_n^2 |S_{\beta - \beta_0}|}{||X||_0^2 \phi_{2,1}(|S_{\beta - \beta_0}|)^2} - D_n ||\beta - \beta_0||_{2,1},
\end{align*}$$

where the inequality holds by the Cauchy-Schwartz inequality. Hence, on the event $\bar{E}_n$, the numerator of (3.17) is bounded by

$$\exp \left\{ \frac{2C_2^2 (K_1 + 1) D_n^2 s_0}{||X||_0^2 \phi_{2,1}((K_1 + 1)s_0)^2} \right\} \sum_{S: s \leq K_1 s_0} \frac{\pi_p(s)}{\prod_{j \in S} \frac{a_j^p}{d_j^s}} \frac{X^{d_s}}{\prod_{j \in S} a_j^s} \int_{\mathbb{R}^d} e^{-D_n ||\beta - \beta_0||_{2,1}} d\{L(\beta_S) \otimes \delta_0(\beta_{S^c})\}$$

$$\leq \exp \left\{ \frac{2C_2^2 (K_1 + 1) D_n^2 s_0}{||X||_0^2 \phi_{2,1}((K_1 + 1)s_0)^2} - \frac{D_n}{2} \inf_{\beta \in A_0 \cap \{\beta: s_\beta \leq K_1 s_0\}} ||\beta - \beta_0||_{2,1} \right\}$$

$$\times \sum_{S: s \leq K_1 s_0} \frac{\pi_p(s)}{\prod_{j \in S} \frac{a_j^p}{d_j^s}} \frac{X^{d_s}}{\prod_{j \in S} a_j^s} \int_{\mathbb{R}^d} e^{-D_n/2} ||\beta - \beta_{0,S}||_{2,1} d\beta_S.$$ 

Since $D_n = \frac{1}{2} \geq C_* \mathcal{X}_r/(2L_3)$ and $A_0 \cap \{\beta \in \mathbb{R}^{d_*} : s_\beta \leq K_1 s_0\} = \{\beta \in \mathbb{R}^{d_*} : s_\beta \leq K_1 s_0, ||\beta - \beta_0||_{2,1} > M s_0 \sqrt{\log p \vee \sqrt{p \log n}}/||X||_0\}$, the right hand side of the display is further bounded by

$$\exp \left\{ \frac{2C_2^2 C_2^2 (K_1 + 1)s_0(\log p \vee \sqrt{p \log n})}{\phi_{2,1}((K_1 + 1)s_0)^2} - \frac{M C_*}{2} s_0(\log p \vee \sqrt{p \log n}) \right\}$$

$$\times \sum_{S: s \leq K_1 s_0} \frac{\pi_p(s)}{\prod_{j \in S} \frac{a_j^p}{d_j^s}} \left( \frac{4L_3/C_*}{A_2} \right) \frac{X^{d_s}}{\prod_{j \in S} a_j^s}.$$ 

Recall that $a_j \geq 2$ for every $j \leq p$. The summation in the last display is thus bounded by

$$\sum_{s=0}^p \pi_p(s) \left( 1 \vee \frac{2L_3}{C_*} \right)^{s \gamma} \leq \sum_{s=0}^p \left( 1 \vee \frac{2L_3}{C_*} \right)^{\frac{\gamma}{\frac{1}{p}} \frac{A_2}{(p \vee \sqrt{p} \vee \gamma) d_A^s}} \leq \left( 1 \vee \frac{2L_3}{C_*} \right)^{\frac{\gamma}{\frac{1}{p}} \frac{A_2}{(p \vee \sqrt{p} \vee \gamma) d_A^s}},$$

which is bounded for large enough $n$ since the term $1 \vee (2L_3/C_*)$ is dominated by $n^{A_1}$. Note also that $\phi_{2,1}((K_1 + 1)s_0)$ is bounded away from zero by the assumption. Therefore, combining (3.19) and (3.20), we see that (3.17) tends to zero on the event $\bar{E}_n$ as long as $M$ is chosen to be sufficiently large.

Proof of Theorem 3.6. Since $E_0 ||\Pi(\beta \in \cdot | Y^{(n)}) - \Pi^\infty(\beta \in \cdot | Y^{(n)})||_{TV}$ tends to zero by Theorem 3.5, it suffices to show that $E_0 \Pi^\infty(\beta : S_\beta \in S_n | Y^{(n)}) \to 0$ for $S_n = \{S : s \leq K_1 s_0, S \supset S_0, S \neq S_0\}$. Define the orthogonal projection $H_S = W^{1/2}X^S \Gamma_S^{-1} X^S W^{1/2}$ for $\Gamma_S = X^S W X_S$. 97
Then, we see that by \((3.9)\), \(\Pi^\infty(\theta : \mathcal{S}_\beta \in \mathcal{S}_n \mid Y^{(n)})\) is bounded by

\[
\sum_{s=s_0+1}^{K_1s_0} \frac{\pi_p(s)(p_{s_0}^\beta p_{s_0}^\theta)^{p_{s_0}-s_0}}{\pi_p(s_0)^{p_s}} \max_{S \in \mathcal{S}_n : |S| = s} \left[ \frac{(X_2 n)^{d_S-d_0}}{\prod_{j \in S \setminus S_0} a_j^{q_j}} \det(\Gamma_{S_0})^{1/2} e^{\|H_S - H_{S_0}\|_2^2/2} \right],
\tag{3.21}
\]

since \((H_S - H_{S_0})W^{1/2}X_{\beta_0} = (H_S - H_{S_0})W^{1/2}X_{\beta_0S_0} = 0\) for every \(S \in \mathcal{S}_n\). Let \(\rho_k(\cdot)\), \(k = 1, 2, \ldots\), stand for eigenvalues of a matrix in decreasing order. Since \(\Gamma_{S_0}\) is a principal submatrix of \(\Gamma_S\) for every \(S \in \mathcal{S}_n\), we have that \(\rho_k(\Gamma_{S_0}) \leq \rho_k(\Gamma_S)\) for \(k = 1, \ldots, d_0\). Therefore, \(\det(\Gamma_{S_0})\) is equal to

\[
\prod_{k=1}^{d_0} \rho_k(\Gamma_{S_0}) \leq \prod_{k=1}^{d_0} \rho_k(\Gamma_S) \leq \frac{\det(\Gamma_S)}{\rho_{\min}(\Gamma_S)^{d_S-d_0}} \leq \frac{\det(\Gamma_S)}{(\rho_{\min}(W)\phi_2^2(s))^{\|X\|_2^2}^{d_S-d_0}},
\]

where the last inequality holds since \(\rho_{\min}(\Gamma_S) \geq \rho_{\min}(W)\rho_{\min}(X_S^TX_S) \geq \rho_{\min}(W)\phi_2^2(s)\|X\|_2^2\) by the definition of \(\phi_2\).

Now, we show that for any fixed \(b > 2B_\ast\),

\[
\mathbb{P}_0\left(\|H_S - H_{S_0}\|_2^2 \leq b(s-s_0)(\log p \lor q \log n), \text{ for every } S \in \mathcal{S}_n\right) \to 1.
\tag{3.22}
\]

By Markov’s inequality, for every \(u > 0\),

\[
\mathbb{P}_0\left(\max_{S \in \mathcal{S}_n : |S| = s} \|H_S - H_{S_0}\|_2^2 > b(s-s_0)(\log p \lor q \log n)\right)
\leq e^{-ub(s-s_0)(\log p \lor q \log n)} \mathbb{E}_0 \max_{S \in \mathcal{S}_n : |S| = s} \exp \left(\frac{u\|H_S - H_{S_0}\|_2^2}{2}\right)
\leq e^{-ub(s-s_0)(\log p \lor q \log n)} N_s \max_{S \in \mathcal{S}_n : |S| = s} \mathbb{E}_0 \exp \left(\frac{u\|H_S - H_{S_0}\|_2^2}{2}\right),
\]

where \(N_s = \binom{p-s_0}{s-s_0}\) is the cardinality of the set \(\{S \in \mathcal{S}_n : |S| = s\}\). Note that \(W_i^{-1} = \text{diag}(1/\mu_i) + \mu_{i0}^{-1}1_{m-1}^T1_{m-1}^T\) by the Sherman-Morrison formula and that \(Y_{\ell t} \in \{0, 1\}, \ell = 1, \ldots, m-1,\) such that \(\sum_{t=1}^{m-1} Y_{\ell t} = 0\) or \(\sum_{t=1}^{m-1} Y_{\ell t} = 1\). Hence, by direct calculations,

\[
\max_{\xi, \xi' \in \supp(U_i)} \|\xi - \xi'\|_2^2 = \max_{y, y' \in \supp(Y_i)} \|W_i^{-1/2}(y - y')\|_2^2 = \max_{0 \leq j \neq k \leq m-1} (1/\mu_{ij} + 1/\mu_{ik}).
\]

Note also that \(\max_{\xi \in \supp(U_i)} \|\xi\|_2^2\) is bounded due to (C3), Lemma 3.6, and the support of \(Y_i\). Plugging in these values, Lemma 3.4 gives that for some \(C_1 \geq 0\) and any \(u \in (0, (2B_\ast)^{-1})\),

\[
\max_{S \in \mathcal{S}_n : |S| = s} \mathbb{E}_0 \exp \left(\frac{u\|H_S - H_{S_0}\|_2^2}{2}\right) \leq \exp \left(\frac{uC_1q(s - s_0)}{1 - 2uB_\ast}\right),
\tag{3.23}
\]

since \(\text{tr}(H_S - H_{S_0}) = d_S - d_0 \leq q(s - s_0)\) and \(\|H_S - H_{S_0}\|_{sp} \leq 1\). By the inequality \(N_s \leq
If \((p - s_0)^{s - s_0} \leq p^{s - s_0}\), we see that for \(T_n\) the event in the relation (3.22), \(\mathbb{P}_0(T_n^c)\) is bounded by

\[
\sum_{s = s_0 + 1}^{K_1 s_0} \exp \left\{ -ub(s - s_0)(\log p \vee q \log n) + (s - s_0) \log p + \frac{uC_1 q(s - s_0)}{1 - 2uB_n} \right\}.
\]

This goes to zero as \(n, p \to \infty\) as long as \(ub > 1\) since

\[
\log p + \frac{uC_1 q}{1 - 2uB_n} \leq \left( 1 + \frac{uC_1}{(1 - 2uB_n) \log n} \right) (\log p \vee q \log n).
\]

Hence we verify (3.22) by combining the condition \(ub > 1\) with the range of \(u\) for (3.23).

Combining (3.21) and (3.22), we see that for some \(s_0 \leq s \leq K_1 s_0\), we also have that

\[
\sum_{s = s_0 + 1}^{K_1 s_0} \frac{\pi_p(s) \left( \frac{p}{s_0} \right)^{s - s_0} \left( \frac{p}{s_0} \right)^{s_0}}{\pi_p(s_0) \left( \frac{p}{s_0} \right)^{s_0}} \leq e^{b(s - s_0)(\log p \vee q \log n)/2} \max_{S \in S_n, |S| = s} \left( \frac{C_2 \sqrt{2\pi}}{\phi_2(s) \|X\|_{\circ}} \right)^{d_s - d_0} \left( \frac{\sqrt{2\pi}}{\phi_2(s_0) \|X\|_{\circ}} \right)^{d_0},
\]

since \(a_j \geq 2\) for every \(j \leq p\). Note that

\[
\frac{\pi_p(s) \left( \frac{p}{s_0} \right)^{s - s_0} \left( \frac{p}{s_0} \right)^{s_0}}{\pi_p(s_0) \left( \frac{p}{s_0} \right)^{s_0}} \leq A_2(s - s_0)(p \vee n)^{-A_4(s - s_0)} \quad \text{and} \quad \left( \frac{p}{s_0} \right)^{s - s_0} / \left( \frac{p}{s_0} \right)^{s_0} = \left( \frac{s}{s_0} \right).
\]

Note also that for every \(S\) such that \(s_0 + 1 \leq s \leq K_1 s_0\),

\[
\left( \frac{C_2 \sqrt{2\pi}}{\phi_2(s) \|X\|_{\circ}} \right)^{d_s - d_0} \leq \left( \frac{C_2 \sqrt{2\pi}}{\phi_2(s_0) \|X\|_{\circ}} \right)^{d_s - d_0} \leq \left( \frac{C_2 L_3 \sqrt{2\pi K_1 s_0}}{\phi_2(s_0) \|X\|_{\circ}} \right)^{d_s - d_0},
\]

which is bounded for sufficiently large \(n\) since \(\phi_2(s_0)\) is bounded away from zero by (C4).

For any \(s \leq K_1 s_0\), we also have that \(\left( \frac{s}{s_0} \right) = \left( \frac{s}{s_0} \right) \leq (K_1 s_0)^{s - s_0} \leq (C_3 p)^{s - s_0}\) for some \(C_3 > 0\). Hence, \(\Pi^\infty(\beta : S_\beta \in S_n \mid Y^{(n)}) \mathbb{1}_{T_n}\) goes to zero provided that \(a - A_4 + b/2 < 0\). This is translated to \(a < A_4 - B_n\) by choosing \(b\) arbitrarily close to \(2B_n\). \(\square\)

### 3.5.3 Proofs for the lemmas

**Proof of Lemma 3.1.** If \(s_0 = 0\), it follows that \(\int \Lambda_n(\beta) d\Pi(\beta) \geq \Lambda_n(0) \pi_p(0) = \pi_p(0)\) which directly gives the desired result. For the case \(s_0 > 0\), let \(\epsilon_n = \sqrt{s_0(\log p \vee q \log n)/n}\) and

\[
B_n = \left\{ \beta \in \mathbb{R}^d : \frac{1}{n} \sum_{i=1}^n K(f_{0,i}, f_{\beta,i}) \leq \epsilon_n^2, \frac{1}{n} \sum_{i=1}^n V(f_{0,i}, f_{\beta,i}) \leq \epsilon_n^2 \right\},
\]

99
where $K$ and $V$ denote the Kullback-Leibler divergence and variation, respectively. By Lemma 10 of Ghosal and van der Vaart (2007), we have that for any $C > 0$,

$$
\mathbb{P}_0 \left( \int_{B_n} \Lambda_n(\beta) d\Pi(\beta) \geq e^{-(1+C)n\epsilon_2^2} \Pi(B_n) \right) \geq 1 - \frac{1}{C^2 n\epsilon_2^2}.
$$

Hence, it suffices to show that $\Pi(B_n)$ is bounded below as in the lemma. By direct calculations, the Kullback-Leibler divergence and variation are then given by

$$
K(f_{0,i}, f_{\beta,i}) := \mathbb{E}_0 \log(f_{0,i}/f_{\beta,i})
= \log(1 + h(\theta_i)^T 1_{m-1}) - \log(1 + h(\theta_{0i})^T 1_{m-1}) - \mu_i^T(\theta_i - \theta_{0i}),
$$
and

$$
V(f_{0,i}, f_{\beta,i}) := \mathbb{E}_0 \left\{ (\log(f_{0,i}/f_{\beta,i}) - K(f_{0,i}, f_{\beta,i}))^2 \right\}
= (\theta_i - \theta_{0i})^T W(\theta_i - \theta_{0i}).
$$

By Taylor’s theorem applied to $\log(1 + h(\theta_i)^T 1_{m-1})$ at $\log(1 + h(\theta_{0i})^T 1_{m-1})$, we obtain that for $\tilde{\theta}_i = t\theta_{0i} + (1 - t)\theta_i$ with some $t \in (0, 1)$,

$$
K(f_{0,i}, f_{\beta,i}) = \frac{1}{2}(\theta_i - \theta_{0i})^T \left\{ \text{diag}(\bar{\mu}_i) - \bar{\mu}_i \bar{\mu}_i^T \right\} (\theta_i - \theta_{0i}),
$$
(3.24)

where $\bar{\mu}_i = (1 + h(\tilde{\theta}_i)^T 1_{m-1})^{-1}h(\tilde{\theta}_i)$. Since the spectral norm of the covariance matrix of a multinomial distribution is bounded by 1 (Lemma 3.6), both $K(f_{0,i}, f_{\beta,i})$ and $V(f_{0,i}, f_{\beta,i})$ are bounded by $\|\theta_i - \theta_{0i}\|_2^2$. Thus, by the inequality $\|X(\beta - \beta_0)\|_2 \leq \|X\|_o \|\beta - \beta_0\|_{2,1} \leq \|X\|_o \|\beta - \beta_0\|_1$, we obtain that

$$
\Pi(B_n) \geq \Pi\{ \beta \in \mathbb{R}^{d_r} : n^{-1/2}\|X\|_o \|\beta - \beta_0\|_1 \leq \epsilon_n \}.
$$

For any $s_0 > 0$ and $\tilde{B}_n := \{ \beta_{S_0} \in \mathbb{R}^{d_0} : n^{-1/2}\|X\|_o \|\beta_{S_0} - \beta_{0,S_0}\|_1 \leq \epsilon_n \}$, it is easy to see that the last display is bounded below by

$$
\frac{\pi_p(s_0)}{(p_0)} \int_{\tilde{B}_n} g_{S_0}(\beta_{S_0}) d\beta_{S_0} \geq \frac{\pi_p(s_0)}{(p_0)} e^{-\lambda\|\beta_0\|_{2,1}} \int_{\tilde{B}_n} g_{S_0}(\beta_{S_0} - \beta_{0,S_0}) d\beta_{S_0},
$$
(3.25)

since $g_{S_0}(\beta_{S_0}) \geq e^{-\lambda\|\beta_0\|_{2,1}} g_{S_0}(\beta_{S_0} - \beta_{0,S_0})$. Observe that

$$
\int_{\tilde{B}_n} g_{S_0}(\beta_{S_0} - \beta_{0,S_0}) d\beta_{S_0} = \int_{\tilde{B}_n} \prod_{j \in S_0} \left\{ \left( \frac{\lambda_j}{a_j} \right)^{q_j} e^{-\lambda_j\|\beta_j - \beta_{0,j}\|^2} \right\} d\beta_{S_0}
\geq \left( \frac{2\lambda/\lambda_j}{\prod_{j \in S_0} a_j^{q_j}} \right)^{d_0} \int_{\tilde{B}_n} \left( \frac{\lambda}{2} \right)^{d_0} e^{-\lambda\|\beta_{S_0} - \beta_{0,S_0}\|_1} d\beta_{S_0},
$$
Therefore, (3.26) is bounded by the known property of the gradient of a convex function (e.g., Theorem 2.1.5 of Nesterov (2004)), the calculation of the Kullback-Leibler divergence in the proof of Lemma 3.1). By the well-diag(\lambda_n) = (\lambda_n \sqrt{n})^d_0\frac{\|X\|_0}{d_0!} \geq e^{-\frac{\lambda_n \sqrt{n}}{L_1}d_0}\frac{\|X\|_0}{d_0!} (p^s_0 \vee n^{d_0})^{L_2},

for every \( s_0 > 0 \). Note also that \( \sum_{j \in s_0} q_j \log a_j \lesssim d_0 \log q \) as \( a_j = O(\sqrt{q}) \) by Stirling’s approximation. Hence, by the inequalities \( (\frac{p}{s_0}) \leq p^s_0 \) and \( \log(d_0) \leq d_0(\log s_0 + \log q) \lesssim s_0 q \log n \), (3.25) is further bounded below by

\[
\pi_p(s_0) \exp \{-\lambda \|\beta_0\|_{2,1} - C_1 s_0 (\log p \vee q \log n)\},
\]

for some \( C_1 > 0 \).

\[\square\]

**Proof of Lemma 3.2.** Observe first that

\[
\sup_{\beta \in A_n} |\log A_n(\beta) - \log A_n^*(\beta)| \lesssim \sup_{\beta \in A_n} \left| \sum_{i=1}^n \|\text{diag}(\mu_i - \bar{\mu}_i) - \mu_i \mu_i^T + \bar{\mu}_i \bar{\mu}_i^T\|_2 \right| \|\theta_i - \theta_0\|_2^2.
\]

and

\[
\|\text{diag}(\mu_i - \bar{\mu}_i) - \mu_i \mu_i^T + \bar{\mu}_i \bar{\mu}_i^T\|_2 \leq \|\mu_i - \bar{\mu}_i\|_\infty + (\|\mu_i\|_2 + \|\bar{\mu}_i\|_2) \|\mu_i - \bar{\mu}_i\|_2 \lesssim \|\mu_i - \bar{\mu}_i\|_2.
\]

Now, to show that the last display is further bounded by a constant multiple of \( \|\theta_i - \theta_0\|_2 \), we shall verify that the function \( \mu : R^{m-1} \rightarrow R^{m-1} \), defined as \( \mu(\cdot) = (1 + h(\cdot)T1_{m-1})^{-1}h(\cdot) \), is Lipschitz continuous; that is, \( \|\mu(x) - \mu(y)\|_2 \lesssim \|x - y\|_2 \) for any \( x, y \in R^{m-1} \). Note that \( \log(1 + h(\cdot)T1_{m-1}) \) is convex and \( \mu(\cdot) \) is its gradient vector. Note also that its Hessian is \( \text{diag}(\mu(\cdot)) - \mu(\cdot) \mu(\cdot)^T \) whose spectral norm is bounded for any \( \cdot \in R^{m-1} \) (see Lemma 3.6 and the calculation of the Kullback-Leibler divergence in the proof of Lemma 3.1). By the well-known property of the gradient of a convex function (e.g., Theorem 2.1.5 of Nesterov (2004)), we reach the Lipschitz continuity of \( \mu \), which leads to the inequality \( \|\mu_i - \bar{\mu}_i\|_2 \lesssim \|\theta_i - \theta_0\|_2 \). Therefore, (3.26) is bounded by

\[
\sup_{\beta \in A_n} \|X(\beta - \beta_0)\|_2 \max_{1 \leq i \leq n} \|X_i(\beta - \beta_0)\|_2 \leq \|X\|_0^2 \sup_{\beta \in A_n} \|\beta - \beta_0\|_{2,1}^2 \max_{1 \leq i \leq n} \|X_i\|_0 \sup_{\beta \in A_n} \|\beta - \beta_0\|_{2,1}
\]

\[
\lesssim s_0^3(\log p \vee q \log n)^{3/2} \max_{1 \leq i \leq n} \|X_i\|_0
\]

which tends to zero by (C5).

\[\square\]
Proof of Lemma 3.3. First, it is easy to see that the expected value and the variance of $e^T_k \hat{\beta}_{S_0}$ are $e^T_k \beta_0, S_0$ and $s^2_{n,k} := e^T_k (X^T_{S_0} WX_{S_0})^{-1} e_k$, respectively. Note that $e^T_k \hat{\beta}_{S_0}$ can be written as

$$
\sum_{i=1}^{n} e^T_k (X^T_{S_0} WX_{S_0})^{-1} X^T_{S_0,i} W_i^{1/2} (U_i + W_i^{1/2} X_i \beta_0) =: \sum_{i=1}^{n} T_{i,k}.
$$

We verify Lindeberg’s condition. For $\nu_{i,k}$ the expected value of $T_{i,k}$, observe that for every $\epsilon > 0$,

$$
\frac{1}{s^2_{n,k}} \sum_{i=1}^{n} \mathbb{E} \left[ |T_{i,k} - \nu_{i,k}|^2 \mathbbm{1}_{\{|T_{i,k} - \nu_{i,k}| > \epsilon s_{n,k}\}} \right] \leq \frac{1}{s^2_{n,k}} \sum_{i=1}^{n} \left\{ \left\| W_i^{1/2} X_{S_0,i} (X^T_{S_0} WX_{S_0})^{-1} e_k \right\|_2^2 \right.
\times \mathbb{E} \left[ \left( \left\| U_i \right\|_2 \mathbbm{1}_{\left\{ \left\| U_i \right\|_2 > \epsilon s_{n,k} / \left\| W_i^{-1/2} X_{S_0,i} (X^T_{S_0} WX_{S_0})^{-1} e_k \right\|_2 \right\}} \right)^2 \right] \right\}
\leq \max_{1 \leq i \leq n} \mathbb{E} \left[ \left( \left\| U_i \right\|_2 \mathbbm{1}_{\left\{ \left\| U_i \right\|_2 > \epsilon s_{n,k} / \left\| W_i^{-1/2} X_{S_0,i} (X^T_{S_0} WX_{S_0})^{-1} e_k \right\|_2 \right\}} \right)^2 \right].
$$

The indicator variable on the rightmost side goes to zero by (C8). Since $\left\| U_i \right\|_2$ is bounded under the condition (C3), the preceding display goes to zero by Lebesgue’s dominated convergence theorem. The assertion then follows from the Lindeberg-Feller central limit theorem. □

Proof of Lemma 3.4. We first write $Z^T Q Z = \sum_{1 \leq j,k \leq n} Z^T_{j} Q_{jk} Z_k$ using the submatrices $Q_{jk} \in \mathbb{R}^{r_j \times r_k}$, $j, k \in \{1, \ldots, n\}$, such that

$$
Q = \begin{pmatrix}
Q_{11} & \cdots & Q_{1n} \\
\vdots & \ddots & \vdots \\
Q_{n1} & \cdots & Q_{nn}
\end{pmatrix}.
$$

Now, observe that

$$
\mathbb{E} \exp \left\{ t Z^T Q Z \right\} = \mathbb{E} \exp \left\{ t \sum_{j=1}^{n} Z^T_{j} Q_{jj} Z_j + t \sum_{j \neq k} Z^T_{j} Q_{jk} Z_k \right\}
\leq \exp \left\{ t \max_{1 \leq j \leq n} \bar{b}_j^2 \text{tr}(Q) \right\} \mathbb{E} \exp \left\{ t \sum_{j \neq k} Z^T_{j} Q_{jk} Z_k \right\},
$$

since $\sum_{j=1}^{n} \| Q_{jj} \|_{sp} \leq \sum_{j=1}^{n} \text{tr}(Q_{jj}) = \text{tr}(Q)$ by its positive semidefiniteness. Using the decou-
pling inequality in Theorem 3.1.1 of De la Pena and Giné (2012), we obtain that

\[
\mathbb{E} \exp \left\{ t \sum_{j \neq k} Z_j^T Q_{jk} Z_k \right\} \leq \mathbb{E} \exp \left\{ 4t \sum_{j=1}^{n} \sum_{k=1}^{n} Z_j^T Q_{jk} \tilde{Z}_k \right\},
\]

where \( \tilde{Z} = (\tilde{Z}_1^T, \ldots, \tilde{Z}_n^T)^T \) is an independent copy of \( Z \). It is clear that the right hand side of the display is equal to

\[
\mathbb{E} \mathbb{E} \left[ \exp \left\{ 4t \sum_{k=1}^{n} \sum_{j=1}^{n} Z_j^T Q_{jk} \tilde{Z}_k \right\} \right] = \mathbb{E} \prod_{k=1}^{n} \mathbb{E} \left[ \exp \left( 4t Z_k^T Q_k \tilde{Z}_k \right) \right],
\]

where \( Q_k = (Q_{1k}^T, \ldots, Q_{nk}^T)^T \in \mathbb{R}^{n \times r_k} \). Since

\[
\max_{\xi_k \in \text{supp}(Z_k)} Z_j^T Q_{jk} \xi_k - \min_{\xi_k, \xi'_k \in \text{supp}(Z_k)} Z_j^T Q_{jk} (\xi_k - \xi'_k) \leq \|Q_k Z\|_2 \tilde{b}_k,
\]

applying Hoeffding’s lemma to the inner expectation, we bound (3.28) by

\[
\mathbb{E} \prod_{k=1}^{n} \exp \left\{ 2t^2 \|Q_k Z\|_2^2 \tilde{b}_k^2 \right\} \leq \mathbb{E} \exp \left\{ 2t^2 \max_{1 \leq k \leq n} \tilde{b}_k^2 \sum_{k=1}^{n} \|Q_k Z\|_2^2 \right\}.
\]

(3.29)

Since we have that by the symmetry of \( Q \),

\[
\sum_{k=1}^{n} \|Q_k Z\|_2^2 = Z^T \left( \sum_{k=1}^{n} Q_k Q_k^T \right) Z = Z^T Q^2 Z,
\]

the right hand side of (3.29) is bounded by

\[
\mathbb{E} \exp \left\{ 2t^2 \max_{1 \leq k \leq n} \tilde{b}_k^2 Z^T Q^{1/2} Q^{1/2} Z \right\} \leq \mathbb{E} \exp \left\{ 2t^2 \max_{1 \leq k \leq n} \tilde{b}_k^2 \|Q\|_{sp} Z^T Q Z \right\},
\]

by the positive semi-definiteness of \( Q \). By Jensen’s inequality, this is further bounded by

\[
[\mathbb{E} \exp \left\{ t Z^T Q Z \right\}]^{2t \max_{1 \leq k \leq n} \tilde{b}_k^2 \|Q\|_{sp}} \cdot 0 < t < \frac{1}{2 \max_{1 \leq k \leq n} \tilde{b}_k^2 \|Q\|_{sp}}.
\]

Combining the last display and (3.27), we obtain the inequality given in the lemma. \( \Box \)

**Proof of Lemma 3.5.** Note that \( Y^{(n)} - \mu \) has a bounded support. By the Markov inequality
followed by Lemma 3.4, we have that for every $t > 0$ and $u < 1/(C_1 \| X \|_{sp}^2)$ with some $C_1 > 0$,

$$
\begin{align*}
\mathbb{P}_0 \left( \| X_j^T (Y^{(n)} - \mu) \|_2 > t \right) &\leq e^{-ut^2} \mathbb{E}_0 \exp \left\{ u \| X_j^T (Y^{(n)} - \mu) \|_2^2 \right\} \\
&\leq e^{-ut^2} \exp \left\{ C_1 u \cdot \text{tr}(X_j X_j^T) \right\} \\
&\quad \left( 1 - C_1 u \| X_j \|_{sp}^2 \right) \exp \left\{ \frac{t^2}{C_1} \| X \|_0^2 \right\}
\end{align*}
$$

for $k = 1, \ldots, p$. Note that $\text{tr}(X_j X_j^T) \leq q_j \| X_j \|_{sp}^2$ since the rank of $X_j$ is at most $q_j$. Hence, choosing $u = 1/(2C_1 \| X \|_{sp}^2)$, the rightmost side of the last display is further bounded by

$$
\exp \left\{ -\frac{t^2}{C_2 \| X \|_0^2} + C_2 q_j \right\} \leq \exp \left\{ -\frac{t^2}{C_2 \| X \|_0^2} + C_2 \bar{q} \right\},
$$

for some $C_2 > 0$. Choosing $t = 2C_2 \| X \|_0 \sqrt{\log p \vee \bar{q} \log n}$, we obtain

$$
\begin{align*}
\mathbb{P}_0 \left( \max_{1 \leq j \leq p} \| X_j^T W^{1/2} U \|_2 > 2C_2 \| X \|_0 \sqrt{\log p \vee \bar{q} \log n} \right) &\leq p \exp \left\{ -4C_2 (\log p \vee \bar{q} \log n) + C_2 \bar{q} \right\} \\
&\leq \frac{p}{(p \vee n^C) \| X \|_0^2},
\end{align*}
$$

which tends to zero by choosing $C_2$ to be larger than $1/3$. \hfill \square

**Proof of Lemma 3.6.** Let $\bar{\rho}_k$, $k = 1, \ldots, m - 1$, stand for the eigenvalues of $\Sigma$ in decreasing order. Without loss of generality, also rearrange $\pi_k$, $k = 1, \ldots, m - 1$, in decreasing order. We may apply the well-known matrix determinant lemma: for a non-singular matrix $Q \in \mathbb{R}^{d \times d}$ and $u \in \mathbb{R}^d$, $\det(Q + uu^T) = (1 + u^T Q^{-1} u) \det(Q)$ (e.g., Lemma 1.1 of Ding and Zhou (2007)). Similar to Watson (1996), we obtain that

$$
\pi_1 \geq \bar{\rho}_1 \geq \pi_2 \geq \bar{\rho}_2 \geq \cdots \geq \pi_{m-1} \geq \bar{\rho}_{m-1}. \tag{3.30}
$$

(Although Watson (1996) considers the covariance matrix of the full random vector $Z$ rather than the reduced form $\tilde{Z}$, the same approach can still be used.) The first inequality of (3.30) verifies the last inequality of the assertion in the lemma.

To find a lower bound of $\bar{\rho}_{m-1}$, observe that we have $\det(\Sigma) = \pi_m \prod_{k=1}^{m-1} \pi_k$ using the matrix determinant lemma again. Therefore,

$$
\bar{\rho}_{m-1} = \frac{\det(\Sigma)}{\prod_{k=1}^{m-2} \bar{\rho}_k} \geq \pi_m \pi_{m-1},
$$

due to (3.30). This verifies the first inequality of the assertion in the lemma. \hfill \square

**Proof of Lemma 3.7.** The proof consists of two parts. In the first part, we shall see that the
Hellinger distance $H(f_{β,i}, f_{0,i})$ strictly increases as $θ_i$ gets farther from $θ_{0i}$ along a straight line in any fixed direction on $\mathbb{R}^{m-1}$, and it has the global minimum zero at $θ_i = θ_{0i}$ as a function of $θ_i$. This directly verifies the desired assertion if $∥θ_{0i}∥_∞$ is bounded. The second part shows that $v(θ_{0i}) := \lim_{t→∞} \inf_{θ_i:∥θ_i∥_∞=t} H(f_{β,i}, f_{0,i}) = o(M_n^{-1})$ whenever $\exp(∥θ_i∥_∞) = o(M_n^2)$, so that the closeness of $θ_i$ and $θ_{0i}$ can be recovered from $H(f_{β,i}, f_{0,i}) \lesssim M_n^{-1}$.

For the first part, note that the squared Hellinger distance between $f_{β,i}$ and $f_{0,i}$ is given by

$$1 - \frac{H^2(f_{β,i}, f_{0,i})}{2} = \frac{1 + h((θ_i + θ_{0i})/2)Tm_{i-1}}{(1 + h(θ_i)Tm_{i-1})(1 + h(θ_{0i})Tm_{i-1})}.$$ 

For any fixed $θ_i \neq θ_{0i}$, let $θ_i(t) = tθ_i + (1 - t)θ_{0i}$, $t \in (0, 1)$, and define

$$g_i(t) := -\log \left\{ 1 - \frac{H^2(\tilde{f}_{β,i}, f_{0,i})}{2} \right\},$$

where $\tilde{f}_{β,i}$ is the density with $\tilde{θ}_i(t)$. Since $g_i(0) = 0$, it suffices to show that $g_i'(0) = 0$ and $g_i(t)$ is strictly increasing in $t \in (0, 1)$ to prove the claim. Note that by direct calculations,

$$g_i'(t) = \frac{h(θ_{0i} + t(θ_i - θ_{0i}))}{1 + h(θ_{0i} + t(θ_i - θ_{0i}))Tm_{i-1}}.$$ 

It can be easily checked that $k_i(t)$ is strictly increasing in $t \in (0, 1)$. Hence, we have that $g_i'(t) > 0$ for every $t \in (0, 1)$ while $g_i'(0) = 0$ is easily verified.

Now, for the second part, observe that

$$\frac{v^2(θ_{0i})}{2} = 1 - \sup_{a \in \mathbb{R}^{m-1}} \lim_{x → +∞} \frac{1 + \sum_{k=1}^{m-1} e^{(a_k x + θ_{0ik})/2}}{\sqrt{1 + \sum_{k=1}^{m-1} e^{a_k x}}} \left(1 + \sum_{k=1}^{m-1} e^{θ_{0ik}}\right),$$

where $θ_{0ik}$ and $a_k$ are the $k$th entries of $θ_{0i}$ and $a$, respectively. For a given $a \in \mathbb{R}^{m-1}$, split the set $\{1, \ldots, m - 1\}$ into the three disjoint subsets: $K_0 := \{k : a_k = 0, 1 \leq k \leq m - 1\}$, $K_+ := \{k : a_k > 0, 1 \leq k \leq m - 1\}$, and $K_- := \{k : a_k < 0, 1 \leq k \leq m - 1\}$. Then we can write

$$\lim_{x → +∞} \frac{1 + \sum_{k=1}^{m-1} e^{(a_k x + θ_{0ik})/2}}{\sqrt{1 + \sum_{k=1}^{m-1} e^{a_k x}}} = \lim_{x → +∞} \frac{1 + \sum_{k \in K_0} e^{θ_{0ik}}/2 + \sum_{k \in K_+} e^{(a_k x + θ_{0ik})/2} + \sum_{k \in K_-} e^{(a_k x + θ_{0ik})/2}}{\sqrt{1 + \sum_{k \in K_0} 1 + \sum_{k \in K_+} e^{a_k x} + \sum_{k \in K_-} e^{a_k x}}. \quad (3.32)$$

This limit is determined by the value of $a$. We only need to consider the following three cases as $K_0$ cannot be the same as $\{1, \ldots, m - 1\}$.
(i). If $K_+ \neq \emptyset$, the limit is solely determined by the coefficients $a_k$ in $K_+$. In this case, for the set $K_+ := \{ k \in K_+ : a_k = \max_j a_j \}$ and its cardinality $|K_+|$, it can be seen that (3.32) is equal to $|K_+|^{-1/2} \sum_{k \in K_+} e^{\theta_0 k}/2$, which is bounded by $\sqrt{\sum_{k \in K_+} e^{\theta_0 k}}$ by the power mean inequality.

(ii). If $K_0 \neq \emptyset$ and $K_+ = \emptyset$, the limit in (3.32) is $(1 + |K_0|)^{-1/2} (1 + \sum_{k \in K_0} e^{\theta_0 k}/2)$, which is bounded by $\sqrt{1 + \sum_{k \in K_0} e^{\theta_0 k}}$ by the power mean inequality.

(iii). If $K_0 = \emptyset$ and $K_+ = \emptyset$, the limit is 1.

Plugging in these values, we see that (3.31) is bounded below by

$$\sqrt{\frac{1 \wedge e^{\min_k \theta_0 k}}{1 + \sum_{k=1}^{m-1} e^{\theta_0 k}}} \geq \sqrt{\frac{1 \wedge e^{-\|\theta_0\|_{\infty}}}{1 + (m - 1)e^{\|\theta_0\|_{\infty}}}.$$ 

Hence, if $M_n^{-2}$ decreases faster than the right hand side of the display (say, $M_n^{-2} = o(RHS)$ for RHS the right hand side), then the lemma follows. We only need to consider increasing $\|\theta_0\|_{\infty}$ as the claim holds with any increasing $M_n$ if $\|\theta_0\|_{\infty}$ is bounded. We see that the required condition is satisfied if $\exp(\|\theta_0\|_{\infty}) = o(M_n^2)$. \hfill \square

**Proof of Lemma 3.8.** By directly computing the second and third derivatives of $\eta(t)$, we obtain that

$$e^{2\eta(t)} \eta''(t) = \sum_{j \neq k} e^{w_j + w_k + t(v_j + v_k)} (v_j - v_k)^2 + \sum_{j=1}^{m-1} v_j^2 e^{w_j + tv_j}, \quad (3.33)$$

and

$$e^{3\eta(t)} \eta'''(t) = \sum_{j \neq k} e^{w_j + w_k + t(v_j + v_k)} (2v_j^3 + 2v_j^2 + 3v_j v_k) + \sum_{j=1}^{m-1} v_j^2 e^{w_j + tv_j} (1 - e^{w_j + tv_j})$$

$$+ \sum_{j \neq k} e^{w_j + w_k + t(v_j + v_k)} (e^{w_j + tv_j} (v_k - v_j)^3 + e^{w_k + tv_k} (v_j - v_k)^3).$$

Multiplying $e^{\eta(t)}$ to both sides of (3.33), we have that

$$e^{3\eta(t)} \eta''(t) = \sum_{j \neq k} e^{w_j + w_k + t(v_j + v_k)} ((v_j - v_k)^2 + v_j^2 + v_k^2) + \sum_{j=1}^{m-1} v_j^2 e^{w_j + tv_j} (1 + e^{w_j + tv_j})$$

$$+ \sum_{j \neq k} e^{w_j + w_k + t(v_j + v_k)} (v_j - v_k)^2 \left(e^{w_j + tv_j} + e^{w_k + tv_k}\right).$$

Note that $2v_j^3 + 2v_j^2 - 3v_j v_k - 3v_j v_k^2 = 2(v_j - v_k)^2(v_j + v_k) - v_j^2 v_k - v_j v_k^2$. The assertion thus follows by comparing the last two displays. \hfill \square

**Proof of Lemma 3.9.** We first verify the first assertion of the lemma. Let $H(p_1, p_2) := H_{1/2}(p_1, p_2)$
be the affinity between two densities \( p_1 \) and \( p_2 \), where \( \mathcal{H}_\alpha(p_1, p_2) = \int p_1^\alpha p_2^{1-\alpha} \) is the Hellinger transform for \( \alpha \in (0, 1) \). By Lemma 6.1 of Kleijn and van der Vaart (2006), there exists a test function \( \varphi_n^* \) such that

\[
\mathbb{E}_0 \varphi_n^* + \sup_{q \in Q_1} \int (1 - \varphi_n^*) q d\nu_n \leq \sup_{q \in \text{conv}(Q_1)} \mathcal{H}(f_0, q),
\]

where \( \text{conv}(Q_1) \) is the convex hall of \( Q_1 \). Note that any \( q \in \text{conv}(Q_1) \) can be expressed by finite convex combination of \( q_{\beta_k} \in Q_1, k = 1, 2, \ldots \); that is, \( q = \sum_k a_k q_{\beta_k} \), where \( \sum_k a_k = 1 \) and \( a_k \geq 0 \) for every \( k \). We shall show that for every \( q_{\beta} \in Q_1 \), it follows that \( q_{\beta}/f_0 \leq e^{-n\delta^2/32} \) uniformly over the sample space. Then, by the definition of \( q_{\beta} \), the affinity satisfies

\[
\mathcal{H}(f_0, q) = \int \sqrt{\sum_k a_k \mathbb{E}_n} \frac{f_{\beta_k}}{f_0} f_0 d\nu_n \leq \sqrt{\sum_k a_k e^{-n\delta^2/32}} \leq e^{-n\delta^2/64}. \tag{3.34}
\]

To complete the proof of the first assertion, we only need to show that \( q_{\beta}/f_0 \leq e^{-n\delta^2/32} \) for every \( q_{\beta} \in Q_1 \), uniformly over the sample space. Observe that

\[
\sum_{i=1}^n \left\{ \log \frac{f_{\beta,i}}{f_{0,i}}(Y_i) - \left( Y_i - \frac{h(\theta_{0i})}{1 + h(\theta_{0i})^T \Lambda_{m-1}} \right)^T (\theta_i - \theta_{0i}) \right\} = -\sum_{i=1}^n \left\{ \log \frac{1 + h(\theta_i)^T \Lambda_{m-1}}{1 + h(\theta_{0i})^T \Lambda_{m-1}} - \frac{h(\theta_{0i})^T (\theta_i - \theta_{0i})}{1 + h(\theta_{0i})^T \Lambda_{m-1}} \right\}.
\]

Let \( b(\cdot) := \log(1 + h(\cdot)^T \Lambda_{m-1}) \) and write its gradient vector and Hessian matrix of \( b \) as \( \nabla b \) and \( \nabla^2 b \), respectively. Using Proposition 1 of Bach (2010) and Lemma 3.8 above, the expression in the last display can be written as

\[
-\sum_{i=1}^n \left\{ b(\theta_i) - b(\theta_{0i}) - \nabla b(\theta_{0i})^T (\theta_i - \theta_{0i}) \right\} \\
\leq -\sum_{i=1}^n \frac{(\theta_i - \theta_{0i})^T \nabla^2 b(\theta_{0i})(\theta_i - \theta_{0i})}{R^2 \|	heta_i - \theta_{0i}\|^2_2} (e^{-R\|	heta_i - \theta_{0i}\|_2} + R\|	heta_i - \theta_{0i}\|_2 - 1) \\
\leq -\sum_{i=1}^n \frac{(\theta_i - \theta_{0i})^T \nabla^2 b(\theta_{0i})(\theta_i - \theta_{0i})}{2 + R\|	heta_i - \theta_{0i}\|_2},
\]

where the second inequality holds since \( e^{-x} + x - 1 \geq x^2/(2 + x) \) for every \( x \geq 0 \). Hence, we
obtain that

$$
\sum_{i=1}^{n} \log \frac{f_{\beta,i}}{f_{0,i}}(Y_i) \leq - \sum_{i=1}^{n} \left\{ \frac{(\theta_i - \theta_{0i})^T [\nabla^2 b(\theta_{0i})] (\theta_i - \theta_{0i})}{2 + R \| \theta_i - \theta_{0i} \|_2} + (Y_i - \nabla b(\theta_{0i}))^T (\theta_i - \theta_{0i}) \right\} \quad (3.35)
$$

$$
\leq - \frac{(\beta - \beta_0)^T X^T W X (\beta - \beta_0)}{2 + R \max_i \| X_i \| \| \beta - \beta_0 \|_2} + \max_{1 \leq j \leq p} \| X_j^T W^{1/2} U \|_2 \| \beta - \beta_0 \|_{2,1},
$$

where the second inequality follows by the identity $\nabla^2 b(\theta_{0i}) = \text{diag}(\mu_i) - \mu_i \mu_i^T$. We see that on the event $\tilde{\mathcal{E}}_n$, (3.35) is bounded by

$$
- \frac{\| X \|_o^2 \| \beta - \beta_0 \|_2 \psi_2((K_1 + 1)s_0)}{2 + R \sqrt{(K_1 + 1)s_0} \max_i \| X_i \| \| \beta - \beta_0 \|_2} + \frac{\delta \sqrt{n}}{16} \| X \|_o \| \beta - \beta_0 \|_2. \quad (3.36)
$$

If $\psi_2((K_1 + 1)s_0) \| X \|_o \geq 4RC_* (K_1 + 1)s_0 \sqrt{\log p \vee q \log n} \max_i \| X_i \|_o$ and $\psi_2((K_1 + 1)s_0) \| X \|_o \| \beta - \beta_0 \|_2 \geq \delta \sqrt{n}/2$, the denominator of the first term in (3.36) satisfies

$$
2 + R \sqrt{(K_1 + 1)s_0} \max_{1 \leq i \leq n} \| X_i \| \| \beta - \beta_0 \|_2 \\
\leq \frac{4\psi_2((K_1 + 1)s_0) \| X \|_o \| \beta - \beta_0 \|_2}{\delta \sqrt{n}} + \frac{\psi_2((K_1 + 1)s_0) \| X \|_o \| \beta - \beta_0 \|_2}{4C_* \sqrt{(K_1 + 1)s_0} (\log p \vee q \log n)} \\
\leq \frac{8\psi_2((K_1 + 1)s_0) \| X \|_o \| \beta - \beta_0 \|_2}{\delta \sqrt{n}},
$$

and hence (3.36) is bounded by

$$
- \frac{\delta \sqrt{n}}{16} \| X \|_o \| \beta - \beta_0 \|_2 \leq - \frac{n\delta^2}{32 \psi_2((K_1 + 1)s_0)}.
$$

This is further bounded by $-n\delta^2/32$ since

$$
\psi_2((K_1 + 1)s_0) \leq \psi_2(1) \leq \phi_2(1) \leq \inf_{|\beta|:S_\beta=1} \frac{\| \beta \|_{2,1}}{\| \beta \|_2} \leq \inf_{|\beta|:S_\beta=1} |S_\beta|^{1/2} = 1. \quad (3.37)
$$

This verifies the first assertion.

Now, we similarly verify the second assertion. By Lemma 6.1 of Kleijn and van der Vaart (2006) again, there exists a test function $\varphi_n^*$ such that

$$
\mathbb{E}_0 \varphi_n^* + \sup_{q \in \mathcal{Q}_1} \int (1 - \varphi_n^*) q d\nu_n \leq \sup_{q \in \text{conv}(\mathcal{Q}_1)} \mathcal{H}(f_0, q).
$$

Similar to (3.34), the right hand side of the display is further bounded by $e^{-n\delta^2/64}$ as long as $q_{\beta}/f_0 \leq e^{-n\delta^2/32}$ for every $q_{\beta} \in \mathcal{Q}_1$, uniformly over the sample space. To this end, observe that

108
on the event $\mathcal{E}_n$, (3.35) is bounded by

$$
- \frac{\|X\|_2^2 \|\beta - \beta_0\|_{2,1}^2 ((K_1 + 1)s_0)}{(K_1 + 1)s_0(2 + R \max_i \|X_i\|_o \|\beta - \beta_0\|_{2,1})} + \frac{\delta \sqrt{n}}{16 \sqrt{(K_1 + 1)s_0}} \|X\|_o \|\beta - \beta_0\|_{2,1}.
$$

(3.38)

If $\psi_{2,1}^2((K+1)s_0) \|X\|_o \geq 4RC_\ast (K+1)s_0 \sqrt{\log p \sqrt{\log n}} \max_i \|X_i\|_o$ and $\psi_{2,1}^2((K+1)s_0) \|X\|_o \|\beta - \beta_0\|_{2,1} \geq \sqrt{(K+1)s_0 \|\beta - \beta_0\|_{2,1}}$, it follows that

$$
2 + R \max_{1 \leq i \leq n} \|X_i\|_o \|\beta - \beta_0\|_{2,1} \leq \frac{4\psi_{2,1}^2((K+1)s_0) \|X\|_o \|\beta - \beta_0\|_{2,1}}{\delta \sqrt{(K+1)s_0n}} + \frac{\psi_{2,1}^2((K+1)s_0) \|X\|_o \|\beta - \beta_0\|_{2,1}}{4C_\ast (K+1)s_0 \sqrt{\log p \sqrt{\log n}}}
$$

$$
\leq \frac{8\psi_{2,1}^2((K+1)s_0) \|X\|_o \|\beta - \beta_0\|_{2,1}}{\delta \sqrt{(K+1)s_0n}}.
$$

By plugging in this, we bound (3.38) by

$$
- \frac{\delta \sqrt{n}}{16 \sqrt{(K+1)s_0}} \|X\|_o \|\beta - \beta_0\|_{2,1} \leq - \frac{n\delta^2}{32 \psi_{2,1}^2((K+1)s_0)} \leq - \frac{n\delta^2}{32},
$$

where the second inequality holds since we have that $\psi_{2,1}((K+1)s_0) \leq \phi_{2,1}(1) \leq 1$ similar to (3.37). This proves the second assertion. \qed
Chapter 4

Posterior contraction in sparse generalized linear models

4.1 Introduction

Since its foundational theory was developed in McCullagh and Nelder (1989), a generalized linear model (GLM) has been an indispensable tool for nonlinear regression and has received great attention from practitioners. Although it has gained a lot of empirical success in both low- and high-dimensional situations, extensive theoretical investigations for optimal estimation in high-dimensional GLMs have been far less studied than in the linear model settings due to complications. In a pioneering work on this topic, van de Geer (2008) studied the oracle rate of the empirical risk minimizer with the lasso penalty in high-dimensional GLMs. More recently, Abramovich and Grinshtein (2016) derived convergence rates with respect to the Kullback-Leibler risk with a wide class of penalizing functions, which can be translated into convergence rates relative to the $\ell_2$-norm under certain conditions. One might also refer to Negahban et al. (2012) for estimation properties and convergence rates in high-dimensional GLMs under more general settings.

From the Bayesian point of view, the topic has been rarely studied and the veil has not been lifted yet, whereas there is a substantial body of literature on sparse linear regression (e.g., Castillo et al., 2015; Gao et al., 2015; Martin et al., 2017; Ročková, 2018; Ning et al., 2018; Belitser and Ghosal, 2019; Bai et al., 2019). For GLMs with increasing dimensional parameters, Ghosal (1997) studied asymptotic normality of the posterior distribution without assuming sparsity, but the number of parameters is assumed to be much less than the number of observations. To our best knowledge, Jiang (2007) is the only available Bayesian investigation of asymptotic properties for the posterior distributions in the high-dimensional setting of GLMs. In this setting, the author studied posterior contraction rates relative to the Hellinger
metric for densities by treating the regression model as an i.i.d. (independent and identically distributed) class with a random design matrix. Liang et al. (2013) also obtained comparable posterior contraction rates for maximum a posteriori models with respect to the Hellinger metric. Although the Hellinger metric is useful for comparison of densities, it gives a vague interpretation for the regression coefficients. A much clearer interpretation of convergence is obtained from contraction with respect to $\ell_q$-type metrics directly on the parameters, which has not been addressed so far in the literature. In the setting with an i.i.d. model, it may not be simple to derive such a recovery result, as the regression coefficients are intertwined with a random design matrix which is not a fixed object. Our study aims to fill this gap.

We consider high-dimensional GLMs with fixed designs and investigate the desired posterior contraction rates with respect 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 natural form; that is, the density (with respect to a dominating measure) of the $i$th observation has the form of

$$f_i(y_i) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{\tau_i} + k(y_i, \tau_i) \right\}, \quad (4.1)$$

where $\theta_i$ is a natural parameter lying on a space $\Theta \subset \mathbb{R}$, $\tau_i$ is a known dispersion factor, and $b$ and $k$ are known functions. Following the convention, the function $b$ is assumed to be twice-differentiable and strictly convex on $\Theta$ such 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 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 coefficients $\beta \in \mathbb{R}^p$ with $p > n$, we consider a link function $h$ such that $(h \circ b') : \Theta \mapsto \mathbb{R}$ is monotone increasing and

$$h(b'(\theta_i)) = X_i^T \beta. \quad (4.2)$$

As $p > n$, we assume that the coefficient $\beta$ is actually sparse, where ‘sparsity’ means that many of the components in $\beta$ are zero. We note that our formulation does not require $h$ to be $b'^{-1}$, the inverse function of $b'$, so that high-dimensional GLMs with non-canonical link functions can also be of our interest. We believe that this generalization is also a substantial contribution of this chapter. This allows the main results developed in this chapter to be used for some standard but non-canonical high-dimensional GLMs, such as probit regression and gamma regression with the logarithmic link.

Ghosal and van der Vaart (2007) developed the general posterior contraction theory for non-i.i.d. models. Based on Birgé (1983), they showed that there always exists an exponentially powerful local-test relative to the root-average-squared Hellinger metric for independent but possibly non-identically distributed models, regardless of the underlying model-likelihood.
With a proper class of prior distributions, we employ their general recipes to derive the posterior contraction rate with respect to the root-average-squared Hellinger metric in high-dimensional GLMs, by treating the class as an independent regression model. In this sense, our study is fundamentally different from Jiang (2007), where the model was treated as an i.i.d. class. The $\ell_q$-contraction rates ($q = 1, 2$) for the regression coefficients are then obtained 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 $s_0 \sqrt{(\log p)/n}$ and $\sqrt{(s_0 \log p)/n}$ (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).

The rest of this chapter is organized as follows. In Section 4.2, we introduce the notations used throughout the chapter and specify required assumptions and an appropriate class of prior distributions for $\beta$. Section 4.3 provides our main results on posterior contraction relative to the root-average-squared Hellinger metric and the recovery. All the technical proofs are deferred to Section 4.4.

4.2 Setup and prior specification

4.2.1 Notation

Here we describe the notations used in this chapter. We have $n$ independent observations denoted by $Y^{(n)} = (Y_1, \ldots, Y_n)^T$ generated from the model in (4.1) and (4.2) with the true parameter $\beta_0$. 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 $\beta$ and a set $S \subset \{1, \ldots, p\}$ of indices, let $\beta_S = \{\beta_j, j \in S\}$ and $\beta_{S^c} = \{\beta_j, j \notin S\}$ to separate $\beta$ into zero and nonzero coefficients using $S$. We also write $S_\beta = \{j : \beta_j \neq 0\} \subset \{1, \ldots, p\}$ for the effective support determined by a vector $\beta$. 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.

For two sequences $a_n$ and $b_n$, $a_n \lesssim b_n$ (or $b_n \gtrsim a_n$) implies $a_n \leq Cb_n$ for some constant $C > 0$ independent of $n$. This notation is also used for constant sequences. For norms of vectors, let $\|\cdot\|_q$ denote the $\ell_q$-norm and $\|\cdot\|_\infty$ denote the max-norm. The max-norm notation $\|\cdot\|_\infty$ is also used for a matrix, 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. In particular, the spectral norm of a matrix is denoted by $\|\cdot\|_{sp}$.

Let $f_{\beta,i}$ and $f_{0,i}$ denote densities with an arbitrary parameter $\beta$ and the true parameter $\beta_0$, respectively. The joint densities are then defined by $f_\beta = \prod_{i=1}^n f_{\beta,i}$ and $f_0 = \prod_{i=1}^n f_{0,i}$, and the likelihood ratio is written by $\Lambda_n(\beta) = (f_\beta/f_0)(Y^{(n)})$. The root-average-squared Hellinger metric
is defined by $H_n(\beta, \beta_0) = \{n^{-1} \sum_{i=1}^n H^2(f_{\beta,i}, f_{0,i})\}^{1/2}$ for the Hellinger distance $H(f_{\beta,i}, f_{0,i}) = \{\int (\sqrt{f_{\beta,i}} - \sqrt{f_{0,i}})^2\}^{1/2}$ between the two densities. The notations $E_0$ and $P_0$ 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 $\phi_2$:

$$
\phi_1(s) = \inf_{\beta: |S_\beta| \leq s} \frac{\|X\beta\|_2 |S_\beta|^{1/2}}{\sqrt{n} \|\beta\|_1}, \quad \phi_2(s) = \inf_{\beta: |S_\beta| \leq s} \frac{\|X\beta\|_2}{\sqrt{n} \|\beta\|_2}.
$$

These definitions for compatibility are required to obtain the $\ell_1$- and $\ell_2$-contraction rates from the rate relative to the squared loss of the linear predictor. Similar definitions are widely used in the Bayesian high-dimensional literature (e.g., Castillo et al., 2015; Martin et al., 2017; Belitser and Ghosal, 2019; Chae et al., 2019).

### 4.2.2 Model assumption

To proceed our investigation, further assumptions are required for the model given in (4.1) and (4.2). The assumptions are solely made through the function $b$ and the link function $h$ as follows.

(A1) For any compact subset $K \subset \Theta$, there exist constants $c_s, c^* > 0$ such that $c_s \leq \inf_{x \in K} b''(x) \leq \sup_{x \in K} b''(x) \leq c^*$.

(A2) The composite function $(h \circ b')$ is locally bi-Lipschitz continuous; that is, for any compact subset $K \subset \Theta$, $|x - y| \lesssim h(b'(x - y)) \lesssim |x - y|$ for every $x, y \in K$.

The assumption (A1) is made mainly to relate the Hellinger metric with the difference of canonical parameters. In the high-dimensional GLM context, even stronger assumptions are often made such that the second derivative $b''$ is bounded uniformly over $\Theta$ (e.g., Abramovich and Grinshtein, 2016). One can easily see that such a stronger limitation is too restrictive for some popular GLMs; for example, Poisson regression possesses $b''(x) = e^x$ which is not bounded.

The assumption (A2) is required to link the canonical parameter $\theta_i$ and the linear predictor $X_i^T \beta$ in an effective way. This assumption is met as soon as $h$ is locally bi-Lipschitz itself, as $b'$ is already locally bi-Lipschitz continuous by (A1). It is trivial that the assumption is satisfied for canonical link functions.

Note that the assumptions on $b$ and $h$ above are very mild and satisfied in almost every practical application of the GLMs, as invertible smooth link functions are almost always used for a homeomorphism in $\mathbb{R}$ and a violation is extremely rare. The conditions are satisfied in all the popular GLM settings including logistic regression, probit regression, Poisson regression, gamma regression with the logarithmic link, negative binomial regression, etc. In what follows we only consider the model in (4.1) and (4.2) endowed with the assumptions (A1) and (A2).
4.2.3 Prior specification

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

Similar to the literature, we first choose a dimension $s$ from a prior $\pi_p$, and then a subset $S \subset \{1, \ldots, p\}$ is randomly chosen for given $s$. Then we select a nonzero part $\beta_S$ from a prior $g_S$ on $\mathbb{R}^s$ while $\beta_{S^c}$ is fixed to zero. The resulting prior specification for $(S, \beta)$ is summarized as

$$(S, \beta) \mapsto \frac{\pi_p(s)}{p} g_S(\beta_S)\delta_0(\beta_{S^c}),$$

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

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

$$\pi_p(s) = 0, \quad s = s + 1, \ldots, p,$$  

where $s$ is a predetermined boundedness level. Examples of priors satisfying (4.3) can be found in Castillo and van der Vaart (2012) and Castillo et al. (2015). If $g_S$ is well defined for every $S \subset \{1, \ldots, p\}$, there is no need to make a restriction and the bound $s$ can be chosen to be $p$ as in Castillo et al. (2015), so that the second line of (4.3) vanishes. Otherwise, putting a restriction $s < p$ may be required, coupled with a choice of $g_S$ (see the examples below). Even in such a case, it still makes sense to consider a sequence $s$ increasing at a suitable order. Indeed, we need to assume that $s_0 \leq s$ for large enough $n$ to recover the true sparsity level.

For the prior $g_S$, we consider a class of distributions satisfying certain conditions. For a given $g_S$, let $G_S$ be the probability operator with $g_S$ and $Z_S \in \mathbb{R}^s$ be a random variable that possesses the density function $g_S$. Motivated by our targeted rate, a prior distribution should meet the following conditions: for every $S$ such that $s \leq M_1 s_0$ with any large $M_1$ and $s_0 > 0$, there exists a constant $M_2$ such that

$$\log G_S\left(\|Z_S - \beta_0, S\|_1 \leq M_2\|X\|^{-1}_\infty \sqrt{(s_0 \log p)/n}\right) \gtrsim -s_0 \log p,$$  

$$\log G_S\left(\|Z_S - \beta_0, S\|_\infty > \|X\|^{-1}_\infty p^{M_2} - \|\beta_0\|_\infty\right) \lesssim -r_n,$$  

where $r_n$ is a sequence such that $(s_0 \log p)/r_n \to 0$ as $n \to \infty$. The constant $M_2$ can be made as large as desired since a larger value makes the conditions weaker. By reading the proof, one can see that we actually need (4.4) to be satisfied with only $S = S_0$. Yet, the given form is more sensible as the true sparsity class is unknown, and is easily satisfied provided that the density
$g_S$ belongs to the same class for every $S \subset \{1, \ldots, p\}$ such that $s \leq \bar{s}$.

The conditions (4.4) and (4.5) require some magnitude restriction of the true signal $\beta_0$ and the design matrix $X$. In particular, the condition (4.5) directly implies that $\|\beta_0\|_\infty$ and $\|X\|_\infty$ should be at most a polynomial in $p$. The latter is certainly mild. In Bayesian high-dimensional regression with Gaussian errors, the restriction on $\|\beta_0\|_\infty$ is often eliminated by choosing a proper prior distribution (e.g., Castillo et al., 2015; Gao et al., 2015). In our GLM setting, this size restriction can be allowed as it will be covered by another condition (see the examples below and the paragraphs following them for a more detailed discussion). Indeed, it is still very mild to assume that $\|\beta_0\|_\infty$ is at most a polynomial in $p$ if a suitably large power is chosen.

Plugging in the upper bound for $\|X\|_\infty$, the conditions (4.4) and (4.5) can usually be verified by stronger inequalities. As we noted, the upper bound should satisfy the mild restriction $\log\|X\|_\infty \lesssim \log p$, while a much stronger condition is required for the recovery; see (4.7). Note that in the typical regression settings, we commonly have $\|X\|_\infty \lesssim 1$.

The examples below are prior distributions possessing exponentially decaying tails, but tails can be thicker than those of normal distributions. The distributions may need to possess dispersion that increases at a proper rate to relax the magnitude of the true coefficients $\beta_0$.

Then the boundedness requirement for $\beta_0$ is comparable to that for sparse linear regression in Section 4 of Castillo et al. (2015).

**Example 4.1** (Uniform). Consider the prior in (4.3) with $\bar{s} = p$. For a sufficiently large $B > 0$, suppose that $g_S(\beta_S) = (2p^{-B})^s$ if $\|\beta_S\|_\infty \leq p^B$, and $g_S(\beta_S) = 0$ otherwise. If $\log\|X\|_\infty \lesssim \log p$ and $\|\beta_0\|_\infty \leq p^{(B-a)}$ for some $a > 0$, then the conditions (4.4) and (4.5) are satisfied.

**Example 4.2** (Laplace). Consider the prior in (4.3) with $\bar{s} = p$. Suppose that $g_S(\beta_S) = (\lambda/2)^s e^{-\lambda \|\beta_S\|_1}$ with $\lambda$ assumed to satisfy $1/(B_1p^{B_2}) \leq \lambda \leq B_3$ for some constants $B_1, B_2, B_3 > 0$. If $\log\|X\|_\infty \lesssim \log p$ and $\|\beta_0\|_\infty \lesssim \lambda^{-1} \log p$, then the conditions (4.4) and (4.5) are satisfied.

**Example 4.3** (Elliptical Laplace). Consider the prior in (4.3) for some predefined sequence $\pi \leq p$ with the assumption $s_0 \leq \pi$, and a positive definite matrix $V_S \in \mathbb{R}^{s \times s}$ such that $p^{-a} \leq \rho_{\min}(V_S) \leq \rho_{\max}(V_S) \leq p^a$ with some constant $a > 0$, for every $S$ such that $s \leq \bar{s}$. We choose $g_S(\beta_S) = (\lambda/\sqrt{\pi})^s \det(V_S)^{-1/2} \Gamma(s/2+1)/\Gamma(s+1)] \exp(-\lambda \|V_S^{-1/2} \beta_S\|_2)$ with $\lambda$ assumed to satisfy $1/(B_1p^{B_2}) \leq \lambda \leq B_3 \sqrt{\rho_{\min}(V_S)}$ for some constants $B_1, B_3 > 0$ and $B_2 > a/2$. If $\log\|X\|_\infty \lesssim \log p$ and $\|\beta_0\|_\infty \lesssim \lambda^{-1} \sqrt{\rho_{\min}(V_S)} \log p$ for every $S$ such that $s \leq \bar{s}$, then the conditions (4.4) and (4.5) are satisfied.

**Example 4.4** (Multivariate normal). Consider the prior in (4.3) for some predefined sequence $\pi \leq p$ with the assumption $s_0 \leq \pi$, and a positive definite matrix $V_S \in \mathbb{R}^{s \times s}$ such that $p^{-a} \leq \rho_{\min}(V_S) \leq \rho_{\max}(V_S) \leq p^a$ with some constant $a > 0$, for every $S$ such that $s \leq \bar{s}$. We choose $g_S(\beta_S) = (\lambda/\sqrt{2\pi})^s \det(V_S)^{-1/2} \exp(-\lambda \|V_S^{-1/2} \beta_S\|_2^2)$ with $\lambda$ assumed to satisfy $1/(B_1p^{B_2}) \leq
\[ \lambda \leq B_3 \rho_{\min}(V_S) \] for some constants \( B_1, B_2 > 0 \) and \( B_2 > a \), that is, the normal density with mean zero and covariance matrix \( \lambda^{-1} V_S \). If \( \log \|X\|_\infty \lesssim \log p \) and \( \|\beta_0\|_\infty \lesssim \sqrt{\lambda^{-1} \rho_{\min}(V_S) \log p} \) for every \( S \) such that \( s \leq \bar{s} \), then the conditions (4.4) and (4.5) are satisfied.

Although the uniform prior in Example 4.1 satisfies all the asymptotic requirements in the assumption, it is less appealing from a practical point of view as it requires to truncate the space \( \mathbb{R}^p \) for inference with given data. In Example 4.3 and Example 4.4, if the densities are well defined for every \( S \), there is no need to consider the restriction \( \bar{s} < p \) and one may choose \( \bar{s} = p \). A typical example is to use a multiple of an identity matrix for \( V_S \). Still, putting a restriction may be needed in some cases. For example, the choice \( V_S = (X_S^T X_S)^{-1} \) is often preferred for computational reasons (e.g., Lee et al., 2003; Zhou et al., 2004; Hanson et al., 2014), where the bound should satisfy \( \bar{s} \leq n \).

The prior distributions given in the examples above require restricted magnitude of \( \beta_0 \) in terms of the max-norm. In sparse linear regression with Gaussian errors, this may be eliminated for deriving posterior contraction by directly handling the denominator and numerator of the posterior distribution with a suitably chosen prior distribution (e.g., Castillo et al., 2015; Gao et al., 2015; Martin et al., 2017; Belitser and Ghosal, 2019). However, we will need to assume that \( \|X\beta_0\|_\infty \lesssim 1 \) for the true signal as in other high-dimensional GLM literature (e.g., Ghosal, 1997), and even stronger conditions are often made (e.g., Abramovich and Grinshtein, 2016). The boundedness requirement for \( \|\beta\|_\infty \) given in each example can be made milder than the restriction \( \|X\beta_0\|_\infty \lesssim 1 \) by choosing \( \lambda \) appropriately, since we have the inequality \( \|\beta_0\|_\infty \leq \sqrt{n} \| (X_{S_0}^T X_{S_0})^{-1/2} \|_{sp} \|X\beta_0\|_\infty \). Therefore, any attempt to relax the condition on \( \|\beta\|_\infty \) in each example is not very useful in our GLM context.

The proofs of the claims made in Examples 4.1–4.4 are provided in Section 4.4.

### 4.3 Posterior contraction rates

Once a prior distribution \( \Pi \) for \( \beta \) is specified, the posterior distribution \( \Pi(\cdot | Y^{(n)}) \) is defined by Bayes’ rule as

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

In this section, we study the posterior contraction rates in high-dimensional GLMs under suitable assumptions on the design matrix \( X \) and the true regression coefficients \( \beta_0 \).

As mentioned above, we need to assume that the true parameter satisfies \( \|X\beta_0\|_\infty \leq M \) for a sufficiently large \( M > 0 \). This is required for the true natural parameters to lie in a compact subset of \( \Theta \), so that extreme cases that the true variance is zero or infinity can be
excluded by (A1). This also guarantees the locally bi-Lipschitz continuity of \((h \circ b')\) near the true natural parameters by (A2). If (A1) and (A2) are globally satisfied rather than locally, as in the Gaussian regression setup, then the magnitude of \(\|X\beta_0\|_\infty\) needs not be restricted. Such a case is not of interest in this study and we make the condition \(\|X\beta_0\|_\infty \leq M\) to simplify the arguments. Under the Bayesian framework, note that Jiang (2007) assumes \(\|\beta_0\|_1 \leq M\), which is much stronger than ours and is rather restrictive unless \(s_0\) is bounded. On the other hand, our condition is easier to be met if there is some offset relation between \(\beta_0\) and \(X\), even if \(s_0\) is increasing. For the reason mentioned above, we mention that our main results below are not meant to be applied to sparse linear regression with Gaussian errors, as its theoretical properties have been investigated under milder conditions. One can easily refer to the abundant literature on it. Our main contribution is focused on more complicated GLMs.

We first state a lemma showing that the denominator of the posterior distribution is bounded below by an appropriate level with probability tending to one. The bounding level is directly related to our targeted contraction rate for the model.

**Lemma 4.1 (Evidence lower bound).** For a sufficiently large \(M > 0\) and a given prior distribution, suppose that \(\beta_0\) satisfies \(\|X\beta_0\|_\infty \leq M\), \(s_0 \log p = o(n)\) and (4.4). Then there exists a constant \(K_0 > 0\) such that

\[
P_0 \left( \int_{\mathbb{R}^p} \Lambda_n(\beta) d\Pi(\beta) \geq \pi_p(s_0) e^{-K_0 s_0 \log p} \right) \to 1. \tag{4.6} \]

As in other Bayesian settings, Lemma 4.1 plays an important role in deriving the posterior contraction rates. The result is used to derive our main results on the effective dimension and the posterior contraction rate in Theorem 4.1 and Theorem 4.2 below.

Next, we provide a theorem which states that the effective dimension is not much higher than the true one. The result allows us to restrict the attention to models with relatively small sizes, rather than the full dimension \(p\), by which we can effectively control the closeness of the two densities with respect to the root-average-squared Hellinger metric. The claim is formalized in the following theorem below.

**Theorem 4.1 (Effective dimension).** For a sufficiently large \(M > 0\) and a given prior distribution, suppose that \(\beta_0\) satisfies \(\|X\beta_0\|_\infty \leq M\), \(s_0 \log p = o(n)\), and (4.4). Then there exists a constant \(K_1 > 0\) such that

\[
\mathbb{E}_0 \Pi \left( \beta : s_\beta > K_1 s_0 \mid Y^{(n)} \right) \to 0. \]
It should be noted that the constant $K_1$ in the threshold is free of any assumption about the design matrix. A similar result will also be shown for the contraction rate relative to the squared loss of the linear predictor in Theorem 4.3. 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 without the identifiability assumption on the design matrix in Section 4 of Castillo et al. (2015) can be related to ours, where certain size restriction of the true linear predictor $\|X\beta_0\|_\infty$ is also required. Note that Gao et al. (2015) improves this finding by obtaining assumption-free results for sparse linear regression in Section 5.3 of their paper.

Theorem 4.1 rules out the models of large dimension with probability tending to one. With the help of this assertion, we are now ready to study the posterior contraction of the high-dimensional GLMs. The next theorem shows that the posterior distribution of $\beta$ contracts to the true one at the desired rate relative to the root-average-squared Hellinger metric. Derivation of this posterior contraction requires enough separation of the truth and other values in some suitable sieve, a subset of the parameter space $\mathbb{R}^p$. For the root-average-squared Hellinger metric, there always exists an exponentially powerful test for the truth against convex hulls that are sufficiently far away from the true parameter value (Ghosal and van der Vaart, 2007). The claim thus follows by controlling the metric entropy of a sieve measured by the root-average-squared Hellinger metric and verifying that a sieve grows sufficiently fast so that the residual probability is exponentially small.

**Theorem 4.2 (Contraction rate, Hellinger).** For a sufficiently large $M > 0$ and a given prior distribution, suppose that $\beta_0$ satisfies $\|X\beta_0\|_\infty \leq M$, $s_0 \log p = o(n)$, (4.4), and (4.5). Then there exists a constant $K_2 > 0$ such that

$$E_0 \Pi \left( \beta : H_n(\beta, \beta_0) > K_2 \sqrt{s_0 \log p \over n} \bigg| Y^{(n)} \right) \to 0.$$ 

It is worthwhile to compare our results in Theorem 4.2 with the posterior contraction rates obtained in Jiang (2007). Corollary 1 of Jiang (2007) implies that if $s_0 = o((\log n)^{b_1})$ 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 i.i.d. model in the chapter is given by $n^{-(1-b_2)/2} (\log n)^{b_1/2}$. The same rate is obtained by plugging in those bounds for the sequences in the rate given in Theorem 4.2, although implication of our rate is not exactly the same since we use the root-average-squared Hellinger metric with the model treated as a non-i.i.d. class. Similarly, we can also recover the rates in Corollary 2 of Jiang (2007). Note again that the assumption $\|\beta_0\|_1 \lesssim 1$ in Jiang (2007) is much stronger than ours.

Theorem 4.2 implies posterior consistency and the contraction rates in terms of the root-
average-squared Hellinger metric between the fitted density with $\beta$ and the true one. However, due to the vagueness of the metric used above, the assertion says nothing explicitly about the closeness of $\beta$ and $\beta_0$ in its current form. Our main purpose is to go beyond Jiang (2007) and Liang et al. (2013) who only studied posterior contraction relative to the Hellinger metric. We thus recover the contraction rates for $\beta$ with respect to more concrete metrics, under an additional boundedness requirement of the sparsity of the true parameter $\beta_0$. The recovery is relatively easy in our fixed design setup.

We clarify the required condition below:

$$\frac{s_0\sqrt{\log p}||X||_\infty}{\phi_1((K_1+1)s_0)\sqrt{n}} \to 0. \quad (4.7)$$

The compatibility number in the denominator is bounded below under a mild condition on the design matrix (e.g., see Example 7 of Castillo et al. (2015)). The quantity $||X||_\infty$ is often assumed to be uniformly bounded. Then the sufficient condition for the given one is $s_0\sqrt{\log p} = o(\sqrt{n})$, which is typically mild and necessary for a good interpretation for the $\ell_1$-contraction rate in the high-dimensional scenario. The recovery theorem is formally stated below.

**Theorem 4.3 (Recovery).** For a sufficiently large $M > 0$ and a given prior distribution, suppose that $\beta_0$ satisfies $||X\beta_0||_\infty \leq M$, $s_0\log p = o(n)$, (4.4), (4.5) and (4.7). Then there exists a constant $K_3 > 0$ such that

$$E_0\Pi(\beta : ||\beta - \beta_0||_1 > \frac{K_3s_0}{\phi_1((K_1+1)s_0)\sqrt{\log p/n}} Y^{(n)}) \to 0,$$

$$E_0\Pi(\beta : ||\beta - \beta_0||_2 > \frac{K_3}{\phi_2((K_1+1)s_0)\sqrt{s_0\log p/n}} Y^{(n)}) \to 0,$$

$$E_0\Pi(\beta : ||X(\beta - \beta_0)||_2 > K_3\sqrt{s_0\log p} Y^{(n)}) \to 0,$$

Similar to the dimensionality recovery result in Theorem 4.1, the constant in the threshold in the third assertion of Theorem 4.3 is free of any assumption on the design matrix. As mentioned earlier, our results are thus more in line with those without the identifiability assumption 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. These numbers are, however, usually bounded away from zero under typical regression settings; we refer to Example 7 of Castillo et al. (2015). The compatibility numbers are then removed from the rates, by which we have a clearer interpretation of the assertion.

119
It is also worth to compare the rates in Theorem 4.3 with those studied in the frequentist literature to glance the optimality of our results. Recently, Abramovich and Grinshtein (2016) obtained the convergence rate \( s_0 \log(ep/s_0) \) with respect to the Kullback-Leibler risk for the high-dimensional GLMs, under generally stronger conditions than ours. They only considered canonical link functions and claimed that the rate is minimax-optimal under certain limited situations where \( p = O(n) \) and the columns of the design matrix are not highly correlated (see their Corollary 1). It is readily seen that from the calculation of the Kullback-Leibler divergence in the proof of Lemma 4.1, the Kullback-Leibler divergence of \( f_\beta \) from \( f_{0} \) can be matched with \( \|X(\beta - \beta_0)\|_2^2 \) up to constants under (A1) and (A2). This implies that our rates are minimax-optimal up to logarithmic factors in the same limited situations.

4.4 Proofs

In this section, the technical proofs of the main results of this chapter are provided. Let 
\[ K(f_{0,i}, f_{\beta,i}) = E_{0} \log(\frac{f_{0,i}}{f_{\beta,i}}) \]
and 
\[ V(f_{0,i}, f_{\beta,i}) = E_{0}\{ (\log(\frac{f_{0,i}}{f_{\beta,i}}) - K(f_{0,i}, f_{\beta,i}))^2 \} \]
denote the Kullback-Leibler divergence and variation between \( f_{0,i} \) and \( f_{\beta,i} \), respectively. The notation \( E_{\beta} \) denotes the expectation operator with an arbitrary parameter \( \beta \). For a metric space \((F, d)\), let 
\[ N(\epsilon, F, d) \]
denote the \( \epsilon \)-covering number, the minimal number of \( \epsilon \)-balls that is needed to cover \( F \). Hereafter we write \( \eta_i := X_i^T \beta \) and \( \eta_{0i} := X_i^T \beta_0 \). We also write \( \theta_{0i} = b^{-1}(h^{-1}(\eta_{0i})) \) for the natural parameter with true \( \beta_0 \).

Proof of the assertions in Examples 4.1–4.4. We first verify (4.4) for the prior distributions in the examples. Note that for every \( S \subset \{1, \ldots, p\} \) that is not an empty set,

\[ G_S \left( \|Z_S - \beta_{0, S}\|_1 \leq M_2\|X\|^{-1}_{\infty} \sqrt{(s_0 \log p)/n} \right) \geq G_S (\|Z_S - \beta_{0, S}\|_{\infty} \leq c_n), \]

where \( c_n = M_2(s\|X\|_{\infty})^{-1} \sqrt{(s_0 \log p)/n} \). As the right hand side of the display is a probability on an \( s \)-dimensional hypercube with edge length \( 2c_n \), it is easy to see that

\[ G_S (\|Z_S - \beta_{0, S}\|_{\infty} \leq c_n) \geq (2c_n)^s \inf_{b \in \mathbb{R}^s: \|b\|_{\infty} \leq c_n} g_S(\beta_{0, S} + b). \]

Since \( s \log(2c_n) \geq -s \log p \geq -M_1 s_0 \log p \), it suffices to show that the logarithm of the infimum in the display is bounded below by a constant multiple of \( -s_0 \log p \), using the four prior distributions as follows.

(i) Uniform prior. The claim easily follows from the density function as sup\{\|\beta_{0, S} + b\|_{\infty} : b \in \mathbb{R}^s, \|b\|_{\infty} < c_n\} \leq \|\beta_0\|_{\infty} + c_n = o(p^B) \), which means that the domain is included in the support of the prior.
(ii) Laplace prior. We have that
\[
\inf_{b \in \mathbb{R}^s} \left( \frac{\lambda}{2} \right)^s \exp(-\lambda \|\beta_0, S + b\|_1) \geq \left( \frac{1}{2 B_1 p B_2} \right)^s \exp(-\lambda \|\beta_0\|_1 - \lambda s c_n). \\
\]
and hence the claim follows since \(\lambda s c_n \leq B_3 s c_n \leq s_0 \log p\) and \(s \leq s_0\).

(iii) Elliptical Laplace prior. We have that
\[
\inf_{b \in \mathbb{R}^s : \|b\|_\infty < c_n} \left( \frac{\lambda}{\sqrt{2\pi}} \right)^s \det(V_S)^{-1/2} \frac{\Gamma(s/2 + 1)}{\Gamma(s + 1)} \exp \left\{ -\lambda \|V_S^{-1/2}(\beta_{0,S} + b)\|_2 \right\} \geq \left( \frac{1}{\sqrt{2\pi B_1 p B_2}} \right)^s \|V_S\|_{sp}^{-s/2} (s!)^{-1} \exp \left\{ -\lambda \|V_S^{-1/2}\beta_{0,S}\|_2 + \sqrt{s} \|V_S^{-1/2}\|_{sp} c_n \right\}. \\
\]
Note that \(\lambda \sqrt{s} \|V_S^{-1/2}\|_{sp} c_n \leq B_3 \sqrt{s} c_n \leq s_0 \log p\) and \(\lambda \|V_S^{-1/2}\beta_{0,S}\|_2 \leq \lambda \|V_S^{-1/2}\|_{sp} \|\beta_0\|_2 \leq s_0 \log p\). The claim follows by the inequalities \(\log \|V_S\|_{sp} \leq \log p\), \(s! \leq p^s\), and \(s \leq M_1 s_0\).

(iv) Multivariate normal prior. We similarly have that
\[
\inf_{b \in \mathbb{R}^s : \|b\|_\infty < c_n} \left( \frac{\lambda}{\sqrt{2\pi}} \right)^s \det(V_S)^{-1/2} \frac{\Gamma(s/2 + 1)}{\Gamma(s + 1)} \exp \left\{ -\lambda \|V_S^{-1/2}(\beta_{0,S} + b)\|_2 \right\} \geq \left( \frac{1}{\sqrt{2\pi B_1 p B_2}} \right)^s \|V_S\|_{sp}^{-s/2} \exp \left\{ -\lambda \left( \|V_S^{-1/2}\beta_{0,S}\|_2 + s \|V_S^{-1/2}\|_{sp} c_n^2 \right) \right\}. \\
\]
Note that \(\lambda s \|V_S^{-1/2}\|_{sp} c_n^2 \leq B_3 s^2 c_n^2 \leq s_0 \log p\) and \(\lambda \|V_S^{-1/2}\beta_{0,S}\|_2 \leq \lambda \|V_S^{-1/2}\|_{sp} \|\beta_0\|_2 \leq s_0 \log p\). Therefore, the claim holds as \(\log \|V_S\|_{sp} \leq \log p\) and \(s \leq M_1 s_0\).

Next, we verify the second inequality of (4.5) for the prior distributions in Examples 4.1–4.4 as follows.

(i) Uniform prior. If \(M_2\) is chosen sufficiently larger than \(B\) such that the influence of \(\|X\|_\infty\) vanishes, then we have that \(G_S(\|Z_S - \beta_{0,S}\|_\infty > p M_2 / \|X\|_\infty - \|\beta_0\|_\infty) = 0\) and the claim holds with \(r_n = \infty\).

(ii) Laplace prior. Using the triangle inequality and the tail probability of Laplace distributions given by \(\int_{|x|>t} 2^{-1} e^{-\lambda |x|} dx = \exp(-\lambda t)\) for every \(t > 0\), we bound \(G_S(\|Z_S - \beta_{0,S}\|_\infty > p M_2 / \|X\|_\infty - \|\beta_0\|_\infty)\) by
\[
e^{-\lambda \|\beta_0\|_1} \int_{\{\beta_S: \|\beta_S - \beta_{0,S}\|_\infty > p M_2 / \|X\|_\infty - \|\beta_0\|_\infty \}} g_S(\beta_S - \beta_{0,S}) d\beta_S \\
\leq s \exp \left\{ -\lambda p M_2 / \|X\|_\infty + 2 s_0 \lambda \|\beta_0\|_\infty \right\}. \\
\]
Since \(s_0 \|\beta_0\|_\infty \leq \lambda^{-1} s_0 \log p \leq s_0 B_1 p B_2 \log p\), the claim holds by choosing \(M_2\) sufficiently large.

(iii) Elliptical Laplace prior. For the set \(\mathcal{B}_S = \{\beta_S : \|V_S^{-1/2}\|_{sp} \|V_S^{-1/2}(\beta_S - \beta_{0,S})\|_\infty > \\
\(p_{M_2}/\|X\|_{\infty} - \|\beta_0\|_{\infty}\), we have that by the triangle inequality,

\[
\int_{\mathcal{B}_S} g_S(\beta_S)d\beta_S \leq e^{\lambda\|V_{S}^{-1/2}\beta_{0,S}\|_2} \int_{\mathcal{B}_S} g_S(\beta_S - \beta_{0,S})d\beta_S.
\]

By the change of variable \(w_S = V_{S}^{-1/2}(\beta_S - \beta_{0,S})\), the right hand side is equal to

\[
e^{\lambda\|V_{S}^{-1/2}\beta_{0,S}\|_2} \int_{w_S:\|V_{S}^{-1/2}\beta_{0,S}\|_2 > p_{M_2}/\|X\|_{\infty} - \|\beta_0\|_{\infty}} \left(\frac{\lambda}{\sqrt{\pi}}\right)^s \frac{\Gamma(s/2 + 1)}{\Gamma(s + 1)} e^{-\lambda\|w_S\|^2_{\infty}} dw_S
\]

\[
\leq e^{\lambda\|V_{S}^{-1/2}\beta_{0,S}\|_2} \int_{w_S:\|w_S\|_{\infty} > p_{M_2}/\|X\|_{\infty} - \|\beta_0\|_{\infty}} \left(\frac{\lambda\sqrt{s}}{2}\right)^s e^{-\lambda\sqrt{s}\|w_S\|_{1_{\infty}}} dw_S,
\]

since \(\pi^{-s/2}\Gamma(s/2+1)/\Gamma(s+1) \leq 2^{-s}\) for \(s \geq 1\). Using the tail probability of Laplace distributions above, the display is further bounded by

\[
e^{\lambda\|V_{S}^{-1/2}\beta_{0,S}\|_2} s \exp \left\{ - \frac{\lambda}{\|V_{S}^{-1/2}\beta_{0,S}\|_2} \left( \frac{p_{M_2}}{\|X\|_{\infty} - \|\beta_0\|_{\infty}} \right) \right\}
\]

\[
\leq s \exp \left\{ - \frac{\lambda}{\|V_{S}^{-1/2}\beta_{0,S}\|_2} \frac{p_{M_2}}{\|X\|_{\infty} - \|\beta_0\|_{\infty}} + 2\lambda s_0 \|V_{S}^{-1/2}\beta_{0,S}\|_2 \right\},
\]

since \(\|V_{S}^{-1}\|_{sp} \geq \|V_{S}\|_{sp}^{-1}\). Note that \(\|\beta_0\|_{\infty} \lesssim \lambda^{-1} \sqrt{\rho_{\min}(V_{S})}\log p\) and consider the bounds for \(\lambda\) and \(\|V_{S}^{-1}\|_{sp}\). The claim thus follows if \(M_2\) is sufficiently large.

(iv) Multivariate normal prior. By the inequality \(\|a - b\|_2/2 - \|b\|_2^2 \leq \|a\|_2^2\), the probability \(G_S(\|Z_S - \beta_0,S\|_{\infty} > p_{M_2}/\|X\|_{\infty} - \|\beta_0\|_{\infty})\) is bounded by

\[
\int_{\mathcal{B}_S} g_S(\beta_S)d\beta_S \leq \exp \left\{ \frac{\lambda}{2}\|V_{S}^{-1/2}\beta_{0,S}\|_2^2 \right\} 2^{s/2} \int_{\mathcal{B}_S} \bar{g}_S(\beta_S)d\beta_S,
\]

where \(\bar{g}_S\) is the \(s\)-dimensional multivariate normal density with mean \(\beta_{0,S}\) and covariance matrix \(2\lambda^{-1}V_{S}\). Using the tail probability of normal distributions, the integral term in the right hand side is bounded by

\[
2s \exp \left\{ - \frac{\lambda}{4\|V_{S}\|_{sp}} \left( \frac{p_{M_2}}{\|X\|_{\infty} - \|\beta_0\|_{\infty}} \right) \right\} \lesssim s \exp \left\{ - \frac{1}{4B_1p^{B_2+a}} \left( \frac{p_{M_2}}{\|X\|_{\infty} - \|\beta_0\|_{\infty}} \right) \right\}.
\]

Since \(\|\beta_0\|_{\infty} \lesssim \sqrt{\lambda^{-1}\rho_{\min}(V_{S})}\log p \leq \sqrt{B_1p^{B_2+a}}\log p\), the claim follows as long as \(M_2\) is chosen sufficiently large.

**Proof of Lemma 4.1.** If \(s_0 = 0\), it is easy to see that \(\int \Lambda_n(\beta)d\Pi(\beta) \geq \Lambda_n(0)\pi_p(0) = \pi_p(0)\).

122
Hence we only consider the case $s_0 > 0$ in what follows. Let $\epsilon_n = \sqrt{(s_0 \log p)/n}$ and

$$
B_n = \left\{ \beta \in \mathbb{R}^p : \frac{1}{n} \sum_{i=1}^{n} K(f_{0,i}, f_{\beta,i}) \leq C_1 \epsilon_n, \frac{1}{n} \sum_{i=1}^{n} V(f_{0,i}, f_{\beta,i}) \leq C_1 \epsilon_n^2 \right\},
$$

for a sufficiently large $C_1$. By Lemma 10 of Ghosal and van der Vaart (2007), we have that for any $C > 0$,

$$
\mathbb{P}_0 \left( \int_{B_n} \Lambda_n(\beta) d\Pi(\beta) \geq e^{-(1+C)C_1 n \epsilon_n^2} \Pi(B_n) \right) \geq 1 - \frac{1}{C^2 C_1^2 n \epsilon_n^2}.
$$

We first show that if $C_1$ is sufficiently large, we have that $B_n \supset \{ \beta \in \mathbb{R}^p : \|\beta - \beta_0\|_1 \leq M_2 \|X\|^{-1}_\infty \epsilon_n \}$ for $M_2$ in (4.4). Note that if $\|X\|_\infty \beta - \beta_0\|_1 \leq M_2 \epsilon_n$, then

$$
\epsilon_n \gtrsim \max_{1 \leq i \leq n} \|X_i\|_\infty \|\beta - \beta_0\|_1 \geq \max_{1 \leq i \leq n} |X_i^T(\beta - \beta_0)| \gtrsim \max_{1 \leq i \leq n} |\theta_i - \theta_0i|, \quad (4.8)
$$

since $\epsilon_n$ is decreasing, $\|X\beta_0\|_\infty$ is bounded, and the function $b^{-1}(h^{-1}(\cdot))$ is locally Lipschitz continuous by (A2). By direct calculations, the Kullback-Leibler divergence and variation of the exponential family of distributions are given by

$$
K(f_{0,i}, f_{\beta,i}) = \mathbb{E}_0 \log \frac{f_{0,i}}{f_{\beta,i}} = \frac{1}{\tau_i} \left\{ (\theta_{0i} - \theta_i) b'(\theta_{0i}) - b(\theta_{0i}) + b(\theta_i) \right\},
$$

$$
V(f_{0,i}, f_{\beta,i}) = \mathbb{E}_0 \left\{ \left( \log \frac{f_{0,i}}{f_{\beta,i}} - K(f_{0,i}, f_{\beta,i}) \right)^2 \right\} = \frac{b''(\theta_{0i})}{\tau_i} (\theta_i - \theta_0i)^2.
$$

Note that by the Taylor expansion, $K(f_{0,i}, f_{\beta,i}) = b''(\theta^*_i)(\theta_i - \theta_0i)^2/(2\tau_i)$ for some $\theta^*_i$ between $\theta_{0i}$ and $\theta_i$. Since $b''(\theta^*_i)$ is bounded if $\theta_i$ is close to $\theta_{0i}$ due to $\|X\beta_0\|_\infty \lesssim 1$ and (A1), both $K(f_{0,i}, f_{\beta,i})$ and $V(f_{0,i}, f_{\beta,i})$ are bounded by a constant multiple of $(\theta_i - \theta_0i)^2$ for $\theta_i$ that is close to $\theta_{0i}$. Thus, (4.8) implies that if $C_1$ is sufficiently large,

$$
\Pi(B_n) \geq \Pi \left\{ \beta \in \mathbb{R}^p : \|\beta - \beta_0\|_1 \leq M_2 \|X\|^{-1}_\infty \sqrt{(s_0 \log p)/n} \right\}.
$$

For any $s_0 > 0$, it is easy to see that the last display is bounded below by

$$
\pi_p(s_0) \frac{1}{(p)^s_0} \int_{\mathbb{R}^p} g_{s_0}(\beta) d\beta_{s_0} \leq \pi_p(s_0) \frac{1}{(p)^s_0} \mathcal{G}_{s_0} \left( \|Z_{s_0} - \beta_{0,s_0}\|_1 \leq M_2 \|X\|^{-1}_\infty \sqrt{(s_0 \log p)/n} \right).
$$

Since $(p)^s_0 \leq p^{s_0}$, by (4.4), the display is further bounded below by $\pi_p(s_0) \exp(-C_2 s_0 \log p)$ for some $C_2 > 0$. This completes the proof.

\[\square\]
Proof of Theorem 4.1. Define the set $\mathcal{B} = \{\beta : s_{\beta} > \bar{s}\}$ for a given integer $\bar{s} \geq s_0$. It follows from (4.3) that
\[
\Pi(\mathcal{B}) = \sum_{s=\bar{s}+1}^{p} \pi_p(s) \leq \sum_{s=\bar{s}+1} \left( \frac{A_2}{p^{A_4}} \right)^{s-s_0} \leq \pi_p(s_0) \left( \frac{A_2}{p^{A_4}} \right)^{\bar{s}+1-s_0} \sum_{j=0}^{\infty} \left( \frac{A_2}{p^{A_4}} \right)^j.
\]

Let $\mathcal{E}_n$ be the event in (4.6). Then by Lemma 4.1, we have that
\[
E_0 \Pi(\mathcal{B} | Y^{(n)}) \mathbb{1}_{\mathcal{E}_n} = E_0 \left[ \frac{\int_{\mathcal{B}} \Lambda_n(\beta) d\Pi(\beta)}{\int \Lambda_n(\beta) d\Pi(\beta)} \mathbb{1}_{\mathcal{E}_n} \right]
\leq \pi_p(s_0)^{-1} \exp(C_1 s_0 \log p) \Pi(\mathcal{B})
\leq \exp \{ (\bar{s} + 1 - s_0)(\log A_2 - A_4 \log p) + C_1 s_0 \log p \},
\]
for some constant $C_1 > 0$ and sufficiently large $p$. Choosing $\bar{s} = C_2 s_0$ for an integer $C_2 > 1$ such that $A_4(C_2 - 1) > C_1$, the rightmost side of the display goes to zero. The proof is then complete since $E_0 \Pi(\mathcal{B} | Y^{(n)}) \leq E_0 \Pi(\mathcal{B} | Y^{(n)}) \mathbb{1}_{\mathcal{E}_n} + P_0(\mathcal{E}_n^c)$ and $P_0(\mathcal{E}_n^c) \rightarrow 0$ by Lemma 4.1. \qed

Proof of Theorem 4.2. Let $\mathcal{A}_n = \{\beta \in \mathbb{R}^p : s_\beta \leq K_1 s_0\}$. Then, for every $\epsilon > 0$,
\[
E_0 \Pi \left( \beta \in \mathbb{R}^p : H_n(\beta, \beta_0) > \epsilon \ | Y^{(n)} \right)
\leq E_0 \Pi \left( \beta \in \mathcal{A}_n : H_n(\beta, \beta_0) > \epsilon \ | Y^{(n)} \right) + E_0 \Pi \left( \mathcal{A}_n^c | Y^{(n)} \right).
\]

Since the second term on the right hand side goes to zero by Theorem 4.1, it suffices to show that the first term goes to zero.

Let $\mathcal{A}_n^* = \{\beta \in \mathcal{A}_n : \|\beta - \beta_0\|_\infty \leq \epsilon p^{-M_2} \|X\|_\infty\}$ for a sufficiently large $M_2$. By Lemma 2 of Ghosal and van der Vaart (2007), there exists a test $\varphi_n$ such that for any $\beta_1$ with $H_n(\beta_0, \beta_1) > \epsilon$,
\[
E_0 \varphi_n \leq \exp(-n \epsilon^2 / 2), \quad \sup_{\beta : H_n(\beta, \beta_0) \leq \epsilon / 18} E_\beta (1 - \varphi_n) \leq \exp(-n \epsilon^2 / 2).
\]

To use Lemma 9 of Ghosal and van der Vaart (2007), we now show that log $N(\epsilon_n/36, \mathcal{A}_n^*, H_n) \lesssim n \epsilon_n^2$ for $\epsilon_n = \sqrt{(s_0 \log p) / n}$. Let
\[
\zeta_i(\eta_i) := \frac{H^2(f_{\beta,i}, f_{0,i})}{2} = 1 - \exp \left\{ \frac{1}{\tau_i} \left[ b \left( \frac{\theta_i + \theta_0}{2} \right) - \frac{b(\theta_i) + b(\theta_0)}{2} \right] \right\}.
\] (4.9)

We first verify that there exists a constant $C_1 > 0$ such that for every $i \leq n$,
\[
\zeta_i(\eta_i) \leq C_1 (\eta_i - \eta_0)^2, \quad \eta_i \in \mathbb{R},
\]
that is, \(\zeta(\eta_i)\) is majorized by a constant multiple of the quadratic term \((\eta_i - \eta_{0i})^2\). This can be expressed as

\[
\bar{\zeta}(\eta_i, \eta_{0i}) := \frac{2\zeta_i(\eta_i)}{(\eta_i - \eta_{0i})^2} \leq 2C_1, \quad \eta_i \in \mathbb{R}.
\] (4.10)

The display is true if \(\eta_i\) is bounded away from \(\eta_{0i}\) as \(\zeta(\eta_i)\) is bounded by 1. Now, note that

\[
\lim_{\eta_i \to \eta_{0i}} \bar{\zeta}(\eta_i, \eta_{0i}) = \zeta''(\eta_{0i}) = \left(\frac{d^2}{d\theta_i^2} \zeta_i(\eta_i) \bigg|_{\eta_i = \eta_{0i}}\right) \left(\frac{d\theta_i}{d\eta_i} \bigg|_{\eta_i = \eta_{0i}}\right)^2,
\]

since \(\left(d\zeta_i(\eta_i)/d\theta_i\right)|_{\eta_i = \eta_{0i}} = 0\). The first factor on the rightmost side is \(b''(\theta_{0i})/(4\tau_i)\) by direct calculations, while the second factor is bounded as \(b^{-1}(h^{-1}(\cdot))\) is locally Lipschitz continuous by (A2) and we assume \(\|X_{\beta_0}\|_\infty \leq 1\). Hence, the last display is bounded due to (A1), which concludes that (4.10) holds for every \(i \leq n\). Therefore,

\[
H_n^2(\beta, \beta_0) \leq \frac{2C_1}{n} \sum_{i=1}^n (\eta_i - \eta_{0i})^2 \leq \frac{2C_1}{n} \sum_{i=1}^n \|X_i\|_\infty^2 \|\beta - \beta_0\|_1^2 \leq p^2 \|X\|_\infty^2 \|\beta - \beta_0\|_\infty^2,
\]

and thus for some \(C_2 > 0, N(\epsilon_n/36, A_n^*, H_n)\) is bounded by

\[
N \left(\frac{C_2\epsilon_n}{36p\|X\|_\infty}, A_n^*, \|\cdot\|_\infty\right) \leq \left(\frac{p}{K_{180}}\right) \left(\frac{3pM_2+1}{C_2\epsilon_n/36}\right)^{K_{180}} \leq \left(\frac{3pM_2+2}{C_2\epsilon_n/36}\right)^{K_{180}}.
\]

It follows from the last display that \(\log N(\epsilon_n/36, A_n^*, H_n) \lesssim n\epsilon_n^2\). Now, Lemma 9 of Ghosal and van der Vaart (2007) implies that for every \(\epsilon > \epsilon_n\), there exists a test \(\bar{\psi}_n\) such that

\[
\mathbb{E}_0\bar{\psi}_n \leq \frac{1}{2} \exp(C_2n\epsilon_n^2 - \epsilon^2/2), \quad \sup_{\beta \in A_n^*: H_n(\beta, \beta_0) > \epsilon} \mathbb{E}_\beta(1 - \bar{\psi}_n) \leq \exp(-n\epsilon^2/2).
\]

Observe that we have \(-\log \pi_p(s_0) \lesssim s_0 \log p\) as \(\pi_p(0)\) is bounded away from zero. Hence for \(\mathcal{E}_n\) the event in (4.6) and some constant \(C_3 > 0\),

\[
\mathbb{E}_0\Pi \left(\beta \in \mathcal{A}_n : H_n(\beta, \beta_0) > \epsilon \mid Y^{(n)}\right) \leq \mathbb{E}_0\Pi \left(\beta \in \mathcal{A}_n : H_n(\beta, \beta_0) > \epsilon \mid Y^{(n)}\right) \mathbb{I}_{\mathcal{E}_n}(1 - \bar{\psi}_n) + \mathbb{E}_0\bar{\psi}_n + \mathbb{P}_0\mathcal{E}_n^c
\]

\[
\leq \left\{ \sup_{\beta \in \mathcal{A}_n^*: H_n(\beta, \beta_0) > \epsilon} \mathbb{E}_\beta(1 - \bar{\psi}_n) + \Pi(\mathcal{A}_n \setminus \mathcal{A}_n^*) \right\} e^{C_3n\epsilon_n^2} + \mathbb{E}_0\bar{\psi}_n + \mathbb{P}_0\mathcal{E}_n^c,
\]

by Lemma 4.1. The last term \(\mathbb{P}_0\mathcal{E}_n^c\) on the right most side goes to zero by Lemma 4.1. The remaining terms other than \(\Pi(\mathcal{A}_n \setminus \mathcal{A}_n^*)e^{C_3n\epsilon_n^2}\) go to zero by choosing \(\epsilon = C_4\epsilon_n\) for a sufficiently
large $C_4$. Next, to complete the proof, note that

$$
\Pi(\mathcal{A}_n \setminus \mathcal{A}_n^* ) = \Pi \{ \beta \in \mathbb{R}^p : s_\beta \leq K_1 s_0 , \| \beta - \beta_0 \|_\infty > p^{M_2} / \| X \|_\infty \} \\
= \sum_{s : s \leq K_1 s_0} \frac{\pi_p(s)}{\binom{p}{s}} \int_{\{ \beta_S : \beta_S \sim \beta_0, s \}} g_S(\beta_S) d\beta_S \\
\leq \sum_{s : s \leq K_1 s_0} \frac{(A_2 p^{-A_4})^s}{\binom{p}{s}} \| Z_S - \beta_0, s \|_\infty > p^{M_2} / \| X \|_\infty - \| \beta_0 \|_\infty \}.
$$

By (4.5), the rightmost side is bounded by $e^{-r n} \sum_{s=0}^p (A_2 p^{-A_4})^s \lesssim e^{-r n}$. Therefore, we see that $\Pi(\mathcal{A}_n \setminus \mathcal{A}_n^*) e^{C_3 n r^2}$ goes to zero as $p \to \infty$, which completes the proof.

To prove Theorem 4.3, we first give the following lemma. The lemma implies that if the Hellinger metric between $f_{\beta,i}$ and $f_{0,i}$ is small enough, then $|X_i^T (\beta - \beta_0)|$ is bounded as long as $|X_i^T \beta_0|$ is bounded.

**Lemma 4.2.** For any bounded $|\theta_0|$, $H(f_{\beta,i}, f_{0,i}) \to 0$ implies that $|\theta_i - \theta_0| \to 0$.

**Proof.** Note that by (4.9),

$$
1 - \frac{H^2(f_{\beta,i}, f_{0,i})}{2} = \exp \left\{ \frac{1}{\tau_i} \left[ b \left( \frac{\theta_i + \theta_0}{2} \right) - b(\theta_i) + b(\theta_0) \right] \right\}.
$$

To prove the assertion, it suffices to show that for $\theta_0$ such that $|\theta_0| \lesssim 1$, the right hand side is strictly quasi-concave in $\theta_i$ and has the global minimum zero at $\theta_i = \theta_0$. For a given $\theta_i \neq \theta_0$, let $\bar{\theta}_i(t) = t \theta_i + (1-t) \theta_0$, $t \in (0,1)$, and define

$$
g_i(t) := -\log \left\{ 1 - \frac{H^2(\bar{f}_{\beta,i}, f_{0,i})}{2} \right\} = \frac{1}{\tau_i} \left\{ b(\bar{\theta}_i(t) + \theta_0) - b \left( \frac{\bar{\theta}_i(t) + \theta_0}{2} \right) \right\},
$$

where $\bar{f}_{\beta,i}$ is the density with $\bar{\theta}_i(t)$ as the parameter. It is clear that $g_i(0) = 0$. Hence, it suffices to show that $g_i'(0) = 0$ and $g_i(t)$ is strictly increasing for every $t \in (0,1)$ to prove the claim. Observe that by direct calculations,

$$
g_i'(t) = \frac{k_i(t) - k_i(t/2)}{2} , \quad k_i(t) = (\theta_i - \theta_0) b'(t(\theta_i - \theta_0) + \theta_0).
$$

Note that $b'$ is increasing on any bounded interval, and hence $k_i(t)$ is increasing in $t \in (0,1)$. This gives that $g_i'(t) > 0$ for $t \in (0,1)$ while $g_i'(0) = 0$ can be easily checked, which leads to the desired assertion.
Proof of Theorem 4.3. By the Taylor expansion, for $\zeta_i$ defined in (4.9),

$$\zeta_i(\eta) = \zeta_i(\eta_0i) + \zeta'_i(\eta_0i)(\eta - \eta_0i) + \frac{\zeta''_i(\eta_0i)}{2}(\eta - \eta_0i)^2,$$

where $\eta_0^*$ lies between $\eta_i$ and $\eta_0i$. As in the proof of Theorem 4.2, it can be easily checked that for every $i \leq n$, $\zeta'_i(\eta_0i) = 0$ and $\zeta''_i(\eta_0^*) \geq 1$ if $|\eta_0^*|$ is bounded due to (A1). Note that by Lemma 4.2, there exists some small $\epsilon > 0$ such that $\zeta_i(\eta_i) \leq \epsilon$ implies $|\eta_i - \eta_0i| \lesssim |\theta_i - \theta_0i| \lesssim 1$ if $|\theta_0i|$ is bounded, since $h(b'(\cdot))$ is locally Lipschitz continuous by (A2). This verifies that $|\eta_0i|$ is bounded if $\zeta_i(\eta_i)$ is small enough. Therefore, there exists a constant $C_1 > 0$ such that for some small $\delta > 0$,

$$\zeta_i(\eta_i) = \zeta_i(\eta) - \zeta_i(\eta_0i) \geq C_1 \{(\eta_i - \eta_0i)^2 \wedge \delta\}.$$ 

Let $I_{n, \delta} = \{i \leq n : (\eta_i - \eta_0i)^2 \geq \delta\}$ and $|I_{n, \delta}|$ be its cardinality. Then,

$$e_n^2 \geq H_n(\beta, \beta_0) \geq \frac{C_1 \delta |I_{n, \delta}|}{n} + \frac{C_1}{n} \sum_{i \notin I_{n, \delta}} (\eta_i - \eta_0i)^2.$$

(4.11)

Note that for some $C_2 > 0$,

$$\frac{1}{n} \sum_{i \notin I_{n, \delta}} (\eta_i - \eta_0i)^2 \geq \frac{1}{n} \sum_{i=1}^{n} (\eta_i - \eta_0i)^2 - \frac{|I_{n, \delta}|}{n} \max_{1 \leq i \leq n} (\eta_i - \eta_0i)^2 \geq \frac{\phi_1^2((K_1 + 1)s_0)}{(K_1 + 1)s_0} \|\beta - \beta_0\|_1^2 - \frac{e_n^2}{C_2 \delta} \max_{1 \leq i \leq n} \|X_i\|_1 \|\beta - \beta_0\|_1^2,$$

using the definition of $\phi_1$ and (4.11). Since $(s_0^2 \log p)\|X\|_1^2 / \phi_1^2((K_1 + 1)s_0) = o(n)$ by (4.7), the last display is further bounded below by a constant multiple of $\phi_1^2((K_1 + 1)s_0)\|\beta - \beta_0\|_1^2 / s_0$, which implies the first assertion of the theorem.

Note also that

$$e_n^2 \geq H_n(\beta, \beta_0) \geq \frac{C_1}{n} \|X(\beta - \beta_0)\|_2^2 - \frac{C_1}{n} \sum_{i \in I_{n, \delta}} |X_i^T(\beta - \beta_0)|^2.$$

Thus, using (4.11), we have that for some $C_3 > 0$,

$$\|X(\beta - \beta_0)\|_2^2 \lesssim n e_n^2 + |I_{n, \delta}| \max_{1 \leq i \leq n} |X_i^T(\beta - \beta_0)|^2 \leq n e_n^2 + C_3 n e_n^2 \|X\|_1^2 \|\beta - \beta_0\|_1^2.$$

Note that we have $\|\beta - \beta_0\|_1^2 \lesssim (s_0^2 \log p) / [n \phi_1((K_1 + 1)s_0)]$ by the first assertion of the theorem. Since $(s_0^2 \log p)\|X\|_1^2 / \phi_1^2((K_1 + 1)s_0) \lesssim n$, the last display is bounded by a multiple of $n e_n^2$, 

127
which concludes the third assertion of the theorem. The second assertion then follows from the
definition of $\phi_2$.\qed
REFERENCES


133


