A General Framework for Bayes Structured Linear Models *

Chao Gao[1], Aad W. van der Vaart[2], and Harrison H. Zhou[1]

[1] *Yale University*
[2] *Leiden University*

June 3, 2015

Abstract

High dimensional statistics deals with the challenge of extracting structured information from complex model settings. Compared with the growing number of frequentist methodologies, there are rather few theoretically optimal Bayes methods that can deal with very general high dimensional models. In contrast, Bayes methods have been extensively studied in various nonparametric settings and rate optimal posterior contraction results have been established. This paper provides a unified approach to both Bayes high dimensional statistics and Bayes nonparametrics in a general framework of structured linear models. With the proposed two-step model selection prior, we prove a general theorem of posterior contraction under an abstract setting. The main theorem can be used to derive new results on optimal posterior contraction under many complex model settings including stochastic block model, graphon estimation and dictionary learning. It can also be used to re-derive optimal posterior contraction for problems such as sparse linear regression and nonparametric aggregation, which improve upon previous Bayes results for these problems. The key of the success lies in the proposed two-step prior distribution. The prior on the parameters is an elliptical Laplace distribution that is capable to model signals with large magnitude, and the prior on the models involves an important correction factor that compensates the effect of the normalizing constant of the elliptical Laplace distribution.

Keywords. Oracle inequality, Stochastic block model, Graphon, Sparse linear regression, Aggregation, Dictionary learning, Posterior contraction

1 Introduction

Theory for posterior distribution has been extensively investigated in Bayes nonparametrics recently. Important works such as [6, 5, 24, 25, 48, 53, 27, 13] established that the posterior distribution contracts to a small neighborhood of the truth under proper conditions on likelihood functions and priors. These works bridge the gap between frequentist and Bayesian views of statistics from a fundamental perspective.

*funding

Despite the success of theoretical advancements of Bayes nonparametrics, there are not many theories developed for Bayes high dimensional statistics. A few exceptions are [14] on sparse Gaussian sequence model, [4] on bandable precision matrix estimation and [22] on sparse PCA. Recently, [15] established posterior contraction rates for sparse linear regression with a spike and slab prior under comparable assumptions of the Lasso estimator [49, 7]. The results of [15] include posterior contraction rates for prediction error and estimation error, oracle inequalities and model selection consistency. However, sparse linear regression is only one example of high dimensional statistics. There is an indispensable demand of a Bayes theory on more complicated model settings such as dictionary learning, stochastic block model and multi-task learning, etc. It is not clear whether the method and the analysis used in [15] can be extended to these more complex settings.

This paper provides a unified approach for both Bayes high dimensional and Bayes nonparametric statistics in a general framework of structured linear models. We first establish a unified view of various high-dimensional and nonparametric models, and then propose a single prior distribution for all models considered in our framework. We establish optimal rates of convergence of the posterior distributions under appropriate conditions. The results directly lead to minimax posterior contraction rates in stochastic block model, biclustering, sparse linear regression, regression with group sparsity, multi-task learning and dictionary learning. Moreover, we also derive a general posterior oracle inequality that allows arbitrary model misspecification. Applications of the posterior oracle inequality let us obtain posterior contraction rates even for models that are not included in our framework. Examples considered in this paper are nonparametric graphon estimation, linear regression with approximate sparsity, wavelet estimation under Besov space and various forms of nonparametric aggregation.

In the heart of our general theory is a proposed two-step prior distribution, which naturally accommodates the structured linear model by first modeling the structure and then modeling the parameters. This two-step modeling strategy was first investigated by [14] for Gaussian sequence models. A key ingredient of the prior distribution is that the tail of the distribution on the model parameter Q cannot be too light [14, 15], which motivates [14, 15] to use the independent Laplace prior with density proportional to $\exp(-\lambda\|Q\|_1)$ on the parameter. Though the prior distribution leads to optimal posterior contraction rates in Gaussian sequence model [14], it requires some excessive assumptions on the design matrix when it is applied to sparse linear regression [15]. The proposal in this paper is the elliptical Laplace distribution with density proportional to $\exp(-\lambda\|\mathscr{X}(Q)\|)$ for some linear operator $\mathscr{X}(\cdot)$. Note that we use the ℓ_2 norm instead of the ℓ_1 norm. With this choice, not only are we able to weaken the assumptions in [15], but we can also solve a more general class of problems in a unified way. To compensate the influence of the normalizing constant of an elliptical Laplace distribution, a correction factor on the prior mass is considered in the model selection step.

The paper is organized as follows. Section 2 introduces the general framework of structured linear models. A general prior distribution is proposed in Section 3. Section 4 presents

the main results of the paper including rate optimal posterior oracle inequality and posterior contraction. The main results are applied to ten examples ranging from nonparametric and high dimensional statistics in Section 5. In Section 6, we present further results on sparse linear regression. All technical proofs are gathered in Section 7-10.

We close this section by introducing some notation. Given an integer d, we use $[d]$ to denote the set $\{1, 2, ..., d\}$. For a set S, $|S|$ denotes its cardinality and $\mathbb{I}_S$ denotes the indicator function. For a vector $u = (u_i)$, $\|u\| = \sqrt{\sum_i u_i^2}$ denotes the ℓ_2 norm. For a matrix $A = (A_{ij}) \in \mathbb{R}^{n\times p}$, and a subset $T \subset [n] \times [p]$, A_T denotes the array $\{A_t\}_{t\in T}$. For any $I \subset [n]$ and $J \subset [p]$, we let $A_{I*} = A_{I\times[p]}$ and $A_{*J} = A_{[n]\times J}$. The Frobenius norm, ℓ_1 norm and ℓ_∞ norm are defined by $\|A\|_{\mathrm{F}} = \sqrt{\sum_{ij} A_{ij}^2}$, $\|A\|_1 = \sum_{ij} |A_{ij}|$ and $\|A\|_\infty = \max_{ij} |A_{ij}|$, respectively. When $A = A^T \in \mathbb{R}^{p\times p}$ is symmetric, the operator norm $\|A\|_{\mathrm{op}}$ is defined by its largest singular value and the matrix ℓ_1 norm $\|A\|_{\ell_1}$ is defined by the maximum row sum. The inner product is defined by $\langle u, v\rangle = \sum_i u_i v_i$ when applied to vectors and is defined by $\langle A, B\rangle = \sum_{ij} A_{ij}B_{ij}$ when applied to matrices. Given two numbers $a, b \in \mathbb{R}$, we use $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. The floor function $\lfloor a\rfloor$ is the largest integer no greater than a, and the ceiling function $\lceil a\rceil$ is the smallest integer no less than a. For two positive sequences $\{a_n\}, \{b_n\}$, $a_n \lesssim b_n$ means $a_n \le Cb_n$ for some constant $C > 0$ independent of n, and $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $b_n \lesssim a_n$. The symbols $\mathbb{P}$ and $\mathbb{E}$ denote generic probability and expectation operators whose distribution is determined from the context.

2 Structured linear models

Let us consider the following structured linear model

$$Y = \mathscr{X}_Z(Q) + W \in \mathbb{R}^N,$$

where $W \in \mathbb{R}^N$ is a noise vector and $\mathscr{X}_Z(\cdot)$ is a linear operator. The signal $\mathscr{X}_Z(Q)$ has two elements, the parameter Q and the structure/model Z that indexes the linear operator $\mathscr{X}_Z(\cdot)$. The structure Z is in some discrete space $\mathcal{Z}_\tau$, which is further indexed by $\tau \in \mathcal{T}$ for some finite set $\mathcal{T}$. We introduce a function $\ell(\mathcal{Z}_\tau)$ that determines the dimension of the parameter Q. In other words, $Q \in \mathbb{R}^{\ell(\mathcal{Z}_\tau)}$, and $\ell(\mathcal{Z}_\tau)$ is referred to as the effective dimension of the structured linear model. The complexity of the model is defined by the quantity

$$\ell(\mathcal{Z}_\tau) + \log|\mathcal{Z}_\tau|, \tag{1}$$

the sum of the effective dimension and the logarithmic cardinality of the structure space. As we are going to show later, (1) will be the posterior contraction rate that we target at. Moreover, in all the examples considered in the paper, (1) will be the minimax rate under the prediction loss. The only requirement we impose on the model is the linearity of the operator $\mathscr{X}_Z(\cdot)$. That is, given any $Z \in \mathcal{Z}_\tau$, we have

$$\mathscr{X}_Z(Q_1 + Q_2) = \mathscr{X}_Z(Q_1) + \mathscr{X}_Z(Q_2), \quad \text{for all } Q_1, Q_2 \in \mathbb{R}^{\ell(\mathcal{Z}_\tau)}. \tag{2}$$

Therefore, we can also view $\mathscr{X}_Z$ as a matrix in $\mathbb{R}^{N\times \ell(\mathcal{Z}_\tau)}$. From now on, whenever we apply a matrix operation with $\mathscr{X}_Z$, the operator $\mathscr{X}_Z$ is understood to be a matrix with slight abuse of notation.

The above framework of structured linear models includes many examples. In this paper, we consider the following six representative instances.

1. *Stochastic block model.* Consider $\mathscr{X}_Z(Q) \in [0,1]^{n\times n}$ to be the mean matrix of a random graph with specification $[\mathscr{X}_Z(Q)]_{ij} = Q_{z(i)z(j)}$. The object $z \in [k]^n$ is the labels of nodes. Moreover, it is easy to see that the parameter Q is of dimension k^2. Therefore, stochastic block model is a special case of our general framework in view of the relation $Z = z$, $\tau = k$, $\mathcal{T} = [n]$, $\mathcal{Z}_k = [k]^n$ and $\ell(\mathcal{Z}_k) = k^2$.

2. *Biclustering.* For a matrix $\mathscr{X}_Z(Q) \in \mathbb{R}^{n\times m}$, a biclustering model means that both rows and columns have clustering structures. That is, $[\mathscr{X}_Z(Q)]_{ij} = Q_{z_1(i)z_2(j)}$ for some $z_1 \in [k]^n$ and $z_2 \in [l]^m$. The parameter Q has dimension kl. Thus, biclustering model is a special case of our general framework by the relation $Z = (z_1, z_2)$, $\tau = (k, l)$, $\mathcal{T} = [n] \times [m]$, $\mathcal{Z}_{k,l} = [k]^n \times [l]^m$ and $\ell(\mathcal{Z}_{k,l}) = kl$.

3. *Sparse linear regression.* A p-dimensional sparse linear regression model refers to $X\beta$, where $\beta \in \mathbb{R}^p$ has a subset of nonzero entries and it can be represented by $\beta^T = (\beta_S^T, 0_{S^c}^T)$ for some $S \subset [p]$. In other words, $X\beta = X_{*S}\beta_S$. It can be represented in a general way by letting $Z = S$, $\tau = s$, $\mathcal{T} = [p]$, $\mathcal{Z}_s = \{S \subset [p] : |S| = s\}$, $\ell(\mathcal{Z}_s) = s$ and $Q = \beta_S$. Moreover, $\mathscr{X}_Z(Q) = X_{*S}\beta_S$.

4. *Linear regression with group sparsity.* It refers to the model XB with $B \in \mathbb{R}^{p\times m}$ being a matrix with nonzero rows in some subset $S \subset [p]$. It can be represented in a general form similarly as the sparse linear regression except that $\ell(\mathcal{Z}_s) = ms$.

5. *Multi-task learning.* Similar to the last example, multi-task learning is the collection of m regression problems. That is, we consider XB for some $B \in \mathbb{R}^{p\times m}$. The jth column of B is represented as $B_{*j} = Q_{*z(j)}$ for some $z \in [k]^m$ and $Q \in \mathbb{R}^{p\times k}$. Thus, it is a special case of our general framework by letting $Z = z$, $\tau = k$, $\mathcal{T} = [m]$, $\mathcal{Z}_k = [k]^m$ and $\ell(\mathcal{Z}_k) = pk$.

6. *Dictionary learning.* Consider the model $\mathscr{X}_Z(Q) = QZ \in \mathbb{R}^{n\times d}$ for some $Z \in \{-1, 0, 1\}^{p\times d}$ and $Q \in \mathbb{R}^{n\times p}$. Each column of Z is assumed to be sparse. Therefore, dictionary learning can be viewed as sparse regression without knowing the design. It can be written in a general form by letting $\tau = (p, s)$, $\mathcal{T} = \{(p, s) \in [n \wedge d] \times [n] : s \le p\}$, $\mathcal{Z}_{p,s} = \{Z \in \{-1, 0, 1\}^{p\times d} : \max_{j\in[d]} |\mathrm{supp}(Z_{*j})| \le s\}$ and $\ell(\mathcal{Z}_{p,s}) = np$.

3 The prior distribution

In this section, we introduce a prior distribution on the structured linear model. The prior distribution has a two-step sampling procedure. First, we are going to sample a structure

Z. Second, given Z, we sample the parameter Q. Let us first present the prior distribution on the parameter $Q \in \mathbb{R}^{\ell(\mathcal{Z}_\tau)}$. We propose the elliptical Laplace distribution with density function proportion to $\exp\left(-\lambda\|\mathscr{X}_Z(Q)\|\right)$. By direct calculation of the normalizing constant, the density function is

$$f_{\ell(\mathcal{Z}_\tau),\mathscr{X}_Z,\lambda}(Q) = \frac{\sqrt{\det(\mathscr{X}_Z^T\mathscr{X}_Z)}}{2}\left(\frac{\lambda}{\sqrt{\pi}}\right)^{\ell(\mathcal{Z}_\tau)}\frac{\Gamma(\ell(\mathcal{Z}_\tau)/2)}{\Gamma(\ell(\mathcal{Z}_\tau))}\exp\left(-\lambda\|\mathscr{X}_Z(Q)\|\right). \tag{3}$$

Recall that $\mathscr{X}_Z$ is understood as a matrix in $\mathbb{R}^{N\times\ell(\mathcal{Z}_\tau)}$ whenever a matrix operation is applied. The elliptical Laplace distribution belongs to the elliptical family [20] with scatter matrix proportional to $(\mathscr{X}_Z^T\mathscr{X}_Z)^{-1}$. Compared with an i.i.d. distribution on Q, the density function (3) involves an extra factor $\frac{\Gamma(\ell(\mathcal{Z}_\tau)/2)}{\Gamma(\ell(\mathcal{Z}_\tau))}$ in the normalizing constant. This factor needs to be corrected in the model selection step.

Let $\epsilon(\mathcal{Z}_\tau)$ be a function satisfying

$$\epsilon(\mathcal{Z}_\tau) \geq \ell(\mathcal{Z}_\tau) + \log|\mathcal{Z}_\tau|, \tag{4}$$

and then the sampling procedure of the prior distribution Π on $\mathscr{X}_Z(Q)$ is given by:

1. Sample $\tau \sim \pi$ from $\mathcal{T}$, where $\pi(\tau) \propto \frac{\Gamma(\ell(\mathcal{Z}_\tau))}{\Gamma(\ell(\mathcal{Z}_\tau)/2)}\exp\left(-D\epsilon(\mathcal{Z}_\tau)\right)$;
2. Conditioning on τ, sample Z uniformly from the set $\bar{\mathcal{Z}}_\tau = \{Z \in \mathcal{Z}_\tau : \det(\mathscr{X}_Z^T\mathscr{X}_Z) > 0\}$;
3. Conditioning on (τ, Z), sample $Q \sim f_{\ell(\mathcal{Z}_\tau),\mathscr{X}_Z,\lambda}$.

Step 1 weighs the structure index τ by the function $\epsilon(\mathcal{Z}_\tau)$ that satisfies (4). For all the examples considered in the paper, $\epsilon(\mathcal{Z}_\tau)$ is chosen to be at the same order of the model complexity (1). The quantity $\frac{\Gamma(\ell(\mathcal{Z}_\tau))}{\Gamma(\ell(\mathcal{Z}_\tau)/2)}$ is called the correction factor that is imposed to compensate the influence of $\frac{\Gamma(\ell(\mathcal{Z}_\tau)/2)}{\Gamma(\ell(\mathcal{Z}_\tau))}$ in the elliptical Laplace distribution. Without the correction factor, $\exp\left(-D\epsilon(\mathcal{Z}_\tau)\right)$ is the complexity prior used by [14, 15] in Gaussian sequence model and sparse linear regression. Since the support $\mathcal{T}$ is a finite set, π is always a valid probability mass function. Step 2 samples a structure Z uniformly in $\bar{\mathcal{Z}}_\tau$. It is sufficient to consider such Z that $\det(\mathscr{X}_Z^T\mathscr{X}_Z) > 0$ for all the examples considered in this paper. Such restriction leads to a proper density function (3) and thus Step 3 is well defined.

After defining the prior, we also need to specify the likelihood function. The six examples in Section 2 have different distributions. For example, stochastic block model usually assumes a Bernoulli random graph, while sparse linear regression often works with general sub-Gaussian noise distributions. To pursue a unified approach, we propose to use the Gaussian likelihood $Y|(Z,Q) \sim N(\mathscr{X}_Z(Q), I_N)$. Then, the posterior distribution is

$$\begin{aligned}
&\Pi\left(\mathscr{X}_Z(Q) \in U|Y\right)\\
=\ &\frac{\sum_{\tau\in\mathcal{T}} e^{-D\epsilon(\mathcal{Z}_\tau)}\sum_{Z\in\bar{\mathcal{Z}}_\tau}\frac{\sqrt{\det(\mathscr{X}_Z^T\mathscr{X}_Z)}}{|\bar{\mathcal{Z}}_\tau|}\left(\frac{\lambda}{\sqrt{\pi}}\right)^{\ell(\mathcal{Z}_\tau)}\int_{\mathscr{X}_Z(Q)\in U} e^{-\frac{1}{2}\|Y-\mathscr{X}_Z(Q)\|^2-\lambda\|\mathscr{X}_Z(Q)\|}dQ}{\sum_{\tau\in\mathcal{T}} e^{-D\epsilon(\mathcal{Z}_\tau)}\sum_{Z\in\bar{\mathcal{Z}}_\tau}\frac{\sqrt{\det(\mathscr{X}_Z^T\mathscr{X}_Z)}}{|\bar{\mathcal{Z}}_\tau|}\left(\frac{\lambda}{\sqrt{\pi}}\right)^{\ell(\mathcal{Z}_\tau)}\int e^{-\frac{1}{2}\|Y-\mathscr{X}_Z(Q)\|^2-\lambda\|\mathscr{X}_Z(Q)\|}dQ}.
\end{aligned}$$

Note that in the above formula of posterior distribution, the factor $\frac{\Gamma(\ell(\mathcal{Z}_\tau)/2)}{\Gamma(\ell(\mathcal{Z}_\tau))}$ in the Laplace normalizing constant has been cancelled out by the correction factor $\frac{\Gamma(\ell(\mathcal{Z}_\tau))}{\Gamma(\ell(\mathcal{Z}_\tau)/2)}$ in the model selection prior.

4 Main results

In this section, we analyze the posterior distribution for the general structured linear model. Though the prior specifies a model $\mathscr{X}_Z(Q)$, we do not need to assume that the data is generated from the same model. Instead, we allow the data to be generated by an arbitrary signal with sub-Gaussian noise. That is,

$$Y = \theta^* + W,$$

where $W = Y - \theta^*$ is the noise vector with a sub-Gaussian tail satisfying

$$\mathbb{P}\left(|\langle W, K\rangle| > t\right) \leq e^{-\rho t^2/2} \text{ for all } \|K\| = 1. \tag{5}$$

The sub-Gaussianity number $\rho > 0$ is assumed to be a constant throughout the paper. We also assume a mild assumption on the function $\epsilon(\mathcal{Z}_\tau)$. That is,

$$|\{\tau \in \mathcal{T} : t - 1 < \epsilon(\mathcal{Z}_\tau) \leq t\}| \leq t \text{ for all } t \in \mathbb{N}. \tag{6}$$

Recall that λ and D are parameters of the prior distribution Π, and the main result of the paper is stated in the following theorem.

Theorem 4.1. *Assume (4), (5) and (6). Given any $\theta^* \in \mathbb{R}^N$, any $\tau^* \in \mathcal{T}$, any $Z^* \in \bar{\mathcal{Z}}_{\tau^*}$, any $Q^* \in \mathbb{R}^{\ell(\mathcal{Z}_{\tau^*})}$, any constants $\lambda, \rho > 0$ and any sufficiently small constant $\delta \in (0,1)$, there exists some constant $D_{\lambda,\delta,\rho} > 0$ only depending on λ, δ, ρ, such that*

$$\begin{aligned} &\mathbb{E}\Pi\left(\epsilon(\mathcal{Z}_\tau) > (1+\delta_1)\epsilon(\mathcal{Z}_{\tau^*}) + \delta_1\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2 \Big| Y\right) \\ \leq\ &\exp\left(-C'\left(\epsilon(\mathcal{Z}_{\tau^*}) + \|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2\right)\right) \end{aligned} \tag{7}$$

and

$$\begin{aligned} &\mathbb{E}\Pi\left(\|\mathscr{X}_Z(Q) - \theta^*\|^2 > (1+\delta_2)\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2 + M\epsilon(\mathcal{Z}_{\tau^*}) \Big| Y\right) \\ \leq\ &\exp\left(-C''\left(\epsilon(\mathcal{Z}_{\tau^*}) + \|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2\right)\right) \end{aligned} \tag{8}$$

for any constant $D > D_{\lambda,\delta,\rho}$ with $\delta_1 = \delta$, $\delta_2 = 8\sqrt{\delta/\rho}$ and some constants M, C', C'' only depending on λ, δ, ρ, D.

Remark 4.1. *The results of Theorem 4.1 hold for all $\epsilon(\mathcal{Z}_\tau)$ satisfying (4). By choosing $\epsilon(\mathcal{Z}_\tau)$ at the same order of (1), we obtain the rate $\ell(\mathcal{Z}_{\tau^*}) + \log|\mathcal{Z}_{\tau^*}|$ under the posterior distribution. From now on, we refer to both (1) and $\epsilon(\mathcal{Z}_\tau)$ as the complexity function.*

Remark 4.2. *By scrutinizing the proof of Theorem 4.1, the assumption (6) can be weakened. In fact, we only require* $|\{\tau \in \mathcal{T} : t-1 < \epsilon(\mathcal{Z}_\tau) \le t\}| \le at^b$ *for arbitrary constants* $a, b > 0$ *for the result of Theorem 4.1 to hold. However, the current form (6) is much simpler and is sufficient for all the examples considered in the paper.*

Theorem 4.1 contains two results of an oracle type, where $\mathscr{X}_{Z^*}(Q^*)$ is understood to be the oracle model that best approximates the true signal θ^*. The first result (7) shows that the model complexity selected by the posterior distribution is not greater than the sum of the complexity of the oracle and a model misspecification term quantified by $\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2$. The second result (8) is a posterior oracle inequality for the squared error loss $\|\mathscr{X}_Z(Q) - \theta^*\|^2$. Compared with that of the oracle $\mathscr{X}_{Z^*}(Q^*)$, the squared error loss of $\mathscr{X}_Z(Q)$ has an extra term proportional to $\epsilon(\mathcal{Z}_{\tau^*})$. It is worth noting that the constant $(1+\delta_2)$ in (8) can be arbitrarily close to 1, as long as D is chosen sufficiently large. Since our procedure involves a model selection step, an oracle inequality with constant exactly 1 is impossible, which is implied by a counter-example in [45] for sparse linear regression. Besides, we do not impose any assumption on the operator $\mathscr{X}_Z(\cdot)$ except its linearity (2). In the regression model, this means the results are assumption-free for the design matrix.

When the model is well specified in the sense that $\theta^* = \mathscr{X}_{Z^*}(Q^*)$, Theorem 4.1 reduces to the following results on posterior contraction.

Corollary 4.1. *Assume (4), (5) and (6). For any* $\theta^* = \mathscr{X}_{Z^*}(Q^*)$ *with any* $Z^* \in \bar{\mathcal{Z}}_{\tau^*}$, *any* $\tau^* \in \mathcal{T}$, *any* $Q^* \in \mathbb{R}^{\ell(\mathcal{Z}_{\tau^*})}$, *any constants* $\lambda, \rho > 0$ *and any sufficiently small constant* $\delta \in (0,1)$, *there exists some constant* $D_{\lambda,\delta,\rho} > 0$ *only depending on* λ, δ, ρ, *such that*

$$\mathbb{E}\Pi\left(\epsilon(\mathcal{Z}_\tau) > (1+\delta)\epsilon(\mathcal{Z}_{\tau^*})\middle| Y\right) \le \exp\left(-C'\epsilon(\mathcal{Z}_{\tau^*})\right)$$

and

$$\mathbb{E}\Pi\left(\|\mathscr{X}_Z(Q) - \theta^*\|^2 > M\epsilon(\mathcal{Z}_{\tau^*})\middle| Y\right) \le \exp\left(-C''\epsilon(\mathcal{Z}_{\tau^*})\right)$$

for any constant $D > D_{\lambda,\delta,\rho}$ *with some constants* M, C', C'' *only depending on* λ, δ, ρ, D.

Therefore, the posterior contraction rate under the squared error loss is $\epsilon(\mathcal{Z}_{\tau^*})$, which can be taken at the order of $\ell(\mathcal{Z}_{\tau^*}) + \log|\mathcal{Z}_{\tau^*}|$. As we are going to show in the next section, the rate is minimax optimal for all the examples considered in the paper.

5 Applications

5.1 Stochastic block model

Stochastic block model was proposed by [28] to model random graphs with a community structure. Given a symmetric adjacency matrix $A = A^T \in \{0,1\}^{n\times n}$ that codes an undirected network with no self loops in the sense that $A_{ii} = 0$ for all $i \in [n]$, stochastic block model assumes $\{A_{ij}\}_{i>j}$ are independent Bernoulli random variables with mean $\theta_{ij} = Q_{z(i)z(j)} \in [0,1]$ with some matrix $Q \in [0,1]^{k\times k}$ and some label vector $z \in [k]^n$. In other words, the

probability that there is an edge between the ith and the jth nodes only depends on their community labels $z(i)$ and $z(j)$. Recently, the problem of estimating the success matrix θ receives much attention. The minimax rate of estimating θ under the Frobenius norm was established by [23]. However, the upper bound in [23] was achieved by a procedure assuming the knowledge of the true number of community k^*, and is not adaptive. The Bayes framework proposed in this paper provides a natural solution to adaptive estimation for stochastic block model.

Let us write the stochastic block model in a general from as $\theta_{ij} = [\mathscr{X}_Z(Q)]_{ij} = Q_{z(i)z(j)}$ for all $i \neq j$. We do not need to model the diagonal entries because $A_{ii} = 0$ for all $i \in [n]$ as convention. Then, $Z = z$, $\tau = k$, $\mathcal{T} = [n]$ and $\mathcal{Z}_k = [k]^n$. Though the true parameter Q^* is symmetric, we do not impose symmetry for the prior distribution. Hence, $\ell(\mathcal{Z}_k) = k^2$ and (4) is satisfied with $\epsilon(\mathcal{Z}_k) = k^2 + n\log k$. The general prior distribution Π can be specialized to this case as

1. Sample $k \sim \pi$ from $[n]$, where $\pi(k) \propto \frac{\Gamma(k^2)}{\Gamma(k^2/2)} \exp\left(-D(k^2 + n\log k)\right)$;
2. Conditioning on k, sample z uniformly from $[k]^n$;
3. Conditioning on (k, z), sample $Q \sim f_{k,z,\lambda}$, where $f_{k,z,\lambda}(Q) \propto e^{-\lambda\sqrt{\sum_{i\neq j} Q^2_{z(i)z(j)}}}$;
4. Set $\theta_{ij} = Q_{z(i)z(j)}$ for all $i \neq j$ and $\theta_{ii} = 0$ for all $i \in [n]$.

Note that in Step 2, we use $\mathcal{Z}_k = [k]^n$ instead of $\bar{\mathcal{Z}}_k$. This is because $(Q_1)_{z(i)z(j)} = (Q_2)_{z(i)z(j)}$ for all $i \neq j$ implies $Q_1 = Q_2$, and thus $\mathcal{Z}_k = \bar{\mathcal{Z}}_k = [k]^n$. To better understand the density function $f_{k,z,\lambda}$, consider the case where n/k is an integer and the community sizes $|\{i \in [n] : z(i) = u\}| = n/k$ are equal for all $u \in [k]$. Then $f_{k,z,\lambda}(Q) \propto e^{-\frac{n\lambda}{k}\|Q\|_{\mathrm{F}}}$, if we also include the diagonal entries. The exponent of the general form of $f_{k,z,\lambda}$ involves a weighted norm of Q depending on the community sizes.

To study the posterior distribution, let us assume that the adjacency matrix is generated by the true mean $\theta^*_{ij} = Q^*_{z^*(i)z^*(j)} = Q^*_{z^*(j)z^*(i)} \in [0, 1]$ for $i \neq j$ and $\theta^*_{ii} = 0$ for all $i \in [n]$. Assume $z^* \in [k^*]^n$ for some $k^* \in [n]$. It is easy to see that the noise $W = A - \theta^*$ satisfies (5) for some constant $\rho > 0$ by Hoeffding's inequality. Moreover, the complexity function $\epsilon(\mathcal{Z}_\tau) = k^2 + n\log k$ satisfies (6). Hence, Corollary 4.1 can be specialized for the stochastic block model.

Corollary 5.1. *For any θ^* and k^* specified above, any constant $\lambda > 0$ and any sufficiently small constant $\delta \in (0, 1)$, there exists some constant $D_{\lambda,\delta} > 0$ only depending on λ, δ such that*

$$\mathbb{E}\Pi\left(k^2 + n\log k > (1+\delta)\left((k^*)^2 + n\log k^*\right)\Big|A\right) \leq \exp\left(-C'((k^*)^2 + n\log k^*)\right)$$

and

$$\mathbb{E}\Pi\left(\|\theta - \theta^*\|_{\mathrm{F}}^2 > M((k^*)^2 + n\log k^*)\Big|A\right) \leq \exp\left(-C''((k^*)^2 + n\log k^*)\right)$$

for any constant $D > D_{\lambda,\delta}$ with some constants M, C', C'' only depending on λ, δ, D.

To the best of our knowledge, this is the first Bayes estimator for stochastic block model with theoretical justification. The posterior contraction rate is $(k^*)^2 + n \log k^*$. According to [23], this is the minimax rate of the problem. When $k^* \le \sqrt{n \log n}$, the rate is dominated by $n \log k^*$, which grows only logarithmically as k^* grows. When $k < \sqrt{n \log n}$, the rate is dominated by $(k^*)^2$, corresponding to the number of parameters. Since posterior contraction implies the existence of a point estimator with the same rate [25], the posterior mean is automatically a rate-optimal adaptive estimator.

5.2 Biclustering

The biclustering model, originated in [26], can be viewed as an asymmetric extension of the stochastic block model. The data matrix $Y \in \mathbb{R}^{n \times m}$ is assumed to be generated by a signal matrix $\theta = \{\theta_{ij}\}$ with form $\theta_{ij} = Q_{z_1(i) z_2(j)}$ for some label vectors $z_1 \in [k]^n$ and $z_2 \in [l]^m$. In other words, the rows of θ have k clusters and the columns of θ have l clusters. The values of $\{\theta_{ij}\}$ that belong to the same row-cluster and the same column-cluster are constant. The goal is to recover the true signal matrix θ^* from the observation Y.

To put it in our general form, observe that $Z = (z_1, z_2)$, $\tau = (k, l)$, $\mathcal{T} = [n] \times [m]$, $\mathcal{Z}_{k,l} = [k]^n \times [l]^m$ and $\ell(\mathcal{Z}_{n,l}) = kl$. Moreover, the complexity function is $\epsilon(\mathcal{Z}_{k,l}) = kl + k \log n + l \log m$, which satisfies (4) and (6). The general prior Π can be specialized to this case as

1. Sample $(k, l) \sim \pi$ from $[n] \times [m]$, where $\pi(k, l) \propto \frac{\Gamma(kl)}{\Gamma(kl/2)} \exp\left(-D(kl + n \log k + m \log l)\right)$;
2. Conditioning on (k, l), sample (z_1, z_2) uniformly from $[k]^n \times [l]^m$;
3. Conditioning on (k, l, z_1, z_2), sample $Q \sim f_{k,l,z_1,z_2,\lambda}$ with $f_{k,l,z_1,z_2,\lambda}(Q) \propto e^{-\lambda \sqrt{\sum_{ij} Q^2_{z_1(i) z_2(j)}}}$;
4. Set $\theta_{ij} = Q_{z_1(i) z_2(j)}$ for all (i, j).

In Step 2, we use $\mathcal{Z}_{k,l}$ because $\mathcal{Z}_{k,l} = \bar{\mathcal{Z}}_{k,l}$ for the same reason as we have argued for the stochastic block model. To analyze the posterior distribution, consider data $Y = \theta^* + W$, where the signal θ^* admits a biclustering structure such that $\theta^*_{ij} = Q^*_{z_1^*(i) z_2^*(j)}$ for $Q^* \in \mathbb{R}^{k^* \times l^*}$ and $(z_1^*, z_2^*) \in [k^*]^n \times [l^*]^m$, and the noise W is assume to satisfy (5).

Corollary 5.2. *For any θ^* and (k^*, l^*) specified above, any constants $\lambda, \rho > 0$ and any sufficiently small constant $\delta \in (0, 1)$, there exists some constant $D_{\lambda,\delta,\rho} > 0$ only depending on λ, δ, ρ such that*

$$\begin{aligned} & \mathbb{E}\Pi\left(kl + n \log k + m \log l > (1 + \delta)\left(k^* l^* + n \log k^* + m \log l^*\right) \middle| Y\right) \\ \le\ & \exp\left(-C'(k^* l^* + n \log k^* + m \log l^*)\right) \end{aligned}$$

and

$$\mathbb{E}\Pi\left(\|\theta - \theta^*\|_{\mathrm{F}}^2 > M(k^* l^* + n \log k^* + m \log l^*) \middle| Y\right) \le \exp\left(-C''(k^* l^* + n \log k^* + m \log l^*)\right)$$

for any constant $D > D_{\lambda,\delta,\rho}$ with some constants M, C', C'' only depending on λ, δ, ρ, D.

The posterior contraction rate for recovering a signal matrix with a biclustering structure is $k^*l^* + n\log k^* + m\log l^*$, which is minimax optimal according to [23]. To the best of our knowledge, this is the first adaptive estimation result for biclustering with optimal rate.

5.3 Sparse linear regression

Consider a regression problem with fixed design $X\beta$, where $X \in \mathbb{R}^{n\times p}$ and $\beta \in \mathbb{R}^p$. The regression coefficient is assumed to be sparse so that $\beta^T = (\beta_S^T, 0_{S^c}^T)$ for some $S \subset [p]$. Recovering the mean vector $X\beta$ and the regression vector β with a sparse prior has been considered in [15]. However, the results of [15] imposed strong assumptions that are commonly used for the Lasso estimator [7]. In this section, we show that the general prior distribution that we propose in Section 3 leads to optimal posterior contraction rates with minimal assumptions.

First, we note that the sparse linear regression model is a special case of the general structured linear model by letting $Z = S$, $\tau = s$, $\mathcal{T} = [p]$, $\mathcal{Z}_s = \{S \subset [p] : |S| = s\}$, $\ell(\mathcal{Z}_s) = s$ and $Q = \beta_S$. Then, we have the representation $\mathscr{X}_Z(Q) = X_{*S}\beta_S = X\beta$. Since $\log|\mathcal{Z}_s| = \log\binom{p}{s} \le s\log\frac{ep}{s}$, the complexity function $\epsilon(\mathcal{Z}_s) = 2s\log\frac{ep}{s}$ satisfies the condition (4). It is also easy to check that $\epsilon(\mathcal{Z}_\tau)$ satisfies (6). We specialize the general prior Π in Section 3 as follows.

1. Sample $s \sim \pi$ from $[p]$, where $\pi(s) \propto \frac{\Gamma(s)}{\Gamma(s/2)}\exp\left(-2Ds\log\frac{ep}{s}\right)$;
2. Conditioning on s, sample S uniformly from $\{S \subset [p] : |S| = s, \det(X_{*S}^T X_{*S}) > 0\}$;
3. Conditioning on (s, S), sample $\beta_S \sim f_{s,S,\lambda}$ with $f_{s,S,\lambda}(\beta_S) \propto e^{-\lambda\|X_{*S}\beta_S\|}$ and set $\beta_{S^c} = 0$.

Note that in Step 1, we use $\epsilon(\mathcal{Z}_s) = 2s\log\frac{ep}{s}$ instead of the exact form of $\ell(\mathcal{Z}_\tau) + \log|\mathcal{Z}_\tau|$ in the exponent for simplicity. In Step 2, we sample S from the set $\bar{\mathcal{Z}}_s = \{S \subset [p] : |S| = s, \det(X_{*S}^T X_{*S}) > 0\}$. In this way, the density $f_{s,S,\lambda}$ in Step 3 is not degenerate. Since $X_{*S} \in \mathbb{R}^{n\times s}$, when $s > n$, we must have $\mathcal{Z}_s = \varnothing$. Hence, we may also replace π in Step 1 by its renormalized version supported on $[n]$. Furthermore, note that the exponent on the density of β_S is $-\lambda\|X_{*S}\beta_S\|$, compared to $-\lambda\|\beta_S\|_1$ in [15]. We let the prior depend on the design matrix X to obtain assumption-free optimal posterior prediction rate. The idea of design-dependent prior was also employed by [40] in an empirical pseudo-Bayes framework. Moreover, $e^{-\lambda\|X_{*S}\beta_S\|}$ implies an exponential tail, which is capable of modeling a large regression coefficient.

The prior distribution involves a correction factor $\frac{\Gamma(s)}{\Gamma(s/2)}$ in the model selection step to compensate the normalizing constant of the elliptical Laplace distribution. Without this factor, $\exp\left(-2Ds\log\frac{ep}{s}\right)$ is the common prior distribution on the model dimension used in [45, 14, 22, 15, 40]. Since $\exp\left(-2Ds\log\frac{ep}{s}\right)$ is a decreasing function of s, it gives less weights for more complex models. However, with the correction factor, this is not true because $\pi(s) \propto \frac{\Gamma(s)}{\Gamma(s/2)}\exp\left(-2Ds\log\frac{ep}{s}\right)$ is not necessarily a decreasing function of s. For a large $D > 0$, we have $\pi(\sqrt{p}) < \pi(p)$, which leads to a counter-intuitive prior modeling strategy.

Let us proceed to specify the truth. That is, $Y = X\beta^* + W$ for some β^* with support S^* and sparsity $|S^*| = s^*$. The noise vector is assumed to be sub-Gaussian in the sense of

(5). Without loss of generality, we may assume $S^* \in \bar{\mathcal{Z}}_{s^*}$. This is because if X_{*S^*} is collinear in the sense that $\det(X_{*S^*}^T X_{*S^*}) = 0$, there always exists a β_1 with support S_1 and sparsity $s_1 = |S_1|$ such that $X\beta^* = X\beta_1$ and $\det(X_{*S_1}^T X_{*S_1}) > 0$. We may simply redefine (s^*, S^*) by (s_1, S_1).

Corollary 5.3. *For any β^*, $S^* \in \bar{\mathcal{Z}}_{s^*}$ and s^* specified above, any constants $\lambda, \rho > 0$ and any sufficiently small constant $\delta \in (0,1)$, there exists some constant $D_{\lambda,\delta,\rho} > 0$ only depending on λ, δ, ρ such that*

$$\mathbb{E}\Pi\left(s > (1+\delta)s^* \middle| Y\right) \le \exp\left(-C' s^* \log \frac{ep}{s^*}\right) \tag{9}$$

and

$$\mathbb{E}\Pi\left(\|X\beta - X\beta^*\|^2 > M s^* \log \frac{ep}{s^*} \middle| Y\right) \le \exp\left(-C'' s^* \log \frac{ep}{s^*}\right) \tag{10}$$

for any constant $D > D_{\lambda,\delta,\rho}$ with some constants M, C', C'' only depending on λ, δ, ρ, D.

The result (9) is implied by (7) that $s \log \frac{ep}{s} \le (1+\delta_1) s^* \log \frac{ep}{s^*}$ under the posterior distribution. It improves the corresponding bounds in [14, 15] at a constant level. The result (10) achieves the minimax optimal prediction rate with no assumption on the design matrix X, which is comparable to the frequentist result in [8]. Slight improvement of (10) will be discussed in Section 5.10.

Besides optimal prediction rate, we are ready to obtain optimal estimation rates given (10) and (9). Define

$$\kappa_2 = \min_{\{b \neq 0 : \|b\|_0 \le (2+\delta)s^*\}} \frac{\|Xb\|}{\sqrt{n}\|b\|} \quad \text{and} \quad \kappa_1 = \min_{b \neq 0 : \|b\|_0 \le (2+\delta)s^*} \frac{\sqrt{s^*}\|Xb\|}{\sqrt{n}\|b\|_1}. \tag{11}$$

Note that κ_2 is the restricted eigenvalue constant [12, 7] and κ_1 is the compatibility constant [10].

Corollary 5.4. *Under the setting of Corollary 5.5, we have*

$$\mathbb{E}\left(\|\beta - \beta^*\|^2 > M \frac{s^* \log \frac{ep}{s^*}}{n\kappa_2^2} \middle| Y\right) \le 2\exp\left(-(C' + C'') s^* \log \frac{ep}{s^*}\right)$$

and

$$\mathbb{E}\left(\|\beta - \beta^*\|_1^2 > M \frac{(s^*)^2 \log \frac{ep}{s^*}}{n\kappa_1^2} \middle| Y\right) \le 2\exp\left(-(C' + C'') s^* \log \frac{ep}{s^*}\right)$$

for the same constants M, C', C'' in Corollary 5.5.

Compared with the minimax rates [18, 54], Corollary 5.4 obtains optimal estimation rates for both ℓ_2 and ℓ_1 loss functions. Moreover, the dependence on the quantities κ_2 and κ_1 are optimal [44], compared with the Lasso estimator and the spike and slab prior [15]. When $\kappa \asymp \kappa_1 \asymp \kappa_2$, the rates of Lasso depend on κ through κ^4 for both the loss $\|\cdot\|^2$ [7] and the loss $\|\cdot\|_1^2$ [52], and the rates of the spike and slab prior depend on κ through κ^6 for the loss $\|\cdot\|^2$ and κ^8 for the loss $\|\cdot\|_1^2$ [15], while we obtain the optimal dependence κ^2 in Corollary 5.4.

The results on ℓ_∞ convergence and model selection for sparse linear regression are not implied by the general theory. We are going to treat it separately in Section 6.

5.4 Linear regression with group sparsity

Let us consider a multiple regression set up XB for $X \in \mathbb{R}^{n\times p}$ and $B \in \mathbb{R}^{p\times m}$. The matrix B collects regression coefficients from m regression problems. We assume the m regression coefficients share the same support. That is, there is some $S \subset [p]$ such that $B_{S^c *} = 0$. In other words, S is the nonzero rows of B. The concept of group sparsity was proposed by [3, 58], and frequentist statistical properties were analyzed by [34].

To apply a Bayes procedure, let us write the problem in a general form by $Z = S$, $\tau = s$, $\mathcal{T} = [p]$, $\mathcal{Z} = \{S \subset [p] : |S| = s\}$, $\ell(\mathcal{Z}_s) = ms$ and $Q = B_{S*}$. Then, we have the representation $\mathscr{X}_Z(Q) = X_{*S}B_{S*} = XB$. The choice $\epsilon(\mathcal{Z}_s) = s\left(m + \log\frac{ep}{s}\right)$ satisfies the conditions (4) and (6). The prior distribution Π is similar to that used in Section 5.3.

1. Sample $s \sim \pi$ from $[p]$, where $\pi(s) \propto \frac{\Gamma(s)}{\Gamma(s/2)}\exp\left(-Ds\left(m + \log\frac{ep}{s}\right)\right)$;
2. Conditioning on s, sample S uniformly from $\{S \subset [p] : |S| = s, \det(X_{*S}^T X_{*S}) > 0\}$;
3. Conditioning on (s, S), sample $B_{S*} \sim f_{s,S,\lambda}$ with $f_{s,S,\lambda}(B_{S*}) \propto e^{-\lambda\|X_{*S}B_{S*}\|_{\mathrm{F}}}$ and set $B_{S^c *} = 0$.

Note that we also use $\bar{\mathcal{Z}}_s$ in Step 3 as what we have done for sparse linear regression. Assume the data is generated by $Y = XB^* + W$ for some matrix B^* with support S^* and sparsity s^*. Again, without loss of generality, we assume $S^* \in \bar{\mathcal{Z}}_{s^*}$. The noise matrix W is assumed to be the sub-Gaussian in the sense of (5).

Corollary 5.5. *For any B^*, $S^* \in \bar{\mathcal{Z}}_{s^*}$ and s^* specified above, any constants $\lambda, \rho > 0$ and any sufficiently small constant $\delta \in (0, 1)$, there exists some constant $D_{\lambda,\delta,\rho} > 0$ only depending on λ, δ, ρ such that*

$$\mathbb{E}\Pi\left(s > (1+\delta)s^* \middle| Y\right) \le \exp\left(-C's^*\left(m + \log\frac{ep}{s^*}\right)\right)$$

and

$$\mathbb{E}\Pi\left(\|XB - XB^*\|_{\mathrm{F}}^2 > Ms^*\left(m + \log\frac{ep}{s^*}\right)\middle| Y\right) \le \exp\left(-C''s^*\left(m + \log\frac{ep}{s^*}\right)\right)$$

for any constant $D > D_{\lambda,\delta,\rho}$ with some constants M, C', C'' only depending on λ, δ, ρ, D.

The posterior contraction rate for the prediction loss is $s^*\left(m + \log\frac{ep}{s^*}\right)$, which is minimax optimal according to [34, 39]. Posterior contraction for various estimation loss functions can also be derived in a similar way as in Section 5.3, and we omit the details.

5.5 Multi-task learning

Multi-task learning is another name for multiple linear regression in the form of XB with $X \in \mathbb{R}^{n\times p}$ and $B \in \mathbb{R}^{p\times m}$. As opposed to m independent linear regression problems, a typical multi-task learning setting assumes some dependent structure among the columns of the coefficient matrix B. The group sparsity assumption considered in Section 5.4 is an example where the columns of B share the same support. In this section, we assume a

clustering structure among the columns of B. That is, $B_{*j} = Q_{*z(j)}$ for some $z \in [k]^m$ and $Q \in \mathbb{R}^{p\times k}$. In other words, the m regression coefficient vectors are allowed to choose from k possibilities. When the design X is an identity matrix, it reduces to an ordinary clustering problem.

Let us write the multi-task learning problem in the general form. This can be done by letting $Z = z$, $\tau = k$, $\mathcal{T} = [m]$, $\mathcal{Z}_k = [k]^m$ and $\ell(\mathcal{Z}_k) = pk$. Moreover, we have the representation $[\mathscr{X}_Z(Q)]_{*j} = XQ_{*z(j)}$. The complexity function $\epsilon(\mathcal{Z}_\tau) = pk + m\log k$ satisfies the conditions (4) and (6). The general prior distribution Π can be specialized to this case. Consider a full rank design matrix that $\det(X^TX) > 0$.

1. Sample $k \sim \pi$ from $[p]$, where $\pi(k) \propto \frac{\Gamma(pk)}{\Gamma(pk/2)} \exp\left(-D(pk + m\log k)\right)$;
2. Conditioning on k, sample z uniformly from $[k]^m$;
3. Conditioning on (k, z), sample $Q \sim f_{k,z,\lambda}$ with $f_{k,z,\lambda}(Q) \propto e^{-\lambda\sqrt{\sum_j \|XQ_{z(j)*}\|^2}}$;
4. Set $B_{*j} = Q_{*z(j)}$ for all $j \in [m]$.

Note that in Step 2, we use $\mathcal{Z}_k = [k]^m$ because $\bar{\mathcal{Z}}_k = \mathcal{Z}_k$, which is due to $\det(X^TX) > 0$. The full rankness of the design matrix implicitly implies $p \le n$. In fact, the assumption $\det(X^TX) > 0$ is without loss of generality, because whenever $\det(X^TX) = 0$, one can simply use a subset of the variables that are linearly independent without affecting the prediction error.

To state the result of posterior contraction, let us assume that the data is generated as $Y = XB^* + W$ for some matrix B^* satisfying $B^*_{*j} = Q^*_{*z^*(j)}$ with some Q^* and $z^* \in [k^*]^m$. The noise matrix is assumed to satisfy (5).

Corollary 5.6. *For any B^* and k^* specified above, any constants $\lambda, \rho > 0$ and any sufficiently small constant $\delta \in (0, 1)$, there exists some constant $D_{\lambda,\delta,\rho} > 0$ only depending on λ, δ, ρ such that*

$$\mathbb{E}\Pi\left(pk + m\log k > (1+\delta)(pk^* + m\log k^*)\middle| Y\right) \le \exp\left(-C'(pk^* + m\log k^*)\right)$$

and

$$\mathbb{E}\Pi\left(\|XB - XB^*\|_{\mathrm{F}}^2 > M(pk^* + m\log k^*)\middle| Y\right) \le \exp\left(-C''(pk^* + m\log k^*)\right)$$

for any constant $D > D_{\lambda,\delta,\rho}$ with some constants M, C', C'' only depending on λ, δ, ρ, D.

The posterior contraction rate for multi-task learning is $pk^* + m\log k^*$, which is smaller than the rate pm for m independent linear regressions. When $m\log k^* \le pk^*$, the rate becomes $pk^* + m\log k^* \asymp pk^*$. In this case, the procedure performs as well as when the clustering structure z^* is known. According to [38], the rate $pk^* + m\log k^*$ is minimax optimal.

5.6 Dictionary learning

Dictionary learning can be viewed as a linear regression problem without knowing the design matrix. Mathematically, the signal matrix $\theta \in \mathbb{R}^{n\times d}$ can be represented as $\theta = QZ$ for some $Q \in \mathbb{R}^{n\times p}$ and $Z \in \mathbb{R}^{p\times d}$. Both the dictionary Q and the coefficient matrix Z are unknown. A common assumption is that each column of Z is sparse, and the goal is to learn the latent sparse representation of the signal. Thus, the problem is also referred to as sparse coding [43]. Recently, the minimax rate of dictionary learning has been established by [38] for estimating the true signal matrix θ^*. In this section, we provide a Bayes solution to the adaptive estimation problem of dictionary learning. Following [1], we consider a discrete version of the problem. Namely, $Z \in \{-1, 0, 1\}^{p\times d}$. Then, the problem can be represented in a general form by letting $\tau = (p, s)$, $\mathcal{T} = \{(p, s) \in [n \wedge d] \times [n] : s \le p\}$, $\mathcal{Z}_{p,s} = \{Z \in \{-1, 0, 1\}^{p\times d} : \max_{j\in[d]} |\mathrm{supp}(Z_{*j})| \le s\}$ and $\ell(\mathcal{Z}_{p,s}) = np$. Moreover, we have the representation $\mathscr{X}_Z(Q) = QZ$. The complexity function is $\ell(\mathcal{Z}_{p,s}) + \log|\mathcal{Z}_{p,s}| = np + d\left(\log\binom{p}{s} + 3\log s\right)$. With $\epsilon(\mathcal{Z}_{p,s}) = 3\left(np + ds\log\frac{ep}{s}\right)$, (4) and (6) are satisfied. The general prior distribution Π can be specialized into the following sampling procedures.

1. Sample $(p, s) \sim \pi$ from $\mathcal{T}$ with $\pi(p, s) \propto \frac{\Gamma(np)}{\Gamma(np/2)} \exp\left(-3D\left(np + ds\log\frac{ep}{s}\right)\right)$;
2. Given (p, s), sample Z uniformly from $\bar{\mathcal{Z}}_{p,s} = \left\{Z \in \mathcal{Z}_{p,s} : \det(ZZ^T) > 0\right\}$;
3. Given (p, s, Z), sample $Q \sim f_{p,s,Z,\lambda}$ with $f_{p,s,Z,\lambda}(Q) \propto e^{-\lambda\|QZ\|}$;
4. Set $\theta = QZ$.

Note that we have used $\epsilon(\mathcal{Z}_{p,s}) = 3\left(np + ds\log\frac{ep}{s}\right)$ instead of the exact $\ell(\mathcal{Z}_\tau) + \log|\mathcal{Z}_\tau|$ in Step 1 for simplicity.

In order to state posterior rate of contraction, we assume that the data is generated by $Y = \theta^* + W$ for some noise matrix W satisfying (5). The signal θ^* is assumed to admits a sparse representation $\theta^* = Q^*Z^*$. Without loss of generality, we can always let the matrix Z^* belong to the set $\bar{\mathcal{Z}}_{p^*,s^*}$. This is because when $\det(Z^*(Z^*)^T) = 0$, there must exist some $Q_1 \in \mathbb{R}^{n\times p_1}$ and $Z_1 \in \bar{\mathcal{Z}}_{p_1,s_1}$ such that $\theta^* = Q^*Z^* = Q_1Z_1$.

Corollary 5.7. *For any $\theta^* = Q^*Z^*$ with $Z^* \in \bar{\mathcal{Z}}_{p^*,s^*}$ specified above, any constants $\lambda, \rho > 0$ and any sufficiently small constant $\delta \in (0, 1)$, there exists some constant $D_{\lambda,\delta,\rho} > 0$ only depending on λ, δ, ρ such that*

$$\mathbb{E}\Pi\left(np + ds\log\frac{ep}{s} > (1+\delta)\left(np^* + ds^*\log\frac{ep^*}{s^*}\right)\Big|Y\right) \le \exp\left(-C'\left(np^* + ds^*\log\frac{ep^*}{s^*}\right)\right)$$

and

$$\mathbb{E}\Pi\left(\|\theta - \theta^*\|_{\mathrm{F}}^2 > M\left(np^* + ds^*\log\frac{ep^*}{s^*}\right)\Big|Y\right) \le \exp\left(-C''\left(np^* + ds^*\log\frac{ep^*}{s^*}\right)\right)$$

for any constant $D > D_{\lambda,\delta,\rho}$ with some constants M, C', C'' only depending on λ, δ, ρ, D.

The rate we have obtained from (5.7) is $np^* + ds^* \log \frac{ep^*}{s^*}$, which is minimax optimal according to [38]. The result can be extended to the case where the entries of Z^* are allowed to take values in an arbitrary discrete set with finite cardinality. To the best of our knowledge, this is the first adaptive estimation result for dictionary learning with optimal prediction rate.

5.7 Nonparametric graphon estimation

Consider a random graph with adjacency matrix $\{A_{ij}\} \in \{0,1\}^{n\times n}$, whose sampling procedure is determined by

$$(\xi_1, ..., \xi_n) \sim \mathbb{P}_\xi, \quad A_{ij}|(\xi_i, \xi_j) \sim \text{Bernoulli}(\theta^*_{ij}), \quad \text{where } \theta^*_{ij} = f^*(\xi_i, \xi_j). \tag{12}$$

For $i \in [n]$, $A_{ii} = \theta^*_{ii} = 0$. Conditioning on $(\xi_1, ..., \xi_n)$, $A_{ij} = A_{ji}$ is independent across $i > j$. The function f^* on $[0,1]^2$, which is assumed to be symmetric, is called graphon. The concept of graphon is originated from graph limit theory [29, 37, 17, 36] and the studies of exchangeable arrays [2, 31]. It is the underlying nonparametric object that generates the random graph.

Let us proceed to specify the function class of graphons. Define the derivative operator by

$$\nabla_{jk} f(x,y) = \frac{\partial^{j+k}}{(\partial x)^j (\partial y)^k} f(x,y),$$

and we adopt the convention $\nabla_{00} f(x,y) = f(x,y)$. The Hölder norm is defined as

$$||f||_{\mathcal{H}_\alpha} = \max_{j+k\le \lfloor\alpha\rfloor} \sup_{x,y\in\mathcal{D}} |\nabla_{jk} f(x,y)| + \max_{j+k=\lfloor\alpha\rfloor} \sup_{(x,y)\neq(x',y')\in\mathcal{D}} \frac{|\nabla_{jk} f(x,y) - \nabla_{jk} f(x',y')|}{||(x-x', y-y')||^{\alpha-\lfloor\alpha\rfloor}},$$

where $\mathcal{D} = \{(x,y) \in [0,1]^2 : x \ge y\}$. Then, the graphon class with Hölder smoothness α is defined by

$$\mathcal{F}_\alpha(L) = \{0 \le f \le 1 : \|f\|_{\mathcal{H}_\alpha} \le L, f(x,y) = f(y,x) \text{ for all } x \in \mathcal{D}\},$$

where $L > 0$ is the radius of the class, which is assumed to be a constant. Recently, a minimax optimal estimator of f^* was proposed by [23] given the knowledge of α. In this section, we propose to solve the adaptive graphon estimation via a Bayes procedure.

As argued in [23], it is sufficient to approximate a graphon with Hölder smoothness by a piecewise constant function, which turns out to be the stochastic block model in the random graph setting. Therefore, we apply the prior distribution in Section 5.1 by equating $f(\xi_i, \xi_j) = \theta_{ij}$. The oracle inequality in Theorem 4.1 gives the desired bias-variance tradeoff of the problem.

Corollary 5.8. *Consider the prior distribution specified in Section 5.1. For the class $\mathcal{F}_\alpha(L)$ with $\alpha, L > 0$ define above and any constant $\lambda > 0$, there exists some constant $D_\lambda > 0$ only*

depending on λ such that

$$\sup_{f^*\in\mathcal{F}_\alpha(L)}\sup_{\mathbb{P}_\xi}\mathbb{E}\Pi\left(\frac{1}{n^2}\sum_{i,j\in[n]}(f(\xi_i,\xi_j)-f^*(\xi_i,\xi_j))^2>M\left(n^{-\frac{2\alpha}{\alpha+1}}+\frac{\log n}{n}\right)\Big|A\right)$$
$$\leq \exp\left(-C'\left(n^{\frac{1}{\alpha+1}}+n\log n\right)\right)$$

for any constant $D>D_\lambda$ with some constants M,C' only depending on λ,D,L.

Remark 5.1. *The expectation in Corollary 5.8 is associated with the joint distribution (12) over both $\{A_{ij}\}$ and $\{\xi_i\}$. Moreover, we do not assume any assumption on the distribution on the design, and the result of Corollary 5.8 holds uniformly over all $\mathbb{P}_\xi$.*

The posterior contraction rate we have obtained for graphon estimation is $n^{-\frac{2\alpha}{\alpha+1}}+\frac{\log n}{n}$, which is minimax optimal according to [23]. When $\alpha\in(0,1)$, the rate is dominated by $n^{-\frac{2\alpha}{\alpha+1}}$, which is the typical two-dimensional nonparametric regression rate. When $\alpha\geq 1$, the rate becomes $\frac{\log n}{n}$, which does not depend on α anymore. The key difference between graphon estimation and nonparametric regression lies in the knowledge of the design sequence $\{\xi_i\}$. A nonparametric regression problem observes the pair $\{(\xi_i,\xi_j),A_{ij}\}$, while graphon estimation only observes the adjacency matrix $\{A_{ij}\}$, resulting in an extra term $\frac{\log n}{n}$ in the rate. To the best of our knowledge, Corollary 5.8 is the first adaptive estimation result on graphon estimation with optimal convergence rate.

5.8 Linear regression under weak ℓ_q ball

Section 5.3 studied high dimensional linear regression under exact sparsity. In this section, we assume the regression coefficients are approximately sparse. Theorem 4.1 allows us to derive optimal posterior rates of contraction even when the prior only charges signals with exact sparsity via a bias variance tradeoff argument. Let us assume the data is generated by $Y=X\beta^*+W\in\mathbb{R}^p$ with some design $X\in\mathbb{R}^{n\times p}$ and some noise vector satisfying (5). We assume β^* is approximately sparse by letting

$$\beta^*\in\mathcal{B}_q(k)=\left\{\beta\in\mathbb{R}^p:\max_{j\in[p]}j|\beta|^q_{(j)}\leq k\right\}$$

with some $q\in[0,1]$, where we order the absolute values of the entries of β by $|\beta|_{(1)}\geq|\beta|_{(2)}\geq ...\geq|\beta|_{(p)}$. Namely, β^* is assumed to have weak ℓ_q radius at most k. To facilitate the presentation, let us define the effective sparsity by $s^*=\lceil x^*\rceil$, where

$$x^*=\max\left\{0\leq x\leq p:x\leq k\left(\frac{n}{\log(ep/x)}\right)^{q/2}\right\}.$$

The effective sparsity s^* is a function of q,k,p,n. Note that in the exact sparse case where $q=0$, we have $s^*=k$. Let us use the prior distribution specified in Section 5.3, and we have the following result.

Corollary 5.9. *Assume* $\max_{j\in[p]} n^{-1/2}\|X_{*j}\| \le L$ *for some constant* $L > 0$. *For any* $q \in [0,1]$, k *and* s^* *specified above and any constants* $\lambda, \rho > 0$, *there exists some constant* $D_{\lambda,\rho} > 0$ *only depending on on* λ, ρ *such that*

$$\sup_{\beta^*\in\mathcal{B}_q(k)} \mathbb{E}\Pi\left(\|X\beta - X\beta^*\|^2 > Ms^*\log\frac{ep}{s^*}\Big|Y\right) \le \exp\left(-C's^*\log\frac{ep}{s^*}\right)$$

for any constant $D > D_{\lambda,\rho}$ *with some constants* M, C' *only depending on* λ, ρ, D, L.

With s^* being the effective sparsity, the posterior rate of contraction has the same form as that of Corollary 5.5. The rate is known to be minimax optimal [18, 44]. In the special case when $k \le p^{1-\eta}\left(\frac{\log p}{n}\right)^{q/2}$ for some constant $\eta \in (0,1)$, the rate has an explicit formula in terms of k, which is $\frac{s^*\log(ep/s^*)}{n} \asymp k\left(\frac{\log p}{n}\right)^{1-q/2}$. When X is an identity matrix, Corollary 5.9 reduces to the results for sparse Gaussian sequence model in [14]. Besides the prediction error, estimation error under approximate sparsity can be derived in the same as Corollary 5.4, and we omit this part due to the similarity.

5.9 Wavelet estimation under Besov space

In this section, we apply the general prior distribution in Section 3 to establish optimal Bayes wavelet estimation under Besov space. Assume the data is generated as

$$Y_{jk} = \theta^*_{jk} + \frac{1}{\sqrt{n}}W_{jk}, \quad k = 1, ..., 2^j; \quad j = 0, 1, 2, ..., \tag{13}$$

where $\{W_{jk}\}$ are i.i.d. $N(0,1)$ variables. It is well known that the sequence model is equivalent to Gaussian white noise model [30], and it is closely related to nonparametric regression and density estimation [9, 42]. Under a wavelet basis, $\{\theta_{jk}\}$ are understood as wavelet coefficients. We assume the true signal $\theta^* = \{\theta^*_{jk}\}$ belongs to the Besov ball defined by

$$\Theta^\alpha_{p,q}(L) = \left\{\theta : \sum_j 2^{ajq}\|\theta_{j*}\|_p^q \le L^q\right\} \tag{14}$$

for some $p, q, \alpha, L > 0$ and $a = \alpha + \frac{1}{2} - \frac{1}{p}$. The Besov ball (14) naturally induces a multi-resolution structure of the signal. This inspires to use a sparse prior distribution independently at each resolution level. That is, we consider a prior distribution Π on θ satisfying

$$\Pi(d\theta) = \prod_j \Pi_j(d\theta_{j*}).$$

The prior distribution Π_j on the jth level for $j < \log_2 n$ is specified as follows:

1. Sample $s_j \sim \pi$ from $[2^j]$, where $\pi(s_j) \propto \frac{\Gamma(s_j)}{\Gamma(s_j/2)}\exp\left(-Ds_j\log\frac{e2^j}{s_j}\right)$;
2. Conditioning on s_j, sample S_j uniformly from $\{S_j \subset [2^j] : |S_j| = s_j\}$;

3. Conditioning on (s_j, S_j), sample $\theta_{jS_j} \sim f_{s_j,S_j,\lambda}$ with $f_{s_j,S_j,\lambda}(\theta_{jS_j}) \propto e^{-\lambda\sqrt{n}\|\theta_{jS_j}\|}$ and set $\theta_{jS_j^c} = 0$.

For $j \geq \log_2 n$, let $\Pi_j(\theta_{j*} = 0) = 1$. Using Theorem 4.1 at each resolution level, we are able to establish the posterior contraction rate in the following corollary.

Corollary 5.10. *For any costants p, q, α satisfying $0 < p, q \leq \infty$, $L > 0$ and $\alpha \geq \frac{1}{p}$ and any constant $\lambda > 0$, there exists some constant D_λ only depending on λ such that*

$$\sup_{\theta^* \in \Theta^\alpha_{p,q}(L)} \mathbb{E}\Pi\left(\|\theta - \theta^*\|^2 > Mn^{-\frac{2\alpha}{2\alpha+1}} \middle| Y\right) \leq \exp\left(-C' n^{\frac{1}{2\alpha+1}} / \log n\right).$$

for any $D > 0$ with some constants M, C' only depending on λ, D, α, p, L.

The result of Corollary 5.10 can be regarded as a Bayes version of Theorem 12.1 of [30] under the same condition. The rate $n^{-\frac{2\alpha}{2\alpha+1}}$ is minimax optimal over the class $\Theta^\alpha_{p,q}(L)$. Posterior contraction for (13) over the class $\Theta^\alpha_{p,q}(L)$ has been investigated by [53, 47, 21, 27] only for a restricted configuration of (p, q, α). In comparison, Corollary 5.10 obtains adaptive optimal posterior contraction rates to all possible combinations of (p, q, α) considered in the frequentist literature [30].

When $p = q = 2$, the class $\Theta^\alpha_{p,q}(L)$ is equivalent to a Sobolev ball. It is worth noting that in this case the prior distribution can be greatly simplified. Let us recast (13) into the sequence model with single index. That is, consider data generated by

$$Y_j = \theta^*_j + \frac{1}{\sqrt{n}} W_j, \quad j = 1, 2, 3, ...,$$

with $\{W_j\}$ being i.i.d. $N(0, 1)$ variables. Assume the true signal $\theta^* = \{\theta^*_j\}$ belongs to the Sobolev ball defined by

$$\mathcal{S}_\alpha(L) = \left\{\theta : \sum_j j^{2\alpha}\theta_j^2 \leq L^2\right\}.$$

We use the following version of the general prior Π in Section 3.

1. Sample $k \sim \pi$ from $[n]$, where $\pi(k) \propto \frac{\Gamma(k)}{\Gamma(k/2)} \exp(-Dk)$;

2. Conditioning on k, sample $\theta_{[k]} = (\theta_1, ..., \theta_k) \sim f_{k,\lambda}$ with $f_{k,\lambda}(\theta_{[k]}) \propto e^{-\lambda\sqrt{n}\|\theta_{[k]}\|}$ and set $\theta_j = 0$ for all $j > k$.

Note that the prior distribution has a missing step compared with the general prior in Section 3. This is because $\mathcal{Z}_k = \{[k]\}$ is a set of singleton so that the model is determined by k and we do not need to perform a further selection. Specializing Theorem 4.1 to this case, we obtain the following result.

Corollary 5.11. *For any constants $\alpha, L > 0$ and any constant $\lambda > 0$, there exists some constant D_λ only depending on λ such that*

$$\sup_{\theta^* \in \mathcal{S}_\alpha(L)} \mathbb{E}\Pi\left(\|\theta - \theta^*\|^2 > Mn^{-\frac{2\alpha}{2\alpha+1}} \middle| Y\right) \leq \exp\left(-C' n^{\frac{1}{2\alpha+1}}\right).$$

for any $D > 0$ with some constants M, C' only depending on λ, D, α, L.

Thus, we have obtained rate-optimal adaptive posterior contraction over the Sobolev ball through a very simple prior distribution.

5.10 Aggregation

Aggregation in nonparametric regression has been considered by [41, 50, 16, 56, 33] among others. Let us start with the nonparametric regression setting with fixed design. That is, the data is generated by

$$Y_i = f^*(x_i) + W_i, \quad i = 1, ..., n, \tag{15}$$

where the noise vector $W = \{W_i\}$ is assumed to satisfy (5). The goal of nonparametric regression is to estimate the true regression function f^* by some estimator $\hat{f}$ under the loss

$$\|\hat{f} - f\|_n^2 = \frac{1}{n}\sum_{i=1}^n \left(\hat{f}(x_i) - f^*(x_i)\right)^2,$$

where $\|\cdot\|_n$ stands for the empirical ℓ_2 norm. Assume we are given a collection of functions $\{f_1, ..., f_p\}$, called the dictionary, and we are also given a subset $\Theta \subset \mathbb{R}^p$. For $\beta \in \Theta$, define $f_\beta = \sum_{j=1}^p \beta_j f_j$. The goal of aggregation is to find an estimator $\hat{f}$ such that its error $\|\hat{f} - f^*\|_n^2$ is comparable to that given by the best among the class $\{f_\beta : \beta \in \Theta\}$. To be specific, one seeks an $\hat{f}$ that satisfies the following oracle inequality

$$\|\hat{f} - f^*\|_n^2 \le (1+\delta) \inf_{\beta\in\Theta} \|f_\beta - f^*\|_n^2 + \Delta_{n,p}(\Theta) \tag{16}$$

with high probability with some arbitrarily small constant $\delta \in (0, 1)$ and some optimal rate function $\Delta_{n,p}(\Theta)$ determined by the class Θ. Various types of aggregation problems include linear, convex and model selection aggregation, etc., which is determined by the choice of the class Θ. In this section, we provide a single Bayes solution to various types of aggregation problems simultaneously and establish the oracle inequality (16) under the posterior distribution.

Since the vector $f_\beta = (f_\beta(x_1), ..., f_\beta(x_n))$ can be represented as $X\beta$ with the matrix X having entries $X_{ij} = f_j(x_i)$ for all $(i, j) \in [n] \times [p]$, the aggregation problem can be recast as a linear regression problem. Define $r = \mathrm{rank}(X)$. Without loss of generality, we assume the first r columns of X span the column space of X. That is, $\mathrm{span}(\{X_{*j}\}_{j\in[r]}) = \mathrm{span}(\{X_{*j}\}_{j\in[p]})$. We are going to use a modified version of the prior distribution defined in Section 5.3.

1. Sample $s \sim \pi$ from $[r]$, where $\pi(s) = \mathcal{N}\frac{\Gamma(s)}{\Gamma(s/2)} \exp\left(-Ds \log \frac{ep}{s}\right)$ for $s < r$ and $\pi(r) = \mathcal{N}\frac{\Gamma(r)}{\Gamma(r/2)} \exp(-Dr)$ with some normalizing constant $\mathcal{N}$;

2. Conditioning on s, sample S uniformly from $\bar{\mathcal{Z}}_s = \{S \subset [p] : |S| = s, \det(X_{*S}^T X_{*S}) > 0\}$ if $s < r$ and set $S = [r]$ if $s = r$;

3. Conditioning on (s, S), sample $\beta_S \sim f_{s,S,\lambda}$ with $f_{s,S,\lambda}(\beta_S) \propto e^{-\lambda\|X_{*S}\beta_S\|}$ and set $\beta_{S^c} = 0$.

The prior Π is similar to the exponential weights used for sparsity pattern aggregation by [45, 46]. Compared with the prior in Section 5.3, it has a modified weight on the model $S = [r]$, which captures the intrinsic dimension of the matrix X. Assuming the data generating process (15), we have the following result implied by Theorem 4.1.

Corollary 5.12. *For any β^* with support $S^* \in \bar{\mathcal{Z}}_{s^*}$ and sparsity $s^* = |S^*| \le r$, any f^*, any constants $\lambda, \rho > 0$ and any sufficiently small constant $\delta \in (0,1)$, there exists some constant $D_{\lambda,\delta,\rho}$ only depending on λ, δ, ρ such that*

$$\begin{aligned} & \mathbb{E}\Pi\left(\|f_\beta - f^*\|_n^2 > (1+\delta)\|f_{\beta^*} - f^*\|_n^2 + M\left(\frac{r}{n} \wedge \frac{s^* \log(ep/s^*)}{n}\right) \middle| Y\right) \\ \le \; & \exp\left(-C'\left(n\|f_\beta - f^*\|_n^2 + r \wedge s^* \log\frac{ep}{s^*}\right)\right) \end{aligned}$$

for any constant $D > D_{\lambda,\delta,\rho}$ with some constants M, C' only depending on λ, δ, ρ, D.

Since $\operatorname{rank}(X) = r$, it is sufficient to establish the posterior oracle inequality for all β^* with sparsity $s^* \le r$. Due to the modified prior weight on the model $S = [r]$, Corollary 5.12 has a better convergence rate than Corollary 5.5. The corresponding frequentist results [45, 46] have leading constant 1 instead of the $(1+\delta)$ in Corollary 5.12. Since our prior induces a subset selection procedure, the presence of an extra small constant δ cannot be avoided [45].

Let us specialize Corollary 5.12 to various types of aggregation problems. Following the notation in [51], define the simplex $\Lambda^p = \{\beta \in \mathbb{R}^p : \sum_j \beta_j = 1, \beta_j \ge 0\}$ and the ℓ_0 ball $\mathcal{B}_0(s^*) = \{\beta \in \mathbb{R}^p : |\operatorname{supp}(\beta)| \le s^*\}$. Then, we consider model selection aggregation $\Theta_{(\mathsf{MS})} = \mathcal{B}_0(1) \cap \Lambda^p$, convex aggregation $\Theta_{(\mathsf{C})} = \Lambda^p$, linear aggregation $\Theta_{(\mathsf{L})} = \mathbb{R}^p$, sparse aggregation $\Theta_{(\mathsf{L}_s)} = \mathcal{B}_0(s^*)$ and sparse convex aggregation $\Theta_{(\mathsf{C}_s)} = \mathcal{B}_0(s^*) \cap \Lambda^p$. For these aggregation problems, define the rate function

$$\Delta_{n,p}(\Theta) = \begin{cases} \frac{\log p}{n}, & \Theta = \Theta_{(\mathsf{MS})}; \\ \sqrt{\frac{1}{n}\log\left(1 + \frac{p}{\sqrt{n}}\right)} & \Theta = \Theta_{(\mathsf{C})}; \\ \frac{r}{n}, & \Theta = \Theta_{(\mathsf{L})}; \\ \frac{s^* \log \frac{ep}{s^*}}{n}, & \Theta = \Theta_{(\mathsf{L}_s)}; \\ \sqrt{\frac{1}{n}\log\left(1 + \frac{p}{\sqrt{n}}\right)} \wedge \frac{s^* \log \frac{ep}{s^*}}{n}, & \Theta = \Theta_{(\mathsf{C}_s)}. \end{cases}$$

Corollary 5.13. *Assume $\max_{j\in[p]} \|f_j\|_n \le 1$. For any f^*, any $\Theta \in \{\Theta_{(\mathsf{MS})}, \Theta_{(\mathsf{C})}, \Theta_{(\mathsf{L})}, \Theta_{(\mathsf{L}_s)}, \Theta_{(\mathsf{C}_s)}\}$, any constants $\lambda, \rho > 0$ and any sufficiently small constant $\delta \in (0,1)$, there exists some constant $D_{\lambda,\delta,\rho}$ only depending on λ, δ, ρ such that*

$$\begin{aligned} & \mathbb{E}\Pi\left(\|f_\beta - f^*\|_n^2 > (1+\delta)\inf_{\beta\in\Theta}\|f_\beta - f^*\|_n^2 + M\left(\Delta_{n,p}(\Theta) \wedge \frac{r}{n}\right) \middle| Y\right) \\ \le \; & \exp\left(-C' n\left(\inf_{\beta\in\Theta}\|f_\beta - f^*\|_n^2 + \Delta_{n,p}(\Theta) \wedge \frac{r}{n}\right)\right) \end{aligned}$$

for any constant $D > D_{\lambda,\delta,\rho}$ with some constants M, C' only depending on λ, δ, ρ, D.

Corollary 5.13 provides a universal aggregation result with a single posterior distribution. The rate is minimax optimal according to [45, 55]. Bayes aggregation was recently studied by [57] under the model misspecification framework [32]. Corollary 5.13 is a stronger result of posterior oracle inequality under weaker assumptions compared with that of [57]. Other types of aggregation results such as ℓ_q aggregation can also be derived directly from Corollary 5.12. The details are omitted in this paper.

6 More results on sparse linear regression

In this section, we provide some further results on posterior contraction rates for linear regression under the ℓ_∞ norm $\|\cdot\|_\infty$. First, let us consider the sparse linear regression setting $Y = X\beta + W$ in Section 5.3. Convergence under the ℓ_∞ norm requires stronger assumptions than convergence under the ℓ_2 norm. Following [19, 35], we assume the mutual coherence condition:

$$n^{-1}X_{*j}^T X_{*j} = 1 \text{ for all } j \in [p] \quad \text{and} \quad \max_{j \neq k} n^{-1}X_{*j}^T X_{*l} \leq \tau. \tag{17}$$

Assuming the data is generated by $Y = X\beta^* + W$ for some regression coefficient β^* with sparsity s^* and some noise vector W satisfying (5), the posterior contraction under the ℓ_∞ norm by using the prior distribution specified in Section 5.3 is given in the following theorem.

Theorem 6.1. *For any $\tau > 0$ and any β^* with sparsity s^* satisfying $\tau s^* \leq 1/9$ and any constants $\lambda, \rho > 0$, there exists some constant $D_{\lambda,\rho} > 0$ only depending on λ, ρ such that*

$$\mathbb{E}\Pi\left(\|\beta - \beta^*\|_\infty > M\sqrt{\frac{\log p}{n}}\Big| Y\right) \leq p^{-C'}$$

for any constant $D > D_{\lambda,\rho}$ with some constants M, C' only depending on λ, ρ, D.

The result of convergence under the ℓ_∞ norm is obtained under the assumption $\tau s^* \leq 1/9$. Such assumption was also made in [19, 11, 35, 15], and it implies the restricted eigenvalue κ_2 defined in (11) to be bounded away from 0 [59]. The convergence rate $\sqrt{\frac{\log p}{n}}$ is optimal under the ℓ_∞ norm. Moreover, with a standard minimal signal strength assumption, Theorem 6.1 immediately implies consistent model selection under the posterior distribution.

While the optimal convergence result for ℓ_∞ norm is well known in the frequentist literature for sparse linear regression, an analogous result for regression with group sparsity is perhaps still open. We provide a Bayes solution to this problem. For simplicity of presentation, we consider the case of identity design $Y = B + W \in \mathbb{R}^{p\times m}$, and the result for the case of a more general design can be derived in a similar way. For any subset $T \subset [p] \times [m]$, let $r(T) = \{i \in [p] : (\{i\} \times [m]) \cap T \neq \varnothing\}$ denote the the rows selected by the set T. The prior Π we use is defined through the following sampling procedure.

1. Sample $T \sim \pi$ in $\{T : T \subset [p] \times [m]\}$ with

$$\pi(T) \propto \frac{\Gamma(|T|)}{\Gamma(|T|/2)} \exp\left(-D\left(m|r(T)| + |r(T)|\log\frac{ep}{|r(T)|} + |T|\log\frac{em|r(T)|}{|T|}\right)\right); \tag{18}$$

2. Conditioning on T, sample $B_T \sim f_{T,\lambda}$ with $f_{T,\lambda}(B_T) \propto e^{-\lambda\sqrt{\sum_{(i,j)\in T} B_{ij}^2}}$ and set $B_{T^c} = 0$.

Compared to the prior distribution specified in Section 5.4, the model selection step for the above prior has a two-level structure. Apart from the correction factor $\frac{\Gamma(|T|)}{\Gamma(|T|/2)}$, the probability mass (18) can be viewed as the product of $e^{-D|S|\left(m+\log\frac{ep}{|S|}\right)}$ and $e^{-D|T|\log\frac{em|S|}{|T|}}$ with $S = r(T)$ denote the row support. Therefore, (18) can be understood as first picking a row support S, and then further select a finer support from $S \times [m]$. In comparison, the prior specified in Section 5.4 does not have the second step. While it only produces B with support with the form $S \times [m]$ for some S, (18) can give an arbitrary support T, which is critical to obtain optimal convergence rate under the ℓ_∞ loss. Let us assume the data is generated from $Y = B^* + W$ for some B^* with row support S^* and noise matrix W satisfying (5), the posterior contraction rate is given in the following theorem.

Theorem 6.2. *For any B^* with row support S^* and sparsity $s^* = |S^*|$, any arbitrarily small constant $\delta > 0$ and any constants $\lambda, \rho > 0$, there exists some constant $D_{\lambda,\delta,\rho} > 0$ only depending on λ, δ, ρ such that*

$$\mathbb{E}\Pi\left(|r(T)| > (1+\delta)s^* \middle| Y\right) \le \exp\left(-C's^*\left(m + \log\frac{ep}{s^*}\right)\right), \tag{19}$$

$$\mathbb{E}\Pi\left(\|B - B^*\|_{\mathrm{F}}^2 > Ms^*\left(m + \log\frac{ep}{s^*}\right) \middle| Y\right) \le \exp\left(-C''s^*\left(m + \log\frac{ep}{s^*}\right)\right) \tag{20}$$

and

$$\mathbb{E}\Pi\left(\|B - B^*\|_\infty > M\sqrt{\log(p+m)} \middle| Y\right) \le (pm)^{-C'''} \tag{21}$$

for any constant $D > D_{\lambda,\delta,\rho}$ with some constants M, C', C'', C''' only depending on λ, δ, ρ, D.

To the best our knowledge, this is the first procedure that achieves the optimal rates simultaneously for both ℓ_2 and ℓ_∞ losses. The $e^{-D|S|\left(m+\log\frac{ep}{|S|}\right)}$ part in (18) preserves the group sparse structure and results in the optimal ℓ_2 result (20). The $e^{-D|T|\log\frac{em|S|}{|T|}}$ part in (18) does a further model selection in a finer resolution, thus giving optimal rate for each coordinate in (21). The subtlety of the simultaneous adaptation under both global and local loss functions is not reflected in an ordinary sparsity setting. When $m = 1$, group sparsity reduces to ordinary sparsity and the two-level model selection prior Π is equivalent to the prior in Section 5.3, so that a one-level model selection is sufficient for the task.

7 Proof of Theorem 4.1

Let us first introduce some notation and give the outline of the proof. Using the fact that

$$\frac{e^{-\frac{1}{2}\|Y - \mathscr{X}_Z(Q)\|^2}}{e^{-\frac{1}{2}\|Y - \mathscr{X}_{Z^*}(Q^*)\|^2}} = e^{-\frac{1}{2}\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\|^2 + \langle Y - \mathscr{X}_{Z^*}(Q^*), \mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\rangle},$$

we can rewrite the posterior distribution as

$$\Pi\left(\mathscr{X}_Z(Q) \in U | Y\right) = \frac{\sum_{\tau\in\mathcal{T}} \exp(-D\epsilon(\mathcal{Z}_\tau))\frac{1}{|\bar{\mathcal{Z}}_\tau|}\sum_{Z\in\bar{\mathcal{Z}}_\tau} R(Z,U)}{\sum_{\tau\in\mathcal{T}} \exp(-D\epsilon(\mathcal{Z}_\tau))\frac{1}{|\bar{\mathcal{Z}}_\tau|}\sum_{Z\in\bar{\mathcal{Z}}_\tau} R(Z)}, \tag{22}$$

where $R(Z,U)$ is defined by

$$\sqrt{\det(\mathscr{X}_Z^T\mathscr{X}_Z)}\left(\frac{\lambda}{\sqrt{\pi}}\right)^{\ell(\mathcal{Z}_\tau)}\int_{\mathscr{X}_Z(Q)\in U} e^{-\frac{1}{2}\|\mathscr{X}_Z(Q)-\mathscr{X}_{Z^*}(Q^*)\|^2+\langle Y-\mathscr{X}_{Z^*}(Q^*),\mathscr{X}_Z(Q)-\mathscr{X}_{Z^*}(Q^*)\rangle-\lambda\|\mathscr{X}_Z(Q)\|}dQ,$$

and $R(Z)=R(Z,\mathbb{R}^N)$. Moreover, for a class of structure indexes $\mathcal{A}\subset\mathcal{T}$, its posterior distribution can be written as

$$\Pi\left(\tau\in\mathcal{A}|Y\right)=\frac{\sum_{\tau\in\mathcal{A}}\exp(-D\epsilon(\mathcal{Z}_\tau))\frac{1}{|\bar{\mathcal{Z}}_\tau|}\sum_{Z\in\bar{\mathcal{Z}}_\tau}R(Z)}{\sum_{\tau\in\mathcal{T}}\exp(-D\epsilon(\mathcal{Z}_\tau))\frac{1}{|\bar{\mathcal{Z}}_\tau|}\sum_{Z\in\bar{\mathcal{Z}}_\tau}R(Z)}. \tag{23}$$

We are going to work with the formulas (23) and (22) to prove (7) and (8), respectively. The main strategy is to lower bound $R(Z^*)$ in the denominator and upper bound $R(Z)$ or $R(Z,U)$ in the numerator given some events holding with high probability. For each $Z\in\mathcal{Z}_\tau$, consider the following events

$$\begin{aligned}
E_Z &= \left\{|\langle W,\mathscr{X}_Z(Q)-\mathscr{X}_{Z^*}(Q^*)\rangle|\le\sqrt{\epsilon^*(\mathcal{Z}_\tau)}\|\mathscr{X}_Z(Q)-\mathscr{X}_{Z^*}(Q^*)\|\text{ for all }Q\in\mathbb{R}^{\ell(\mathcal{Z}_\tau)}\right\},\\
F_Z &= \left\{|\langle W,\mathscr{X}_Z(Q)-\mathscr{X}_{Z^*}(Q^*)\rangle|\le\sqrt{\epsilon^*(\mathcal{Z}_{\tau^*})}\|\mathscr{X}_Z(Q)-\mathscr{X}_{Z^*}(Q^*)\|\text{ for all }Q\in\mathbb{R}^{\ell(\mathcal{Z}_\tau)}\right\},
\end{aligned}$$

where $\epsilon^*(\mathcal{Z}_\tau)=C_1\epsilon(\mathcal{Z}_\tau)+C_2\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2$ and $\epsilon^*(\mathcal{Z}_{\tau^*})=C_1\epsilon(\mathcal{Z}_{\tau^*})+C_2\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2$ for some constants C_1,C_2 to be specified later. The next lemma shows that both events hold with high probability.

Lemma 7.1. *For any constants $C_1>1$ and $C_2>0$, the conditions (4) and (5) imply*

$$\begin{aligned}
\mathbb{P}(E_Z^c) &\le 2\exp\left(-(\rho C_1/16-5)\epsilon(\mathcal{Z}_\tau)-\rho C_2\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2/16\right),\\
\mathbb{P}(F_Z^c) &\le 2\exp\left(5\ell(\mathcal{Z}_\tau)-\rho C_1\epsilon(\mathcal{Z}_{\tau^*})/16-\rho C_2\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2/16\right).
\end{aligned}$$

We also need a lemma to characterize the growing rate of $\epsilon(\mathcal{Z}_\tau)$.

Lemma 7.2. *For any $\beta\ge 2$ and $\alpha\ge 1$, the condition (6) implies*

$$\begin{aligned}
\sum_{\{\tau\in\mathcal{T}:\epsilon(\mathcal{Z}_\tau)\le\alpha\}}\exp\left(\beta\epsilon(\mathcal{Z}_\tau)\right) &\le 4\lceil\alpha\rceil\exp(\beta\lceil\alpha\rceil);\\
\sum_{\{\tau\in\mathcal{T}:\epsilon(\mathcal{Z}_\tau)>\alpha\}}\exp\left(-\beta\epsilon(\mathcal{Z}_\tau)\right) &\le 4\alpha\exp\left(-\beta\lfloor\alpha\rfloor\right);\\
\sum_{\{\tau\in\mathcal{T}:\epsilon(\mathcal{T}_\tau)\le\alpha\}}\exp\left(-\beta\epsilon(\mathcal{Z}_\tau)\right) &\le 6.
\end{aligned}$$

The proofs of Lemma 7.1 and Lemma 7.2 are given in Section 9.

Lower bounding $R(Z^*)$. For $Z^* \in \bar{\mathcal{Z}}_{\tau^*}$ with any $\tau^* \in \mathcal{T}$, we lower bound $R(Z^*)$ by

$$
\begin{aligned}
&\left(\frac{\sqrt{\pi}}{\lambda}\right)^{\ell(\mathcal{Z}_{\tau^*})} R(Z^*) \\
&= \sqrt{\det(\mathscr{X}_{Z^*}^T \mathscr{X}_{Z^*})} \int e^{-\frac{1}{2}\|\mathscr{X}_{Z^*}(Q)-\mathscr{X}_{Z^*}(Q^*)\|^2+\langle Y-\mathscr{X}_{Z^*}(Q^*),\mathscr{X}_{Z^*}(Q)-\mathscr{X}_{Z^*}(Q^*)\rangle-\lambda\|\mathscr{X}_{Z^*}(Q)\|} dQ \\
&= \sqrt{\det(\mathscr{X}_{Z^*}^T \mathscr{X}_{Z^*})} \int e^{-\frac{1}{2}\|\mathscr{X}_{Z^*}(Q)\|^2+\langle Y-\mathscr{X}_{Z^*}(Q^*),\mathscr{X}_{Z^*}(Q)\rangle-\lambda\|\mathscr{X}_{Z^*}(Q)+\mathscr{X}_{Z^*}(Q^*)\|} dQ \qquad (24) \\
&\geq e^{-\lambda\|\mathscr{X}_{Z^*}(Q^*)\|} \sqrt{\det(\mathscr{X}_{Z^*}^T \mathscr{X}_{Z^*})} \int e^{-\frac{1}{2}\|\mathscr{X}_{Z^*}(Q)\|^2+\langle Y-\mathscr{X}_{Z^*}(Q^*),\mathscr{X}_{Z^*}(Q)\rangle-\lambda\|\mathscr{X}_{Z^*}(Q)\|} dQ \qquad (25) \\
&= e^{-\lambda\|\mathscr{X}_{Z^*}(Q^*)\|} \int e^{-\frac{1}{2}\|b\|^2+\langle Y-\mathscr{X}_{Z^*}(Q^*),b\rangle-\lambda\|b\|} db \qquad (26) \\
&\geq e^{-\lambda\|\mathscr{X}_{Z^*}(Q^*)\|} \int e^{-\frac{1}{2}\|b\|^2-\lambda\|b\|} db \exp\left(\int \langle Y-\mathscr{X}_{Z^*}(Q^*), b\rangle \frac{e^{-\frac{1}{2}\|b\|^2-\lambda\|b\|}}{\int e^{-\frac{1}{2}\|b\|^2-\lambda\|b\|} db} db\right) \qquad (27) \\
&= e^{-\lambda\|\mathscr{X}_{Z^*}(Q^*)\|} \int e^{-\frac{1}{2}\|b\|^2-\lambda\|b\|} db. \qquad (28)
\end{aligned}
$$

The equalities (24) and (26) are due to changes of variables. We also use triangle inequality and Jensen's inequality to derive (25) and (27), respectively. The last equality (28) uses the fact that the distribution $\frac{e^{-\frac{1}{2}\|b\|^2-\lambda\|b\|}}{\int e^{-\frac{1}{2}\|b\|^2-\lambda\|b\|} db}$ is spherically symmetric so that its mean is zero. Finally, let us lower bound the integral $\int e^{-\frac{1}{2}\|b\|^2-\lambda\|b\|} db$ by

$$
\begin{aligned}
\int e^{-\frac{1}{2}\|b\|^2-\lambda\|b\|} db &= \frac{2\pi^{\ell(\mathcal{Z}_{\tau^*})/2}}{\Gamma(\ell(\mathcal{Z}_{\tau^*})/2)} \int_0^\infty r^{\ell(\mathcal{Z}_{\tau^*})-1} e^{-\frac{1}{2}r^2-\lambda r} dr \\
&\geq \frac{2\pi^{\ell(\mathcal{Z}_{\tau^*})/2}}{\Gamma(\ell(\mathcal{Z}_{\tau^*})/2)} e^{-\frac{1}{2}\ell(\mathcal{Z}_{\tau^*})-\lambda\sqrt{\ell(\mathcal{Z}_{\tau^*})}} \int_0^{\sqrt{\ell(\mathcal{Z}_{\tau^*})}} r^{\ell(\mathcal{Z}_{\tau^*})-1} dr \\
&= \frac{2\pi^{\ell(\mathcal{Z}_{\tau^*})/2}}{\ell(\mathcal{Z}_{\tau^*})} \frac{[\ell(\mathcal{Z}_{\tau^*})]^{\ell(\mathcal{Z}_{\tau^*})/2}}{\Gamma(\ell(\mathcal{Z}_{\tau^*})/2)} e^{-\frac{1}{2}\ell(\mathcal{Z}_{\tau^*})-\lambda\sqrt{\ell(\mathcal{Z}_{\tau^*})}} \\
&\geq \frac{2(2\pi)^{\ell(\mathcal{Z}_{\tau^*})/2}}{\ell(\mathcal{Z}_{\tau^*})} e^{-\frac{1}{2}\ell(\mathcal{Z}_{\tau^*})-\lambda\sqrt{\ell(\mathcal{Z}_{\tau^*})}}.
\end{aligned}
$$

Combining the above lower bound with (28), we reach the conclusion

$$
R(Z^*) \geq e^{-\lambda\|\mathscr{X}_{Z^*}(Q^*)\|-(1+\lambda+\lambda^{-1})\ell(\mathcal{Z}_{\tau^*})}. \qquad (29)
$$

Note that (29) is a deterministic lower bound for the denominator $R(Z^*)$. The arguments we have used to derive (29) are greatly inspired by the corresponding ones in [14, 15].

Upper bounding $R(Z)\mathbb{I}_{E_Z}$. To facilitate the analysis, we introduce the object

$$
\bar{Q}_Z = \operatorname*{argmin}_{Q\in\mathbb{R}^{\ell(\mathcal{Z}_\tau)}} \|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\|^2. \qquad (30)
$$

The property of least squares implies the following Pythagorean identity,

$$
\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\|^2 = \|\mathscr{X}_Z(Q) - \mathscr{X}_Z(\bar{Q}_Z)\|^2 + \|\mathscr{X}_Z(\bar{Q}_z) - \mathscr{X}_{Z^*}(Q^*)\|^2. \qquad (31)
$$

We first analyze the exponent in the definition of $R(Z)$ on the event E_Z by

$$\begin{aligned}
& -\frac{1}{2}\|\mathscr{X}_Z(Q)-\mathscr{X}_{Z^*}(Q^*)\|^2+\langle Y-\mathscr{X}_{Z^*}(Q^*),\mathscr{X}_Z(Q)-\mathscr{X}_{Z^*}(Q^*)\rangle-\lambda\|\mathscr{X}_Z(Q)\| \\
= & -\frac{1}{2}\|\mathscr{X}_Z(Q)-\mathscr{X}_{Z^*}(Q^*)\|^2+\langle W,\mathscr{X}_Z(Q)-\mathscr{X}_{Z^*}(Q^*)\rangle \\
& +\langle \theta^*-\mathscr{X}_{Z^*}(Q^*),\mathscr{X}_Z(Q)-\mathscr{X}_{Z^*}(Q^*)\rangle-\lambda\|\mathscr{X}_Z(Q)\| \\
\leq & -\frac{1}{2}\|\mathscr{X}_Z(Q)-\mathscr{X}_{Z^*}(Q^*)\|^2+(\sqrt{\epsilon^*(\mathcal{Z}_\tau)}+\lambda)\|\mathscr{X}_Z(Q)-\mathscr{X}_{Z^*}(Q^*)\| \qquad (32)\\
& +\|\theta^*-\mathscr{X}_{Z^*}(Q^*)\|\|\mathscr{X}_Z(Q)-\mathscr{X}_{Z^*}(Q^*)\| \\
& -\lambda\|\mathscr{X}_Z(Q)\|-\lambda\|\mathscr{X}_Z(Q)-\mathscr{X}_{Z^*}(Q^*)\| \\
\leq & 2\left(\sqrt{\epsilon^*(\mathcal{Z}_\tau)}+\lambda\right)^2-\left(\frac{1}{2}-\frac{1}{8}\right)\|\mathscr{X}_Z(Q)-\mathscr{X}_{Z^*}(Q^*)\|^2 \qquad (33)\\
& +2\|\theta^*-\mathscr{X}_{Z^*}(Q^*)\|^2+\frac{1}{8}\|\mathscr{X}_Z(Q)-\mathscr{X}_{Z^*}(Q^*)\|^2-\lambda\|\mathscr{X}_{Z^*}(Q^*)\| \\
\leq & (4+2/C_2)\epsilon^*(\mathcal{Z}_\tau)+8\lambda^2-\frac{1}{4}\|\mathscr{X}_Z(Q)-\mathscr{X}_{Z^*}(Q^*)\|^2-\lambda\|\mathscr{X}_{Z^*}(Q^*)\| \qquad (34)\\
\leq & (4+2/C_2)\epsilon^*(\mathcal{Z}_\tau)+8\lambda^2-\frac{1}{4}\|\mathscr{X}_Z(Q)-\mathscr{X}_Z(\bar{Q}_Z)\|^2-\lambda\|\mathscr{X}_{Z^*}(Q^*)\|. \qquad (35)
\end{aligned}$$

We have used Cauchy-Schwarz inequality and the event E_Z to get (32). The inequality (33) is due to the fact $ab\leq 2a^2+b^2/8$ for all $a,b\geq 0$ and triangle inequality. By rearrangement and the fact $C_2\|\theta^*-\mathscr{X}_{Z^*}(Q^*)\|^2\leq\epsilon^*(\mathcal{Z}_\tau)$, we obatin (34). Finally, the inequality (35) is by the identity (31). The above upper bound implies

$$\begin{aligned}
R(Z)\mathbb{I}_{E_Z} \leq & \left(\frac{\lambda}{\sqrt{\pi}}\right)^{\ell(\mathcal{Z}_\tau)} e^{(4+2/C_2)\epsilon^*(\mathcal{Z}_\tau)+8\lambda^2-\lambda\|\mathscr{X}_{Z^*}(Q^*)\|} \\
& \times\sqrt{\det(\mathscr{X}_Z^T\mathscr{X}_Z)}\int e^{-\frac{1}{4}\|\mathscr{X}_Z(Q)-\mathscr{X}_Z(\bar{Q}_Z)\|^2}dQ \\
= & \left(\frac{\lambda}{\sqrt{\pi}}\right)^{\ell(\mathcal{Z}_\tau)} e^{(4+2/C_2)\epsilon^*(\mathcal{Z}_\tau)+8\lambda^2-\lambda\|\mathscr{X}_{Z^*}(Q^*)\|}\int e^{-\frac{1}{4}\|b\|^2}db \\
= & \left(2\lambda\sqrt{\pi}\right)^{\ell(\mathcal{Z}_\tau)} e^{(4+2/C_2)\epsilon^*(\mathcal{Z}_\tau)+8\lambda^2-\lambda\|\mathscr{X}_{Z^*}(Q^*)\|}.
\end{aligned}$$

Using the fact that $\ell(\mathcal{Z}_\tau)\leq\epsilon^*(\mathcal{Z}_\tau)$ by (4), we reach the conclusion

$$R(Z)\mathbb{I}_{E_Z}\leq e^{(4+2/C_2+\log(2\lambda\sqrt{\pi}))\epsilon^*(\mathcal{Z}_\tau)+8\lambda^2-\lambda\|\mathscr{X}_{Z^*}(Q^*)\|}. \qquad (36)$$

Upper bounding $R(Z,U)\mathbb{I}_{F_Z}$. Let us fix U to be

$$U=\left\{\|\mathscr{X}_Z(Q)-\theta^*\|^2>(1+\delta_2)\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2+M\epsilon(\mathcal{Z}_{\tau^*})\right\}.$$

Let $\xi \in (0, 1/4)$ be a constant to be specified later. When both F_Z and U hold, the exponent in the definition of $R(Z, U)$ is bounded by

$$
\begin{aligned}
& -\frac{1}{2}\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\|^2 + \langle Y - \mathscr{X}_{Z^*}(Q^*), \mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\rangle - \lambda\|\mathscr{X}_Z(Q)\| \\
= \; & -\frac{1}{2}\xi\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\|^2 + \langle W, \mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\rangle + \langle \theta^* - \mathscr{X}_{Z^*}(Q^*), \mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\rangle \\
& -\frac{1}{2}(1-\xi)\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\|^2 - \lambda\|\mathscr{X}_Z(Q)\| \\
\le \; & -\frac{1}{2}\xi\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\|^2 + (\sqrt{\epsilon^*(\mathcal{Z}_{\tau^*})} + \lambda)\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\| \qquad (37) \\
& \langle \theta^* - \mathscr{X}_{Z^*}(Q^*), \mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\rangle - \frac{1}{2}(1-\xi)\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\|^2 \\
& -\lambda\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\| - \lambda\|\mathscr{X}_Z(Q)\| \\
\le \; & \xi^{-1}\left(\sqrt{\epsilon^*(\mathcal{Z}_{\tau^*})} + \lambda\right)^2 - \frac{1}{4}\xi\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\|^2 \qquad (38) \\
& -\frac{1}{2}(1-\xi)\|\mathscr{X}_Z(Q) - \theta^*\|^2 + \frac{1}{2}(1+\xi)\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2 + \xi\,\langle \mathscr{X}_Z(Q) - \theta^*, \theta^* - \mathscr{X}_{Z^*}(Q^*)\rangle \\
& -\lambda\|\mathscr{X}_{Z^*}(Q^*)\| \\
\le \; & \xi^{-1}\left(\sqrt{\epsilon^*(\mathcal{Z}_{\tau^*})} + \lambda\right)^2 - \frac{1}{4}\xi\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\|^2 - \lambda\|\mathscr{X}_{Z^*}(Q^*)\| \qquad (39) \\
& -\frac{1}{2}(1-2\xi)\|\mathscr{X}_Z(Q) - \theta^*\|^2 + \frac{1}{2}(1+2\xi)\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2 \\
\le \; & 8\delta_2^{-1}\lambda^2 - \frac{1}{8}M\epsilon(\mathcal{Z}_{\tau^*}) - \frac{1}{2}\delta_2\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2 \qquad (40) \\
& \frac{1}{16}\delta_2\|\mathscr{X}_Z(Q) - \mathscr{X}_Z(\bar{Q}_Z)\|^2 - \lambda\|\mathscr{X}_{Z^*}(Q^*)\|.
\end{aligned}
$$

We have used Cauchy-Schwarz inequality and the event F_Z to get (37). The inequality (38) is due to the fact $ab \le a^2 + b^2/4$ for all $a, b \ge 0$ and triangle inequality. Then, (39) is by rearranging (38). Finally, we have set

$$
\xi = \frac{1}{4}\delta_2 \quad \text{and} \quad C_2 = \frac{1}{32}\delta_2^2 \qquad (41)
$$

and used (31) to obtain (40) on the event U for all $M > 64\delta_2^{-1}C_1$. Using the above bound, we have

$$
\begin{aligned}
R(Z,U)\mathbb{I}_{F_Z} \; \le \; & \left(\frac{\lambda}{\sqrt{\pi}}\right)^{\ell(\mathcal{Z}_\tau)} e^{-\lambda\|\mathscr{X}_{Z^*}(Q^*)\| + 8\delta_2^{-1}\lambda^2 - \frac{1}{8}M\epsilon(\mathcal{Z}_{\tau^*}) - \frac{1}{2}\delta_2\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2} \\
& \times\sqrt{\det(\mathscr{X}_Z^T\mathscr{X}_Z)}\int e^{-\frac{1}{16}\delta_2\|\mathscr{X}_Z(Q) - \mathscr{X}_Z(\bar{Q}_Z)\|^2} dQ \\
= \; & \left(\frac{4\lambda}{\sqrt{\delta_2}}\right)^{\ell(\mathcal{Z}_\tau)} e^{-\lambda\|\mathscr{X}_{Z^*}(Q^*)\| + 8\delta_2^{-1}\lambda^2 - \frac{1}{8}M\epsilon(\mathcal{Z}_{\tau^*}) - \frac{1}{2}\delta_2\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2}.
\end{aligned}
$$

by the same argument in deriving (36). By $\ell(\mathcal{Z}_\tau) \le \epsilon^*(\mathcal{Z}_\tau)$ from (4), we reach the conclusion

$$
R(Z,U)\mathbb{I}_{F_Z} \le e^{-\lambda\|\mathscr{X}_{Z^*}(Q^*)\| - \frac{1}{16}M\epsilon(\mathcal{Z}_{\tau^*}) - \frac{1}{2}\delta_2\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2}, \qquad (42)
$$

for all $M > \max\left\{64\delta_2^{-1}C_1, 16\log(4\lambda/\sqrt{\delta_2}) + 128\delta_2^{-1}\lambda^2\right\}$.

After obtaining the bounds (29), (36) and (42), we are really to prove the main results.

Proof of (7). First, we use (29) and (36) to bound the ratio $R(Z)\mathbb{I}_{E_Z}/R(Z^*)$.

$$\begin{aligned}
|\bar{\mathcal{Z}}_{\tau^*}|\frac{R(Z)\mathbb{I}_{E_Z}}{R(Z^*)} &\leq e^{8\lambda^2}|\mathcal{Z}_{\tau^*}|\frac{e^{\left[4C_1+2C_1/C_2+C_1\log(2\lambda\sqrt{\pi})\right]\epsilon(\mathcal{Z}_\tau)+\left[4C_2+2+C_2\log(2\lambda\sqrt{\pi})\right]\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2}}{e^{-(1+\lambda+\lambda^{-1})\ell(\mathcal{Z}_{\tau^*})}} \\
&\leq e^{8\lambda^2}\exp\left((1+\lambda+\lambda^{-1})\epsilon(\mathcal{Z}_{\tau^*}) + C_1'\epsilon(\mathcal{Z}_\tau) + C_2'\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2\right),
\end{aligned}$$

where $C_1' = 4C_1 + 2C_1/C_2 + C_1\log(2\lambda\sqrt{\pi})$ and $C_2' = 4C_2 + 2 + C_2\log(2\lambda\sqrt{\pi})$. Let us use the formula (23) with

$$\mathcal{A} = \left\{\epsilon(\mathcal{Z}_\tau) > (1+\delta_1)\epsilon(\mathcal{Z}_{\tau^*}) + \delta_1\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2\right\}.$$

By $Z^* \in \bar{\mathcal{Z}}_{\tau^*}$, we have

$$\begin{aligned}
\mathbb{E}\Pi(\tau\in\mathcal{A}|Y) &\leq \sum_{\tau\in\mathcal{A}}\frac{\exp\left(-D\epsilon(\mathcal{Z}_\tau)\right)}{\exp\left(-D\epsilon(\mathcal{Z}_{\tau^*})\right)}\frac{|\bar{\mathcal{Z}}_{\tau^*}|}{|\bar{\mathcal{Z}}_\tau|}\sum_{Z\in\bar{\mathcal{Z}}_\tau}\mathbb{E}\frac{R(Z)\mathbb{I}_{E_Z}}{R(Z^*)} \qquad (43)\\
&\quad + \sum_{\tau\in\mathcal{A}}\sum_{Z\in\bar{\mathcal{Z}}_\tau}\mathbb{P}(E_Z^c). \qquad (44)
\end{aligned}$$

We use Lemma 7.2 to bound (43) by

$$\begin{aligned}
&\exp\left(8\lambda^2 + (D+\lambda+\lambda^{-1}+1)\epsilon(\mathcal{Z}_{\tau^*}) + C_2'\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2\right)\sum_{\tau\in\mathcal{A}}\exp\left(-(D-C_1')\epsilon(\mathcal{Z}_\tau)\right)\\
\leq\ & 4e^{D+8\lambda^2}\exp\left(-\left((D-C_1'-1)\delta_1 - C_2'\right)\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2\right)\\
&\times\exp\left(-\left((D-C_1'-1)(1+\delta_1) - (D+\lambda+\lambda^{-1}+1)\right)\epsilon(\mathcal{Z}_{\tau^*})\right)\\
\leq\ & 4e^{D+8\lambda^2}\exp\left(-\frac{\delta_1 D}{2}\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2 - \frac{\delta_1 D}{2}\epsilon(\mathcal{Z}_{\tau^*})\right),
\end{aligned}$$

for $D > \max\left\{\frac{\lambda+\lambda^{-1}+1+2(C_1'+1)}{\delta_1/2}, 2(C_1'+1) + \frac{2C_2'}{\delta_1}\right\}$. Using Lemma 7.1, Lemma 7.2 and (4), the second term (44) is bounded by

$$\begin{aligned}
&2\exp\left(-C_2\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2/16\right)\sum_{\tau\in\mathcal{A}}\exp\left(-(\rho C_1/16-6)\epsilon(\mathcal{Z}_\tau)\right)\\
\leq\ & 8e^{14}\exp\left(-\frac{\delta_2^2}{512}\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2 - 7\epsilon(\mathcal{Z}_{\tau^*})\right),
\end{aligned}$$

for $C_1 = \max\{1, 224/\rho\}$ and the value of C_2 is set in (41). Letting $\delta_2 = 8\sqrt{\delta_1/\rho} = 8\sqrt{\delta/\rho}$, we obtain the desired result by combining the bounds of (43) and (44). □

Proof of (8). Let us first use (29) and (42) to bound the ratio $R(Z,U)\mathbb{I}_{F_Z}/R(Z^*)$. That is,

$$\begin{aligned}
\frac{R(Z,U)\mathbb{I}_{F_Z}}{R(Z^*)} &\leq \exp\left(-\left(M/16 - (1+\lambda+\lambda^{-1})\right)\epsilon(\mathcal{Z}_{\tau^*}) - \frac{1}{2}\delta_2\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2\right)\\
&\leq \exp\left(-\frac{M}{32}\epsilon(\mathcal{Z}_{\tau^*}) - \frac{1}{2}\delta_2\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2\right),
\end{aligned}$$

for $M > \max\left\{64\delta_2^{-1}C_1, 16\log(4\lambda/\sqrt{\delta_2}) + 128\delta_2^{-1}\lambda^2, 32(1+\lambda+\lambda^{-1})\right\}$. By the formula (22), we have

$$\begin{aligned}
\mathbb{E}\Pi(U|Y) \quad \leq \quad & \sum_{\tau\in\mathcal{T}\cap\mathcal{A}^c} \frac{\exp\left(-D\epsilon(\mathcal{Z}_\tau)\right)}{\exp\left(-D\epsilon(\mathcal{Z}_{\tau^*})\right)} \frac{|\bar{\mathcal{Z}}_{\tau^*}|}{|\bar{\mathcal{Z}}_\tau|} \sum_{Z\in\bar{\mathcal{Z}}_\tau} \mathbb{E}\frac{R(Z,U)\mathbb{I}_{F_Z}}{R(Z^*)} && (45)\\
& + \sum_{\tau\in\mathcal{T}\cap\mathcal{A}^c} \sum_{Z\in\bar{\mathcal{Z}}_\tau} \mathbb{P}(F_Z^c) && (46)\\
& + \mathbb{E}\Pi(\tau\in\mathcal{A}|Y) && (47)
\end{aligned}$$

The bound of (47) has been derived in the proof of (7). Using Lemma 7.2, we bound (45) by

$$\begin{aligned}
& \exp\left(-\left(\frac{M}{32}-D\right)\epsilon(\mathcal{Z}_{\tau^*}) - \frac{1}{2}\delta_2\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2\right) \sum_{\tau\in\mathcal{T}\cap\mathcal{A}^c} \exp\left(-D\epsilon(\mathcal{Z}_\tau)\right)\\
\leq \quad & 6\exp\left(-\frac{M}{64}\epsilon(\mathcal{Z}_{\tau^*}) - \frac{1}{2}\delta_2\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2\right),
\end{aligned}$$

for $M > \max\left\{64\delta_2^{-1}C_1, 16\log(4\lambda/\sqrt{\delta_2}) + 128\delta_2^{-1}\lambda^2, 32(1+\lambda+\lambda^{-1}), 64D\right\}$. Using Lemma 7.1, Lemma 7.2 and (4), the term (46) is bounded by

$$\begin{aligned}
& 2\exp\left(-\rho C_1\epsilon(\mathcal{Z}_{\tau^*})/16 - \rho C_2\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2\right) \sum_{\tau\in\mathcal{T}\cap\mathcal{A}^c} \exp\left(5\epsilon(\mathcal{Z}_\tau)\right)\\
\leq \quad & 8e^6\exp\left(-\left(\frac{\rho C_1}{16}-2\right)\epsilon(\mathcal{Z}_{\tau^*}) - \left(\rho C_2-\delta_1\right)\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2\right)\\
= \quad & 8e^6\exp\left(-12\epsilon(\mathcal{Z}_{\tau^*}) - \delta_1\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2\right),
\end{aligned}$$

by the relation $C_2 = \delta_2^2/32$, $C_1 = \max\{1, 224/\rho\}$ and $\delta_2 = 8\sqrt{\delta_1/\rho} = 8\sqrt{\delta/\rho}$. The proof is complete by combining the bounds of (45), (46) and (47). □

8 Proofs of corollaries

Proofs of Corollary 4.1 and Corollaries 5.1-5.7. Corollary 4.1 is a direct consequence of Theorem 4.1 by letting $\theta^* = \mathscr{X}_{Z^*}(Q^*)$. Except Corollary 5.4, Corollaries 5.1-5.7 are special cases of Corollary 4.1 in different model settings. By the definitions of κ_1 and κ_2, we have $\|\beta-\beta^*\|^2 \leq \kappa_2^{-2}\|X\beta - X\beta^*\|^2/n$ and $\|\beta-\beta^*\|_1^2 \leq \kappa_1^{-2}s^*\|X\beta-X\beta^*\|^2/n$, which implies Corollary 5.4 from Corollary 5.5. □

Proof of Corollary 5.8. For any ξ, recall that $f(\xi_i,\xi_j) = \theta_{ij} = Q_{z(i)z(j)}$. Then, (8) of Theorem 4.1 implies that

$$\sum_{i,j}\left(f(\xi_i,\xi_j) - f^*(\xi_i,\xi_j)\right)^2 \leq (1+\delta_2)\sum_{i,j}\left(Q^*_{z^*(i)z^*(j)} - f^*(\xi_i,\xi_j)\right)^2 + M\left((k^*)^2 + n\log k^*\right)$$

under the posterior distribution for any $k^* \in [n]$, any $z^* \in [k^*]^n$ and any $Q^* \in \mathbb{R}^{(k^*)^2}$. Lemma 2.1 of [23] implies there exist some $z^* \in [k^*]^n$ and some $Q^* \in \mathbb{R}^{(k^*)^2}$ such that

$$\sum_{i,j}\left(Q^*_{z^*(i)z^*(j)} - f^*(\xi_i,\xi_j)\right)^2 \leq C_3L^2n^2\left(\frac{1}{k^*}\right)^{\alpha\wedge 1},$$

for any $f^* \in \mathcal{F}_\alpha(L)$ and some absolute constant $C_3 > 0$. Therefore,

$$\frac{1}{n^2}\sum_{i,j}(f(\xi_i,\xi_j) - f^*(\xi_i,\xi_j))^2 \le M'\left(\left(\frac{1}{k^*}\right)^{\alpha\wedge 1} + \left(\frac{k^*}{n}\right)^2 + \frac{\log k^*}{n}\right).$$

The proof is complete by choosing $k^* = \lceil n^{\frac{1}{\alpha\wedge 1+1}}\rceil$. □

Proof of Corollary 5.9. The case $q = 0$ is Corollary 5.5. We consider $q \in (0, 1]$. For the effective sparsity defined in Section 5.8, (8) of Theorem 4.1 implies that

$$\|X\beta - X\beta^*\|^2 \le (1+\delta_2)\|X\beta_0 - X\beta^*\|^2 + Ms^*\log\frac{ep}{s^*}$$

under the posterior distribution for all $\beta_0 \in \mathcal{B}_0(s^*)$. By the "Maurey argument" (see Lemma 7.2 in [51]),

$$\frac{1}{n}\|X\beta_0 - X\beta^*\|^2 \le L^2 k^{2/q}(s^*)^{1-2/q},$$

for all $\beta^* \in \mathcal{B}_q(k)$. Therefore,

$$\frac{1}{n}\|X\beta - X\beta^*\|^2 \le M'\left(k^{2/q}(s^*)^{1-2/q} + \frac{s^*\log\frac{ep}{s^*}}{n}\right).$$

By the definition of the effective sparsity s^*, we obtain the desired result. □

Proof of Corollary 5.10. First, we note that by slightly modifying the proof of Theorem 4.1, we can have a more general version of (8), which is

$$\begin{aligned}&\mathbb{E}\Pi\left(\|\mathscr{X}_Z(Q) - \theta^*\|^2 > (1+\delta_2)\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2 + M\epsilon(\mathcal{Z}_{\tau^*}) + t\middle| Y\right)\\ \le\ &\exp\left(-C''\left(\epsilon(\mathcal{Z}_{\tau^*}) + \|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2 + t\right)\right),\end{aligned} \tag{48}$$

for all $t \ge 0$. For every $j < \log_2 n$, the model induced by the prior can be represented in the general framework by letting $Z_j = S_j$, $\tau_j = s_j$, $\mathcal{T}_j = [2^j]$, $\mathcal{Z}_{s_j} = \{S_j \subset [2^j] : |S_j| = s_j\}$, $\ell(\mathcal{Z}_{s_j}) = s_j$ and $Q_j = \sqrt{n}\theta_{jS_j}$. Then, we have the representation $\mathscr{X}_{Z_j}(Q_j) = \sqrt{n}(\theta_{jS_j}^T, 0_{jS_j^c}^T)^T$. The complexity function is $\epsilon_j(\mathcal{Z}_{s_j}) = 2s_j\log\frac{e2^j}{s_j}$, which satisfies (4) and (6). By (48) and letting $t = n^{\frac{1}{2\alpha+1}}/\log_2 n$, we have

$$\begin{aligned}&\mathbb{E}\Pi\left(n\|\theta_{j*} - \theta_{j*}^*\|^2 > (1+\delta_2)n\|\bar\theta_{j*} - \theta_{j*}^*\|^2 + 2Ms_j^*\log\frac{e2^j}{s_j^*} + \frac{n^{\frac{1}{2\alpha+1}}}{\log_2 n}\middle| Y_{j*}\right)\\ \le\ &\exp\left(-C''\frac{n^{\frac{1}{2\alpha+1}}}{\log_2 n}\right),\end{aligned}$$

for any $\bar\theta_{j*} \in \mathbb{R}^{2^j}$ with sparsity s_j^*. Since $\theta^* \in \Theta_{p,q}^\alpha(L)$ implies $\|\theta_{j*}^*\|_p \le L2^{-aj}$, we have

$$\|\bar\theta_{j*} - \theta_{j*}^*\|^2 \le C^* r_{2^j,p}(L2^{-aj}, n^{-1/2})$$

for some absolute constant $C^* > 0$ by the proof of Theorem 11.7 in [30], where $r_{2^j,p}(L2^{-aj}, n^{-1/2})$ is the control function defined in Section 11.5 of [30]. Therefore,

$$\mathbb{E}\Pi(G_j^c|Y_{j*}) \leq \exp\left(-C''\frac{n^{\frac{1}{2\alpha+1}}}{\log_2 n}\right)$$

for all $j < \log_2 n$, where

$$G_j = \left\{\|\theta_{j*} - \theta_{j*}^*\|^2 \leq M' r_{2^j,p}(L2^{-aj}, n^{-1/2}) + \frac{n^{-\frac{2\alpha}{2\alpha+1}}}{\log_2 n}\right\}.$$

Moreover, $\Pi(\theta_{j*} = 0|Y_{j*}) = 1$ for all $j \geq \log_2 n$ by the definition of the prior. Using the independence structure of the posterior distribution, we have

$$\begin{aligned}
&\mathbb{E}\Pi\left(\left(\cap_{j<\log_2 n} G_j\right)^c|Y\right) \leq \sum_{j<\log_2 n} \mathbb{E}\Pi(G_j^c|Y) = \sum_{j<\log_2 n} \mathbb{E}\Pi(G_j^c|Y_{j*}) \\
\leq\ & (\log_2 n)\exp\left(-C''\frac{n^{\frac{1}{2\alpha+1}}}{\log_2 n}\right) \leq \exp\left(-\bar{C}\frac{n^{\frac{1}{2\alpha+1}}}{\log n}\right).
\end{aligned}$$

Finally, the event $\cap_{j<\log_2 n} G_j$ and $\theta_{j*} = 0$ for all $j \geq \log_2 n$ implies

$$\begin{aligned}
\|\theta - \theta^*\|^2 &\leq \sum_{j<\log_2 n} \|\theta_{j*} - \theta_{j*}^*\|^2 + \sum_{j\geq\log_2 n} \|\theta_{j*}^*\|^2 \\
&\leq M' \sum_{j<\log_2 n}\left(r_{2^j,p}(L2^{-aj}, n^{-1/2}) + \frac{n^{-\frac{2\alpha}{2\alpha+1}}}{\log_2 n}\right) + \sum_{j\geq\log_2 n} \|\theta_{j*}^*\|^2 \\
&\leq M'' n^{-\frac{2\alpha}{2\alpha+1}},
\end{aligned}$$

where the last inequality follows the proof of Theorem 12.1 in [30] under the assumption $\alpha \geq \frac{1}{p}$. Hence, the proof is complete. □

Proof of Corollary 5.11. Let us write the model induced by the prior distribution in the general framework by letting $Z = [k]$, $\tau = k$, $\mathcal{T} = [n]$, $\mathcal{Z}_k = \{[k]\}$, $\ell(\mathcal{Z}_k) = k$ and $Q = \sqrt{n}\theta_{[k]}$. Then, we have the representation $\mathscr{X}_Z(Q) = \sqrt{n}(\theta_{[k]}^T, 0_{[k]^c}^T)^T$. The complexity function $\epsilon(\mathcal{Z}_k)$ is $2k$, which satisfies (4) and (6). Then, (8) of Theorem 4.1 implies that

$$\mathbb{E}\Pi\left(n\|\theta - \theta^*\|^2 > (1+\delta_2)n\|\bar{\theta} - \theta^*\|^2 + 2Mk^*\middle|Y\right) \leq \exp\left(-C''\left(k^* + \|\bar{\theta} - \theta^*\|^2\right)\right)$$

for any $\bar{\theta}$ satisfying $\bar{\theta}_j = 0$ for $j > k^*$. Since $\theta^* \in \mathcal{S}_\alpha(L)$, there exists some $\bar{\theta}$ satisfying $\bar{\theta}_j = 0$ for $j > k^*$ such that $\|\bar{\theta} - \theta^*\|^2 \leq L^2(k^*)^{-2\alpha}$. Therefore, $\|\theta - \theta^*\|^2 \leq M'\left((k^*)^{-2\alpha} + \frac{k^*}{n}\right)$ under the posterior distribution. Letting $k^* = \lceil n^{\frac{1}{2\alpha+1}} \rceil$, the proof is complete. □

Proof of Corollary 5.12. Note that the model induced by the prior distribution can be written in a general way by letting $Z = S$, $\tau = s$, $\mathcal{T} = [r]$, $\mathcal{Z}_s = \{S \subset [p] : |S| = s\}$ if $s < r$ and $\mathcal{Z}_r = \{[r]\}$, $\ell(\mathcal{Z}_s) = s$ and $Q = \beta_S$. Then, we have the representation $\mathscr{X}_Z(Q) = X_{*S}\beta_S = X\beta$.

The complexity function we choose is $\epsilon(\mathcal{Z}_s) = 2s\log\frac{ep}{s}$ for $s < r$ and $\epsilon(\mathcal{Z}_r) = 2r$. It is easy to check that $\epsilon(\mathcal{Z}_s)$ satisfies (4) and (6). Using (8) of Theorem 4.1, we have

$$\begin{aligned}
&\mathbb{E}\Pi\left(\|f_\beta - f^*\|_n^2 > (1+\delta_2)\|f_{\beta^*} - f^*\|_n^2 + 2M\frac{s^*\log(ep/s^*)}{n}\Big|Y\right)\\
\leq\ &\exp\left(-C''\left(n\|f_{\beta^*} - f^*\|_n^2 + s^*\log\frac{ep}{s^*}\right)\right),
\end{aligned}$$

for any β^* with sparsity s^*. For this β^*, there exists some β_1 such that $\mathrm{supp}(\beta_1) \subset [r]$ and $f_{\beta^*} = f_{\beta_1}$. Therefore, (8) of Theorem 4.1 implies

$$\begin{aligned}
&\mathbb{E}\Pi\left(\|f_\beta - f^*\|_n^2 > (1+\delta_2)\|f_{\beta_1} - f^*\|_n^2 + 2M\frac{r}{n}\Big|Y\right)\\
\leq\ &\exp\left(-C''\left(n\|f_{\beta_1} - f^*\|_n^2 + r\right)\right).
\end{aligned}$$

Combining the two results by union bound, the proof is complete. □

Proof of Corollary 5.13. Using the corresponding arguments in [46, 51], Corollary 5.13 is implied by Corollary 5.12. □

9 Proofs of technical results

Proof of Lemma 7.1. Consider $\bar{Q}_Z$ defined in (30). Then, we have the bound

$$\begin{aligned}
&|\langle W, \mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\rangle|\\
\leq\ &\|\mathscr{X}_Z(Q) - \mathscr{X}_Z(\bar{Q}_Z)\|\left|\left\langle W, \frac{\mathscr{X}_Z(Q) - \mathscr{X}_Z(\bar{Q}_Z)}{\|\mathscr{X}_Z(Q) - \mathscr{X}_Z(\bar{Q}_Z)\|}\right\rangle\right|\\
&+\|\mathscr{X}_Z(\bar{Q}_Z) - \mathscr{X}_{Z^*}(Q^*)\|\left|\left\langle W, \frac{\mathscr{X}_Z(\bar{Q}_Z) - \mathscr{X}_{Z^*}(Q^*)}{\|\mathscr{X}_Z(\bar{Q}_Z) - \mathscr{X}_{Z^*}(Q^*)\|}\right\rangle\right|\\
\leq\ &\max\left\{\left|\left\langle W, \frac{\mathscr{X}_Z(Q) - \mathscr{X}_Z(\bar{Q}_Z)}{\|\mathscr{X}_Z(Q) - \mathscr{X}_Z(\bar{Q}_Z)\|}\right\rangle\right|, \left|\left\langle W, \frac{\mathscr{X}_Z(\bar{Q}_Z) - \mathscr{X}_{Z^*}(Q^*)}{\|\mathscr{X}_Z(\bar{Q}_Z) - \mathscr{X}_{Z^*}(Q^*)\|}\right\rangle\right|\right\}\\
&\times\sqrt{2}\sqrt{\|\mathscr{X}_Z(Q) - \mathscr{X}_Z(\bar{Q}_Z)\|^2 + \|\mathscr{X}_Z(\bar{Q}_Z) - \mathscr{X}_{Z^*}(Q^*)\|^2}\\
=\ &\sqrt{2}\max\left\{\left|\left\langle W, \frac{\mathscr{X}_Z(Q) - \mathscr{X}_Z(\bar{Q}_Z)}{\|\mathscr{X}_Z(Q) - \mathscr{X}_Z(\bar{Q}_Z)\|}\right\rangle\right|, \left|\left\langle W, \frac{\mathscr{X}_Z(\bar{Q}_Z) - \mathscr{X}_{Z^*}(Q^*)}{\|\mathscr{X}_Z(\bar{Q}_Z) - \mathscr{X}_{Z^*}(Q^*)\|}\right\rangle\right|\right\}\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\|,
\end{aligned}$$

where the last equality is due to (31). By (5), $\left|\left\langle W, \frac{\mathscr{X}_Z(\bar{Q}_Z) - \mathscr{X}_{Z^*}(Q^*)}{\|\mathscr{X}_Z(\bar{Q}_Z) - \mathscr{X}_{Z^*}(Q^*)\|}\right\rangle\right| \leq \frac{1}{\sqrt{2}}\sqrt{\epsilon^*(\mathcal{Z}_\tau)}$ with probabilty at least $1 - \exp(-\rho\epsilon^*(\mathcal{Z}_\tau)/4)$. Now it is sufficient to bound.

$$\sup_{Q\in\mathbb{R}^{\ell(\mathcal{Z}_\tau)}}\left|\left\langle W, \frac{\mathscr{X}_Z(Q) - \mathscr{X}_Z(\bar{Q}_Z)}{\|\mathscr{X}_Z(Q) - \mathscr{X}_Z(\bar{Q}_Z)\|}\right\rangle\right| = \sup_{Q\in\mathbb{R}^{\ell(\mathcal{Z}_\tau)}, \|\mathscr{X}_Z(Q)\|\leq 1}|\langle W, \mathscr{X}_Z(Q)\rangle|.$$

A standard discretization argument as Lemma A.1 in [23] gives

$$\sup_{Q\in\mathbb{R}^{\ell(\mathcal{Z}_\tau)}, \|\mathscr{X}_Z(Q)\|\leq 1}|\langle W, \mathscr{X}_Z(Q)\rangle| \leq 2\max_{1\leq l\leq L}|\langle W, \mathscr{X}_Z(Q_l)\rangle|,$$

where $\{Q_l\}_{1\le l\le L}$ is a subset of $\{Q\in\mathbb{R}^{\ell(\mathcal{Z}_\tau)}:\|\mathscr{X}_Z(Q)\|\le 1\}$ such that for any $Q\in\mathbb{R}^{\ell(\mathcal{Z}_\tau)}$ with $\|\mathscr{X}_Z(Q)\|\le 1$, there exists an $l\in[L]$ that satisfies $\|\mathscr{X}_Z(Q-Q_l)\|\le 1/2$ and a covering number argument gives the bound $L\le\exp\left(5\ell(\mathcal{Z}_\tau)\right)$. Using union bound together with (5), we have $\max_{1\le l\le L}|\langle W,\mathscr{X}_Z(Q_l)\rangle|\le\frac{1}{2\sqrt{2}}\sqrt{\epsilon^*(\mathcal{Z}_\tau)}$ with probability at least

$$1-\exp\left(5\ell(\mathcal{Z}_\tau)-\rho\epsilon^*(\mathcal{Z}_\tau)/16\right)\ge 1-\exp\left(-(\rho C_1/16-5)\epsilon(\mathcal{Z}_\tau)-\rho C_2\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2/16\right),$$

where we have used the condition (4). Using union bound again, we have

$$\sqrt{2}\max\left\{\left|\left\langle W,\frac{\mathscr{X}_Z(Q)-\mathscr{X}_Z(\bar{Q}_Z)}{\|\mathscr{X}_Z(Q)-\mathscr{X}_Z(\bar{Q}_Z)\|}\right\rangle\right|,\left|\left\langle W,\frac{\mathscr{X}_Z(\bar{Q}_Z)-\mathscr{X}_{Z^*}(Q^*)}{\|\mathscr{X}_Z(\bar{Q}_Z)-\mathscr{X}_{Z^*}(Q^*)\|}\right\rangle\right|\right\}\le C_1\sqrt{\epsilon^*(\mathcal{Z}_\tau)},$$

with probability at least $1-2\exp\left(-(\rho C_1/16-5)\epsilon(\mathcal{Z}_\tau)-\rho C_2\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2/16\right)$. This leads to the bound $\mathbb{P}(E_Z^c)\le 2\exp\left(-(\rho C_1/16-5)\epsilon(\mathcal{Z}_\tau)-\rho C_2\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2/16\right)$. A similar argument also leads to the bound $\mathbb{P}(F_Z^c)\le 2\exp\left(5\ell(\mathcal{Z}_\tau)-\rho C_1\epsilon(\mathcal{Z}_{\tau^*})/16-\rho C_2\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2/16\right)$. □

Proof of Lemma 7.2. The first inequality holds because

$$\begin{aligned}\sum_{\{\tau\in\mathcal{T}:\epsilon(\mathcal{T}_\tau)\le\alpha\}}\exp\left(\beta\epsilon(\mathcal{Z}_\tau)\right) &\le \sum_{t=1}^{\lceil\alpha\rceil}\sum_{\{\tau\in\mathcal{T}:t-1<\epsilon(\mathcal{Z}_\tau)\le t\}}e^{\beta\epsilon(\mathcal{Z}_\tau)}+e^\beta\\ &\le \sum_{t=1}^{\lceil\alpha\rceil}te^{\beta t}+e^\beta\\ &\le 2\lceil\alpha\rceil\frac{e^\beta}{e^\beta-1}e^{\beta\lceil\alpha\rceil}\\ &\le 4\lceil\alpha\rceil\exp(\beta\lceil\alpha\rceil),\end{aligned}$$

by $\beta\ge 2$. The second inequality holds because

$$\begin{aligned}\sum_{\{\tau\in\mathcal{T}:\epsilon(\mathcal{Z}_\tau)>\alpha\}}\exp\left(-\beta\epsilon(\mathcal{Z}_\tau)\right) &\le \sum_{t=\lfloor\alpha\rfloor}^{\infty}\sum_{\{\tau\in\mathcal{T}:t<\epsilon(\mathcal{Z}_\tau)\le t+1\}}e^{-\beta\epsilon(\mathcal{Z}_\tau)}\\ &\le \sum_{t=\lfloor\alpha\rfloor}^{\infty}(t+1)e^{-\beta t}\\ &\le 2\sum_{t=\lfloor\alpha\rfloor}^{\infty}\exp\left(-\left(\beta-\frac{\log\lfloor\alpha\rfloor}{\lfloor\alpha\rfloor}\right)t\right) \qquad (49)\\ &\le 4\alpha\exp(\beta\lfloor\alpha\rfloor),\end{aligned}$$

for $\beta\ge 2$ and $\alpha\ge 1$. The inequality (49) is because $\log t\le\frac{\log\lfloor\alpha\rfloor}{\lfloor\alpha\rfloor}t$ for all $t\ge\lfloor\alpha\rfloor$. Finally,

$$\begin{aligned}\sum_{\{\tau\in\mathcal{T}:\epsilon(\mathcal{T}_\tau)\le\alpha\}}\exp\left(-\beta\epsilon(\mathcal{Z}_\tau)\right) &\le 1+\sum_{t=1}^{\infty}te^{-\beta(t-1)}\\ &\le 1+e^\beta\sum_{t=1}^{\infty}e^{-(\beta-1)t}\\ &\le 6,\end{aligned}$$

for $\beta \geq 2$. $\square$

10 Proofs in Section 6

Proof of Theorem 6.1. The assumption $\tau s^* \leq 1/9$ and the argument in the proof of Theorem 1 of [35] implies

$$\max_{|S|\leq(2+\delta)s^*} \| \left(n^{-1}X_{*S}^T X_{*S}\right)^{-1} \|_{\text{op}} \leq \max_{|S|\leq(2+\delta)s^*} \| \left(n^{-1}X_{*S}^T X_{*S}\right)^{-1} \|_{\ell_1} \leq 4 \tag{50}$$

for $\delta \leq 1/4$. Define $\hat{\beta}_S = \min_b \|Y - X_{*S}b\|^2$. Then it is easy to see that $\|Y - X_{*S}\beta_S\|^2 = \|Y - X_{*S}\hat{\beta}_S\|^2 + \|X_{*S}(\beta_S - \hat{\beta}_S)\|^2$. Define the distribution $\mathcal{L}(\hat{\beta}_S, X_{*S}, \lambda)$ of β_S that has density function

$$\frac{\exp\left(-\frac{1}{2}\|X_{*S}\beta_S - X_{*S}\hat{\beta}_S\|^2 - \lambda\|X_{*S}\beta_S\|\right)}{\int \exp\left(-\frac{1}{2}\|X_{*S}\beta_S - X_{*S}\hat{\beta}_S\|^2 - \lambda\|X_{*S}\beta_S\|\right) d\beta_S}. \tag{51}$$

Then, according to the formula of the posterior distribution, to sample β from the posterior distribution is equivalent to first sample S from $\Pi(S|Y)$ and then sample $\beta_S \sim \mathcal{L}(\hat{\beta}_S, X_{*S}, \lambda)$ to form $\beta^T = (\beta_S^T, 0_{S^c}^T)$. Hence, the posterior distribution can be represented as

$$\sum_S \Pi(S|Y)\Pi_S(\cdot|Y) = \sum_S \omega(S)\mathcal{L}(\hat{\beta}_S, X_{*S}, \lambda) \otimes \delta_{S^c},$$

where $\Pi(S|Y) = \omega(S)$ and $\Pi_S(\cdot|Y) = \mathcal{L}(\hat{\beta}_S, X_{*S}, \lambda) \otimes \delta_{S^c}$ with

$$\omega(S) \propto \frac{\pi(|S|)}{|\bar{\mathcal{Z}}_{|S|}|}\left(\frac{\lambda}{\sqrt{\pi}}\right)^{|S|} \mathcal{N}_{X_{*S}\hat{\beta}_S,\lambda} e^{-\frac{1}{2}\|Y - X_{*S}\hat{\beta}_S\|^2} \mathbb{I}\{|\bar{\mathcal{Z}}_{|S|}| > 0\}. \tag{52}$$

The number $\mathcal{N}_{y,\lambda}$ for any vector y and any scalar λ is defined as

$$\mathcal{N}_{y,\lambda} = \int \exp\left(-\frac{1}{2}\|t - y\|^2 - \lambda\|t\|\right) dt. \tag{53}$$

Define the event

$$E = \left\{\max_{j\in[p]} \left|\frac{X_j^T W}{\sqrt{n}}\right| \leq C_1\sqrt{\log p}\right\} \tag{54}$$

for some constant $C_1 > 0$ to be determined later. We have

$$\begin{aligned}
&\mathbb{E}\Pi\left(\|\beta - \beta^*\|_\infty > M\sqrt{\frac{\log p}{n}}\Big|Y\right) \\
=\ &\mathbb{E}\sum_{|S|\leq(1+\delta)s^*} \omega(S)\Pi_S\left(\|\beta_S - \beta_S^*\|_\infty \vee \|\beta_{S^c}^*\|_\infty > M\sqrt{\frac{\log p}{n}}\Big|Y\right) + \mathbb{E}\Pi(|S| > (1+\delta)s^*|Y) \\
\leq\ &\mathbb{E}\sum_{\substack{|S|\leq(1+\delta)s^* \\ \|\beta_{S^c}^*\|_\infty \leq C_2\sqrt{\frac{\log p}{n}}}} \omega(S)\Pi_S\left(\|\beta_S - \hat{\beta}_S\|_\infty > \frac{1}{2}M\sqrt{\frac{\log p}{n}}\Big|Y\right)\mathbb{I}_E + \mathbb{E}\sum_{\substack{|S|\leq(1+\delta)s^* \\ \|\beta_{S^c}^*\|_\infty > C_2\sqrt{\frac{\log p}{n}}}} \omega(S)\mathbb{I}_E \\
&+\mathbb{P}(E^c) + \mathbb{E}\Pi(|S| > (1+\delta)s^*|Y)
\end{aligned} \tag{55}$$

for some constant $C_2 > 0$ to be determined later. The inequality (55) is due to the inequality $\|\beta_S - \beta^*_S\|_\infty \le \|\beta_S - \hat\beta_S\|_\infty + \|\hat\beta_S - \beta^*_S\|$ and

$$E \subset \left\{ \|\hat\beta_S - \beta^*_S\|_\infty \le \frac{1}{2} M \sqrt{\frac{\log p}{n}} \right\}, \tag{56}$$

for all S that satisfies $|S| \le (1+\delta)s^*$ and $\|\beta^*_{S^c}\|_\infty \le C_2\sqrt{\frac{\log p}{n}}$. Let us give a proof for (56). By the definition of $\hat\beta_S$, we have $X^T_{*S}X_{*S}\hat\beta_S = X^T_{*S}Y = X^T_{*S}X_{*S}\beta^*_S + X^T_{*S}X_{*S^c}\beta^*_{S^c} + X^T_{*S}W$, which implies

$$\|\hat\beta_S - \beta^*_S\|_\infty \le 4\|X^T_{*S}X_{*S}(\hat\beta_S - \beta^*_S)\|_\infty/n \le \frac{4}{n}\|X^T_{*S}X_{*S^c}\beta^*_{S^c}\|_\infty + \frac{4}{n}\|X^T_{*S}W\|_\infty.$$

Note that $\frac{4}{n}\|X^T_{*S}X_{*S^c}\beta^*_{S^c}\|_\infty = \frac{4}{n}\|X^T_{*S}X_{*S^*\cap S^c}\beta^*_{S^*\cap S^c}\|_\infty \le 8s^*\tau\|\beta^*_{S^c}\|_\infty \le C_2\sqrt{\frac{\log p}{n}}$ due to $\|\beta^*_{S^c}\|_\infty \le C_2\sqrt{\frac{\log p}{n}}$. We also have $\frac{4}{n}\|X^T_{*S}W\|_\infty \le \frac{4}{n}\max_{j\in[p]}|X^T_j W| \le 4C_1\sqrt{\frac{\log p}{n}}$. Therefore, (56) is proved for some $M/2 \ge 4C_1 + C_2$.

In view of (55), it is sufficient to bound the four terms in (55). The last term is bounded as a result of (9). The third term is bounded by $p^{-\left(\frac{C_1\rho}{2}-1\right)}$ using (5) and a union bound argument. Let us give a bound for the first term.

$$\begin{aligned}
&\Pi_S\left(\|\beta_S - \hat\beta_S\|_\infty > \frac{1}{2}M\sqrt{\frac{\log p}{n}}\Big|Y\right) \\
\le\ &\sum_{j\in S}\Pi_S\left(|\beta_j - \hat\beta_j| > \frac{1}{2}M\sqrt{\frac{\log p}{n}}\Big|Y\right) \\
\le\ &\sum_{j\in S}\exp\left(-\frac{1}{2}tM\sqrt{\log p}\right)\mathbb{E}_{\Pi_S}\left(e^{\sqrt{n}t|\beta_j-\hat\beta_j|}\Big|Y\right),
\end{aligned} \tag{57}$$

where $\mathbb{E}_{\Pi_S}(\cdot|Y)$ is the posterior expectation with the distribution $\Pi_S(\cdot|Y) = \mathcal{L}(\hat\beta_S, X_{*S}, \lambda)$ and $t > 0$ is some number to be specified later. Using the formula of the density (51), for

any unit vector $v \in \mathbb{R}^{|S|}$, we have

$$
\begin{aligned}
& \mathbb{E}_{\Pi_S}\left(e^{\sqrt{n}tv^T(\beta_S-\hat{\beta}_S)}\Big|Y\right) \\
= & \frac{\int \exp\left(\sqrt{n}tv^T(\beta_S-\hat{\beta}_S)-\frac{1}{2}\|X_{*S}\beta_S-X_{*S}\hat{\beta}_S\|^2-\lambda\|X_{*S}\beta_S\|\right)d\beta_S}{\int \exp\left(-\frac{1}{2}\|X_{*S}\beta_S-X_{*S}\hat{\beta}_S\|^2-\lambda\|X_{*S}\beta_S\|\right)d\beta_S} \\
= & e^{\frac{1}{2}t^2\|(n^{-1}X_{*S}^TX_{*S})^{-1/2}v\|^2}\frac{\int \exp\left(-\frac{1}{2}\|X_{*S}(\beta_S-\hat{\beta}_S-tn^{-1/2}(n^{-1}X_{*S}^TX_{*S})^{-1}v)\|^2-\lambda\|X_{*S}\beta_S\|\right)d\beta_S}{\int \exp\left(-\frac{1}{2}\|X_{*S}(\beta-\hat{\beta}_S)\|^2-\lambda\|X_{*S}\beta_S\|\right)d\beta_S} \\
\leq & \exp\left(\frac{1}{2}t^2\|(n^{-1}X_{*S}^TX_{*S})^{-1/2}v\|^2+\lambda t\|(n^{-1}X_{*S}^TX_{*S})^{-1/2}v\|\right) \qquad (58) \\
\leq & \exp\left(\frac{1}{2}\lambda^2+t^2\|(n^{-1}X_{*S}^TX_{*S})^{-1/2}v\|^2\right) \\
\leq & \exp\left(\frac{1}{2}\lambda^2+16t^2\right), \qquad (59)
\end{aligned}
$$

where the inequality (58) is due to a change of variable and triangle inequality and the inequality (59) is by (50). Specializing v so that $v^T(\beta_S-\hat{\beta}_S)=\pm(\beta_j-\hat{\beta}_j)$, we have

$$
\mathbb{E}_{\Pi_S}\left(e^{\sqrt{n}t|\beta_j-\hat{\beta}_j|}\Big|Y\right) \leq \mathbb{E}_{\Pi_S}\left(e^{\sqrt{n}t(\beta_j-\hat{\beta}_j)}\Big|Y\right)+\mathbb{E}_{\Pi_S}\left(e^{-\sqrt{n}t(\beta_j-\hat{\beta}_j)}\Big|Y\right) \leq 2e^{\frac{1}{2}\lambda^2+16t^2}.
$$

Letting $t=\sqrt{\log p}$, we have

$$
\Pi_S\left(\|\beta_S-\hat{\beta}_S\|_\infty > \frac{1}{2}M\sqrt{\frac{\log p}{n}}\Big|Y\right) \leq 2e^{\lambda^2/2}p^{-(\frac{M}{2}-17)},
$$

which bounds the first term of (55).

Now, let us give a bound for the second term of (55). Given $j=\text{argmax}_{l\in[p]}|\beta_j^*|$, for any $S\subset[p]$ such that $j\notin S$, define $S'=S\cup\{j\}$. We are going to provide a bound for $\omega(S)/\omega(S')$ on the event E to argue the model S' is favored over the model S under the posterior distribution if $|\beta_j^*|$ is large. Because of (50), $|\bar{\mathcal{Z}}_{|S|}|=|\mathcal{Z}_{|S|}|=\binom{p}{|S|}$ for all $|S|\leq(1+\delta)s^*$. By (52), we have

$$
\frac{\omega(S)}{\omega(S')}=\frac{\pi(|S|)}{\pi(|S'|)}\frac{\binom{p}{|S'|}}{\binom{p}{|S|}}\frac{\sqrt{\pi}}{\lambda}\frac{\mathcal{N}_{X_{*S}\hat{\beta}_S,\lambda}}{\mathcal{N}_{X_{*S'}\hat{\beta}_{S'},\lambda}}e^{-\frac{1}{2}\|Y-X_{*S}\hat{\beta}_S\|^2+\frac{1}{2}\|Y-X_{*S'}\hat{\beta}_{S'}\|^2}.
$$

Since $\frac{\pi(|S|)}{\pi(|S'|)}\leq\exp\left(2D\log(ep)\right)$, $\frac{\binom{p}{|S'|}}{\binom{p}{|S|}}\leq p$, $\frac{\mathcal{N}_{X_{*S}\hat{\beta}_S,\lambda}}{\mathcal{N}_{X_{*S'}\hat{\beta}_{S'},\lambda}}\leq e^{\lambda\|X_{*S}\hat{\beta}_S-X_{*S'}\hat{\beta}_{S'}\|}$ by the definition (53) and a change of variable, and $-\frac{1}{2}\|Y-X_{*S}\hat{\beta}_S\|^2+\frac{1}{2}\|Y-X_{*S'}\hat{\beta}_{S'}\|^2=\frac{1}{2}\|X_{*S}\hat{\beta}_S\|^2-\frac{1}{2}\|X_{*S'}\hat{\beta}_{S'}\|^2$, we have

$$
\frac{\omega(S)}{\omega(S')}\leq\frac{\sqrt{\pi}}{\lambda}(ep)^{2D+1}e^{\lambda\|X_{*S}\hat{\beta}_S-X_{*S'}\hat{\beta}_{S'}\|+\frac{1}{2}\|X_{*S}\hat{\beta}_S\|^2-\frac{1}{2}\|X_{*S'}\hat{\beta}_{S'}\|^2}. \qquad (60)
$$

Let P_S and $P_{S'}$ stand for the projection matrix onto the column spaces of X_{*S} and $X_{*S'}$, respectively. Then $X_{*S}\hat{\beta}_S = P_S Y$ and $X_{*S'}\hat{\beta}_{S'} = P_{S'}Y$. Let F be the orthogonal complement of the columns space of X_{*S} in the column space of $X_{*S'}$, and then define P_F to be the associated projection matrix. It is easy to see that $P_{S'} = P_S + P_F$ and $P_S P_F = 0$. Thus, the exponent of (60) equals $\lambda\|P_F Y\| - \frac{1}{2}\|P_F Y\|^2 \le -\frac{1}{4}\|P_F Y\|^2 + \lambda^2 \le -\frac{1}{8}\|P_F X\beta^*\|^2 + \frac{1}{4}\|P_F W\|^2 + \lambda^2$. We are going to give a lower bound on $\|P_F X\beta^*\|^2$ and an upper bound on $\|P_F W\|^2$. To facilitate the proof, we bound $\|P_S X_{*j}\|^2$ as

$$\begin{aligned}\|P_S X_{*j}\|^2 &= X_{*j}^T X_{*S}(X_{*S}^T X_{*S})^{-1} X_{*S}^T X_{*j} \\ &\le 4n\|n^{-1} X_{*S}^T X_{*j}\|^2 \\ &\le 8ns^*\tau^2 \le \frac{n}{10} \end{aligned} \tag{61}$$

by (50) and $\tau s^* \le 1/9$. The noise part $\|P_F W\|^2$ is bounded as

$$\begin{aligned}\|P_F W\|^2 &= \left\| \frac{(I-P_S)X_{*j}X_{*j}^T(I-P_S)}{\|(I-P_S)X_{*j}\|^2} W \right\|^2 \\ &\le \frac{|X_{*j}^T(I-P_S)W|^2}{\|(I-P_S)X_{*j}\|^2} \\ &\le \frac{2|X_{*j}^T W|^2 + 2|X_{*j}^T P_S W|^2}{9n/10} \qquad (62) \\ &\le 8C_1^2 \log p, \qquad (63)\end{aligned}$$

where (62) is because of (61) and (63) is derived from the event E and the following argument that

$$\begin{aligned}|X_{*j}^T P_S W|^2 &= |X_{*j}^T X_{*S}(X_{*S}^T X_{*S})^{-1} X_{*S}^T W|^2 \\ &\le 16n\|X_{*j}^T X_{*S}/n\|^2 \|X_{*S}^T W/\sqrt{n}\|^2 \\ &\le 32C_1^2 (s^*\tau)^2 n\log p \le \frac{1}{2}C_1^2 n\log p\end{aligned}$$

by (50) and the event E. The signal part $\|P_F X\beta^*\|^2$ is lower bounded by

$$\|P_F X\beta^*\| \ge \|(I-P_S)X_{*j}\||\beta_j^*| - \sum_{l\in S^*\cap(S\cup\{j\})^c} \frac{|X_{*j}^T(I-P_S)X_{*l}|}{\|(I-P_S)X_{*j}\|}|\beta_l^*|,$$

where the first term on the right hand side above is lower bounded by $\sqrt{9n/10}|\beta_j^*|$ by (61), and the second term is upper bounded by

$$\sum_{l\in S^*\cap\{j\}^c} |\beta_l^*|\frac{|X_{*j}^T X_{*l}|}{\|(I-P_S)X_{*j}\|} + \sum_{l\in S^*\cap S^c} |\beta_l^*|\frac{|X_{*j}^T P_S X_{*l}|}{\|(I-P_S)X_{*j}\|} \le 7\sqrt{n}|\beta_j^*|/9$$

due to (61), (50), $\tau s^* \le 1/9$ and the fact $|\beta_j^*| = \max_{l\in[p]}|\beta_l^*|$. Therefore, $\|P_F X\beta^*\| \ge \sqrt{n}|\beta_j^*|/7$. When $|\beta_j^*| \ge 400C_1\sqrt{\frac{\log p}{n}}$, we have $-\frac{1}{8}\|P_F X\beta^*\|^2 + \frac{1}{4}\|P_F W\|^2 \le -2C_1^2\log p$.

Plugging this bound into (60), we have $\frac{\omega(S)}{\omega(S')} \le \frac{\sqrt{\pi}}{\lambda} e^{2D+1+\lambda^2} p^{-(2C_1^2-2D-1)}$, which implies

$$\sum_{\substack{|S|\le(1+\delta)s^* \\ j\notin S}} \omega(S)\mathbb{I}_E = \sum_{\substack{|S|\le(1+\delta)s^* \\ j\notin S}} \frac{\omega(S)}{\omega(S\cup\{j\})}\omega(S\cup\{j\})\mathbb{I}_E \le \frac{\sqrt{\pi}}{\lambda} e^{2D+1+\lambda^2} p^{-(2C_1^2-2D-1)}.$$

By letting $C_2 = 400C_1$, a mathematical induction argument in [15] leads to a bound on the second term of (55) that

$$\sum_{\substack{|S|\le(1+\delta)s^* \\ \|\beta^*_{S^c}\|_\infty > C_2\sqrt{\frac{\log p}{n}}}} \omega(S)\mathbb{I}_E \le p^{-C_3},$$

for some constant C_3 depending on C_1, D, λ. Moreover, C_3 is increasing with C_1.

Finally, combining the bounds for the four terms in (55), we get

$$\begin{aligned} & \mathbb{E}\Pi\left(\|\beta-\beta^*\|_\infty > M\sqrt{\frac{\log p}{n}}\Big| Y\right) \\ \le\ & 2e^{\lambda^2/2}p^{-\left(\frac{M}{2}-17\right)} + p^{-C_3} + p^{-\left(\frac{C_1\rho}{2}-1\right)} + e^{-C's^*\log\frac{ep}{s^*}} \\ \le\ & p^{-C_4}, \end{aligned}$$

for some M, C_4 depending on ρ, λ, D. □

Proof of Theorem 6.2. For B_T, we use $\|\cdot\|$ to denote the ℓ_2 norm as $\|B_T\| = \sqrt{\sum_{(i,j)\in T} B_{ij}^2}$. Let us first establish (19) and (20). The proof is close to that of Theorem 4.1. By the definition of the prior, the posterior distribution has formula

$$\Pi(B\in U|Y) = \frac{\sum_T \alpha(T)R(T,U)}{\sum_T \alpha(T)R(T)}, \tag{64}$$

where $R(T,U)$ is defined by

$$\left(\frac{\lambda}{\sqrt{\pi}}\right)^{|T|} \int_{(B_T,0_{T^c})\in U} e^{-\frac{1}{2}\|(B_T,0_{T^c})-B^*\|^2+\langle W,(B_T,0_{T^c})-B^*\rangle-\lambda\|B_T\|} dB_T,$$

$R(T) = R(T,\mathbb{R}^{p\times m})$ and

$$\alpha(T) = \exp\left(-D\left(|r(T)|\log\frac{ep}{|r(T)|} + |T|\log\frac{em|r(T)|}{|T|}\right)\right).$$

Moreover, for a set of subsets $\mathcal{A}$, the posterior distribution can be written as

$$\Pi(T\in\mathcal{A}|Y) = \frac{\sum_{T\in\mathcal{A}}\alpha(T)R(T)}{\sum_T \alpha(T)R(T)}. \tag{65}$$

We need to give a lower bound for $R(T^*)$ with $T^* = S^*\times[m]$ and give upper bounds for $R(T)$ and $R(T,U)$. For each subset T, define the following events

$$\begin{aligned} E_T &= \left\{|\langle W,(B_T,0_{T^c})-B^*\rangle| \le \sqrt{C_1\left(m|r(T)|+|r(T)|\log\frac{ep}{|r(T)|}\right)}\|(B_T,0_{T^c})-B^*\| \text{ for all } B_T\in\mathbb{R}^{|T|}\right\}, \\ F_T &= \left\{|\langle W,(B_T,0_{T^c})-B^*\rangle| \le \sqrt{C_1\left(ms^*+s^*\log\frac{ep}{s^*}\right)}\|(B_T,0_{T^c})-B^*\| \text{ for all } B_T\in\mathbb{R}^{|T|}\right\} \end{aligned}$$

for some constant $C_1 > 0$ to be determined later. A special case of Lemma 7.1 gives

$$\mathbb{P}(E_T^c) \le 2e^{-(\rho C_1/16-6)\left(m|r(T)|+|r(T)|\log\frac{ep}{|r(T)|}\right)} \quad \text{and} \quad \mathbb{P}(F_T^c) \le 2e^{5m|r(T)|-\frac{\rho C_1}{16}\left(ms^*+s^*\log\frac{ep}{s^*}\right)}. \tag{66}$$

The same arguments used for deriving (29), (36) and (42) imply

$$R(T^*) \ge e^{-\lambda\|B^*\|-(1+\lambda+\lambda^{-1})ms^*}, \tag{67}$$

$$R(T)\mathbb{I}_{E_T} \le (2\lambda)^{|T|}e^{2\lambda^2-\lambda\|B^*\|+2C_1\left(m|r(T)|+|r(T)|\log\frac{ep}{|r(T)|}\right)}, \tag{68}$$

$$R(T,U)\mathbb{I}_{F_T} \le (2\sqrt{2}\lambda)^{|T|}e^{2\lambda^2-\lambda\|B^*\|-\left(\frac{1}{8}M-2C_1\right)\left(ms^*+s^*\log\frac{ep}{s^*}\right)}, \tag{69}$$

with $U = \left\{\|B-B^*\| > M\left(ms^* + s^*\log\frac{ep}{s^*}\right)\right\}$. Let $\mathcal{A} = \{|r(T)| > (1+\delta)s^*\}$. By the formula (65) and the inequalities (67) and (68), we have

$$\begin{aligned}
&\mathbb{E}\Pi(T\in\mathcal{A}|Y)\\
\le\ & \sum_{T\in\mathcal{A}}\frac{\alpha(T)}{\alpha(T^*)}\mathbb{E}\frac{R(T)}{R(T^*)}\mathbb{I}_{E_T} + \sum_{T\in\mathcal{A}}\mathbb{P}(E_T^c)\\
\le\ & e^{(C_2+D)\left(ms^*+s^*\log\frac{ep}{s^*}\right)}\sum_{s>(1+\delta)s^*}\sum_{S:|S|=s}e^{-(D-C_2)(ms+s\log\frac{ep}{s})}\sum_{T:r(T)=S}e^{-D|T|\log\frac{ems}{|T|}}\\
&+2\sum_{s>(1+\delta)s^*}\sum_{S:|S|=s}e^{-(\rho C_1/16-7)\left(ms+s\log\frac{ep}{s}\right)}\\
\le\ & e^{-C'\left(ms^*+s^*\log\frac{ep}{s^*}\right)}
\end{aligned}$$

for some sufficiently large D with C_2 only depending on C_1 and λ and C' only depending on D, ρ, λ. By the formula (64) and the inequalities (67) and (42), we have

$$\begin{aligned}
&\mathbb{E}\Pi(B\in U|Y)\\
\le\ & \sum_{T\in\mathcal{A}^c}\frac{\alpha(T)}{\alpha(T^*)}\mathbb{E}\frac{R(T,U)}{R(T^*)}\mathbb{I}_{F_T} + \sum_{T\in\mathcal{A}^c}\mathbb{P}(F_T^c) + e^{-C'\left(ms^*+s^*\log\frac{ep}{s^*}\right)}\\
\le\ & e^{-\left(\frac{1}{8}M-C_3-D\right)\left(ms^*+s^*\log\frac{ep}{s^*}\right)}\sum_{s\le(1+\delta)s^*}\sum_{S:|S|=s}e^{-D(ms+s\log\frac{ep}{s})}\sum_{T:r(T)=S}e^{-D|T|\log\frac{ems}{|T|}}\\
&+2e^{-\frac{\rho C_1}{16}\left(ms^*+s^*\log\frac{ep}{s^*}\right)}\sum_{s\le(1+\delta)s^*}\sum_{S:|S|=s}e^{6ms} + e^{-C'\left(ms^*+s^*\log\frac{ep}{s^*}\right)}\\
\le\ & e^{-C''\left(ms^*+s^*\log\frac{ep}{s^*}\right)}
\end{aligned}$$

for some sufficiently large M with C_3 only depending on C_1 and λ and C'' only depending on D, ρ, λ. Hence, (19) and (20) are proved.

Now let us proceed to prove (21). We are going to use the similar argument as that of Theorem 6.1. Note that the posterior distribution can be represented as

$$\sum_T \Pi(T|Y)\Pi_T(\cdot|Y) = \sum_T \omega(T)\mathcal{L}(Y_T,\lambda)\otimes\delta_{T^c},$$

where $\Pi(T|Y) = \omega(T)$ and $\Pi_T(\cdot|Y) = \mathcal{L}(Y_T, \lambda) \otimes \delta_{T^c}$ with

$$\omega(T) \propto \left(\frac{\lambda}{\sqrt{\pi}}\right)^{|T|} \alpha(T) \mathcal{N}_{Y_T,\lambda} e^{\frac{1}{2}\|Y_T\|^2}.$$

The distribution $B_T \sim \mathcal{L}(Y_T, \lambda)$ is defined through the density function

$$\mathcal{N}_{Y_T,\lambda}^{-1} e^{-\frac{1}{2}\|B_T - Y_T\|^2 - \lambda\|B_T\|},$$

where $\mathcal{N}_{Y_T,\lambda}$ is the normalizing constant defined in (53). Define the event

$$E = \left\{ \max_{(i,j)\in[p]\times[m]} |W_{ij}| \le C_1 \sqrt{\log(pm)} \right\}$$

for some constant $C_1 > 0$. We have

$$\begin{aligned}
& \mathbb{E}\Pi\left(\|B - B^*\|_\infty > M\sqrt{\log(pm)} \middle| Y\right) \\
\le\ & \mathbb{E} \sum_{|r(T)|\le(1+\delta)s^*} \omega(T)\Pi_T\left(\|B_T - Y_T\|_\infty > \frac{1}{2}M\sqrt{\log(pm)} \middle| Y\right)\mathbb{I}_E + \mathbb{E} \sum_{\substack{|r(T)|\le(1+\delta)s^* \\ \|B^*_{T^c}\|_\infty > C_2\sqrt{\log(pm)}}} \omega(T)\mathbb{I}_E \\
& + \mathbb{P}(E^c) + \mathbb{E}\Pi\left(|r(T)| > (1+\delta)s^* | Y\right). \qquad (70)
\end{aligned}$$

It is sufficient to bound the four terms in (70). The last term is bounded by (19). Using (5) and a union bound argument, we bound the third term in (70) as $\mathbb{P}(E^c) \le (pm)^{-\left(\frac{\rho C_1^2}{2} - 1\right)}$. Using the same arguments in deriving (57) and (59), we have

$$\begin{aligned}
& \Pi_T\left(\|B_T - Y_T\|_\infty > \frac{1}{2}M\sqrt{\log(pm)} \middle| Y\right) \\
\le\ & \sum_{(i,j)\in T} \exp\left(-\frac{1}{2}tM\sqrt{\log(pm)}\right) \mathbb{E}_{\Pi_T}\left(e^{\sqrt{n}t|B_{ij} - Y_{ij}|} \middle| Y\right) \\
\le\ & 2e^{\lambda^2/2} pm e^{-\frac{1}{2}tM\sqrt{\log(pm)} + t^2} \le 2e^{\lambda^2/2}(pm)^{-\left(\frac{M}{2} - 2\right)}
\end{aligned}$$

by choosing $t = \sqrt{\log(pm)}$. This bounds the first term of (70). Now let us provide a bound for the first term of (70). Given some $(i,j) \in [p] \times [m]$, for any subset T such that $(i,j) \notin T$, use the notation $T' = T \cup \{(i,j)\}$. To facilitate the proof, we need an upper bound for $\omega(T)/\omega(T')$ on the event E. Direct calculation gives

$$\frac{\omega(T)}{\omega(T')} = \frac{\sqrt{\pi}}{\lambda} \frac{\alpha(T)}{\alpha(T')} \frac{\mathcal{N}_{Y_T,\lambda}}{\mathcal{N}_{Y_{T'},\lambda}} e^{-\frac{1}{2}Y_{ij}^2}.$$

Since $\frac{\alpha(T)}{\alpha(T')} \le (epm)^{3D}$, and

$$\frac{\mathcal{N}_{Y_T,\lambda}}{\mathcal{N}_{Y_{T'},\lambda}} \le C_\lambda e^{\lambda|Y_{ij}|} \qquad (71)$$

for some constant C_λ only depending on λ, we have $\omega(T)/\omega(T') \le C_\lambda \frac{\sqrt{\pi}}{\lambda}(epm)^{3D} e^{\lambda|Y_{ij}| - \frac{1}{2}Y_{ij}^2}$. The inequality (71) will be established in the end of the proof. Since

$$\begin{aligned} \lambda|Y_{ij}| - \frac{1}{2}Y_{ij}^2 &\le \lambda^2 - \frac{1}{4}Y_{ij}^2 \\ &\le \lambda^2 - \frac{1}{8}(B_{ij}^*)^2 + \frac{1}{4}W_{ij}^2 \le \lambda^2 - \frac{1}{4}C_1^2 \log(pm) \end{aligned}$$

when $|B_{ij}^*| > 2C_1\sqrt{\log(pm)}$ on the event E. Hence,

$$\frac{\omega(T)}{\omega(T')}\mathbb{I}_E \le C_\lambda \frac{\sqrt{\pi}}{\lambda} e^{3D+\lambda^2}(pm)^{-\left(\frac{1}{4}C_1^2 - 3D\right)}. \tag{72}$$

Let $C_2 = 2C_1$ and define $\{(i_1, j_1), ..., (i_q, j_q)\}$ to be the set such that $|B_{i_l j_l}^*| > 2C_1\sqrt{\log(pm)}$ for all $l \in [q]$. Then, we have

$$\{\|B_{T^c}^*\|_\infty > C_2\sqrt{\log(pm)}\} \subset \cup_{l\in[q]}\{(i_l, j_l) \notin T\},$$

which implies

$$\begin{aligned} \sum_{\substack{|r(T)|\le(1+\delta)s^* \\ \|B_{T^c}^*\|_\infty > C_2\sqrt{\log(pm)}}} \omega(T) &\le \sum_{l\in[q]} \sum_{T\in\{T:(i_l,j_l)\notin T\}} \frac{\omega(T)}{\omega(T\cup\{(i_l,j_l)\})}\omega(T\cup\{(i_l,j_l)\}) \\ &\le C_\lambda \frac{\sqrt{\pi}}{\lambda} e^{3D+\lambda^2}(pm)^{-\left(\frac{1}{4}C_1^2-3D\right)} \sum_{l\in[q]} \sum_{T\in\{T:(i_l,j_l)\notin T\}} \omega(T\cup\{(i_l,j_l)\}) \\ &\le (pm)^{-\bar{C}} \end{aligned}$$

by (72) for some constant $\bar{C}$ with sufficiently large C_1. Combining the bounds for the four terms in (55), we reach the conclusion (21).

Finally, let us establish (71) to close the proof. By change of variable, we have

$$\mathcal{N}_{Y_T,\lambda} = \int_{\mathbb{R}^{|T|-1}} \int_{\mathbb{R}} e^{-\frac{1}{2}b_1^2 - \frac{1}{2}\|b_2\|^2 - \lambda\sqrt{(b_1+\|Y_T\|)^2+\|b_2\|^2}} db_1 db_2,$$

and

$$\mathcal{N}_{Y_{T'},\lambda} = \int_{\mathbb{R}} \int_{\mathbb{R}^{|T|-1}} \int_{\mathbb{R}} e^{-\frac{1}{2}(b_1^2+b_3^2) - \frac{1}{2}\|b_2\|^2 - \lambda\sqrt{(b_1+\|Y_{T'}\|)^2+\|b_2\|^2+b_3^2}} db_1 db_2 db_3.$$

Therefore, triangle inequality implies

$$\mathcal{N}_{Y_{T'},\lambda} \ge \mathcal{N}_{Y_T,\lambda} \int_{\mathbb{R}} e^{-\frac{1}{2}b^2 - \lambda|b|} db e^{-\lambda|\|Y_T\| - \|Y_{T'}\||} \ge C_\lambda^{-1} e^{-\lambda|Y_{ij}|},$$

where $C_\lambda = \left(\int_{\mathbb{R}} e^{-\frac{1}{2}b^2-\lambda|b|}db\right)^{-1}$. Thus, the proof is complete. □

References

[1] Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli. Exact recovery of sparsely used overcomplete dictionaries. *arXiv preprint arXiv:1309.1952*, 2013.

[2] David J Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

[3] Sergey Bakin. Adaptive regression and model selection in data mining problems. 1999.

[4] Sayantan Banerjee and Subhashis Ghosal. Posterior convergence rates for estimating large precision matrices using graphical models. *Electronic Journal of Statistics*, 8(2): 2111–2137, 2014.

[5] Andrew Barron, Mark J Schervish, and Larry Wasserman. The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561, 1999.

[6] Andrew R Barron. *The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions.* Univ.of Illinois, 1988.

[7] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

[8] Lucien Birgé and Pascal Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268, 2001.

[9] Lawrence D Brown and Mark G Low. Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics*, 24(6):2384–2398, 1996.

[10] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media, 2011.

[11] Florentina Bunea et al. Consistent selection via the lasso for high dimensional approximating regression models. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, pages 122–137. Institute of Mathematical Statistics, 2008.

[12] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

[13] Ismaël Castillo. On bayesian supremum norm contraction rates. *The Annals of Statistics*, 42(5):2058–2091, 2014.

[14] Ismaël Castillo and Aad van der Vaart. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069–2101, 2012.

[15] Ismael Castillo, Johannes Schmidt-Hieber, and Aad W van der Vaart. Bayesian linear regression with sparse priors. *arXiv preprint arXiv:1403.0735*, 2014.

[16] Olivier Catoni. Statistical learning theory and stochastic optimization. *Lecture Notes in Mathematics*, 1851, 2004.

[17] Persi Diaconis and Svante Janson. Graph limits and exchangeable random graphs. *arXiv preprint arXiv:0712.2749*, 2007.

[18] David L Donoho and Iain M Johnstone. Minimax risk overl p-balls forl p-error. *Probability Theory and Related Fields*, 99(2):277–303, 1994.

[19] David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *Information Theory, IEEE Transactions on*, 52(1):6–18, 2006.

[20] Kai-Tai Fang, Samuel Kotz, and Kai Wang Ng. *Symmetric multivariate and related distributions*. Chapman and Hall, 1990.

[21] Chao Gao and Harrison H Zhou. Adaptive bayesian estimation via block prior. *arXiv preprint arXiv:1312.3937*, 2013.

[22] Chao Gao and Harrison H Zhou. Rate-optimal posterior contraction for sparse pca. *The Annals of Statistics*, 43(2):785–818, 2015.

[23] Chao Gao, Yu Lu, and Harrison H Zhou. Rate-optimal graphon estimation. *arXiv preprint arXiv:1410.5837*, 2014.

[24] Subhashis Ghosal, Jayanta K Ghosh, and RV Ramamoorthi. Posterior consistency of dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1):143–158, 1999.

[25] Subhashis Ghosal, Jayanta K Ghosh, and Aad W Van Der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531, 2000.

[26] John A Hartigan. Direct clustering of a data matrix. *Journal of the american statistical association*, 67(337):123–129, 1972.

[27] Marc Hoffmann, Judith Rousseau, and Johannes Schmidt-Hieber. On adaptive posterior concentration rates. *arXiv preprint arXiv:1305.5270*, 2013.

[28] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[29] Douglas N Hoover. Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*, 2, 1979.

[30] Iain M Johnstone. Gaussian estimation: Sequence and wavelet models. 2011.

[31] Olav Kallenberg. On the representation theorem for exchangeable arrays. *Journal of Multivariate Analysis*, 30(1):137–154, 1989.

[32] Bas JK Kleijn and Aad W van der Vaart. Misspecification in infinite-dimensional bayesian statistics. *The Annals of Statistics*, pages 837–877, 2006.

[33] Gilbert Leung and Andrew R Barron. Information theory and mixing least-squares regressions. *Information Theory, IEEE Transactions on*, 52(8):3396–3410, 2006.

[34] Karim Lounici, Massimiliano Pontil, Sara Van De Geer, and Alexandre B Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, 2011.

[35] Karim Lounici et al. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of statistics*, 2:90–102, 2008.

[36] László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012.

[37] László Lovász and Balázs Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957, 2006.

[38] Yu Lu and Harrison H Zhou. Minimax rates for product of three matrices. 2015.

[39] Zongming Ma and Yihong Wu. Volume ratio, sparsity, and minimaxity under unitarily invariant norms. *arXiv preprint arXiv:1306.3609*, 2013.

[40] Ryan Martin, Raymond Mess, and Stephen G Walker. Empirical bayes posterior concentration in sparse high-dimensional linear models. *arXiv preprint arXiv:1406.7718*, 2014.

[41] Arkadi Nemirovski. Topics in non-parametric statistics. 2000.

[42] Michael Nussbaum. Asymptotic equivalence of density estimation and gaussian white noise. *The Annals of Statistics*, pages 2399–2430, 1996.

[43] Bruno A Olshausen. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[44] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over-balls. *Information Theory, IEEE Transactions on*, 57(10):6976–6994, 2011.

[45] Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

[46] Philippe Rigollet, Alexandre B Tsybakov, et al. Sparse estimation by exponential weighting. *Statistical Science*, 27(4):558–575, 2012.

[47] Vincent Rivoirard and Judith Rousseau. Posterior concentration rates for infinite dimensional exponential families. *Bayesian Analysis*, 7(2):311–334, 2012.

[48] Xiaotong Shen and Larry Wasserman. Rates of convergence of posterior distributions. *The Annals of Statistics*, pages 687–714, 2001.

[49] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[50] Alexandre B Tsybakov. Optimal rates of aggregation. In *Learning Theory and Kernel Machines*, pages 303–313. Springer, 2003.

[51] Alexandre B Tsybakov. Aggregation and minimax optimality in high-dimensional estimation. In *Proceedings of the International Congress of Mathematicians*, 2014.

[52] Sara A Van De Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[53] AW Van der Vaart and JH Van Zanten. Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, pages 1435–1463, 2008.

[54] Nicolas Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90, 2012.

[55] Zhan Wang, Sandra Paterlini, Frank Gao, and Yuhong Yang. Adaptive minimax estimation over sparse ℓ_q-hulls. *arXiv preprint arXiv:1108.1961*, 2011.

[56] Yuhong Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10(1):25–47, 2004.

[57] Yun Yang and David B Dunson. Minimax optimal bayesian aggregation. *arXiv preprint arXiv:1403.1345*, 2014.

[58] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[59] Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, pages 1567–1594, 2008.