The Annals of Statistics
2020, Vol. 48, No. 5, 2848–2878
https://doi.org/10.1214/19-AOS1909

A GENERAL FRAMEWORK FOR BAYES STRUCTURED LINEAR MODELS

BY CHAO GAO[1], AAD W. VAN DER VAART[2] AND HARRISON H. ZHOU[3]

[1]*Department of Statistics, University of Chicago, chaogao@galton.uchicago.edu*
[2]*Mathematical Institute, Faculty of Science, Leiden University, avdvaart@math.leidenuniv.nl*
[3]*Department of Statistics, Yale University, huibin.zhou@yale.edu*

High dimensional statistics deals with the challenge of extracting structured information from complex model settings. Compared with a large number of frequentist methodologies, there are rather few theoretically optimal Bayes methods for high dimensional models. This paper provides a unified approach to both Bayes high dimensional statistics and Bayes nonparametrics in a general framework of structured linear models. With a proposed two-step prior, we prove a general oracle inequality for posterior contraction under an abstract setting that allows model misspecification. The general result can be used to derive new results on optimal posterior contraction under many complex model settings including recent works for stochastic block model, graphon estimation and dictionary learning. It can also be used to improve upon posterior contraction results in literature including sparse linear regression and nonparametric aggregation. The key of the success lies in the novel two-step prior distribution: one for model structure, that is, model selection, and the other one for model parameters. The prior on the parameters of a model is an elliptical Laplace distribution that is capable of modeling signals with large magnitude, and the prior on the model structure involves a factor that compensates the effect of the normalizing constant of the elliptical Laplace distribution, which is important to attain rate-optimal posterior contraction.

1. Introduction. Theory for posterior distribution has been extensively investigated in Bayes nonparametrics recently. Important works such as [6, 7, 15, 26, 27, 29, 51, 57] 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.

Despite the success of theoretical advancements of Bayes nonparametrics, there are not many theories developed for Bayes high dimensional statistics. A few exceptions are [17] on the sparse Gaussian sequence model, [4] on bandable precision matrix estimation and [24] on sparse PCA. Recently, [16] established posterior contraction rates for sparse linear regression with a spike and slab prior under comparable assumptions for theoretical justification of the Lasso estimator [9, 52]. The results of [16] include posterior contraction rates for prediction error, estimation error, oracle inequalities and model selection consistency. However, sparse linear regression is just 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, multitask learning, etc. It is not clear whether the theory in [16] can be extended to these more complex settings.

This paper provides a unified methodology and theory for both Bayes high dimensional statistics and Bayes nonparametric statistics in a general framework of structured linear models. We first introduce a unified view of various high dimensional and nonparametric models,

Received July 2016; revised July 2019.
MSC2020 subject classifications. Primary 62C10; secondary 62F15.
Key words and phrases. Oracle inequality, stochastic block model, graphon, sparse linear regression, aggregation, dictionary learning, posterior contraction.

and then propose a single prior distribution for all models considered in our framework. Optimal rates of convergence of the posterior distributions are established under appropriate conditions. The results directly lead to exact minimax posterior contraction rates in stochastic block model, biclustering, sparse linear regression, regression with group sparsity, multitask learning and dictionary learning. Moreover, we also derive a general posterior oracle inequality that allows arbitrary model misspecification. Applications of the posterior oracle inequality help us obtain posterior contraction rates even for models that are not included in our framework of structured linear models. Examples considered in this paper include nonparametric graphon estimation, various forms of nonparametric aggregation, linear regression with approximate sparsity and wavelet estimation under Besov spaces.

In the heart of the general theory is a novel 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 [17] for Gaussian sequence models. A key ingredient of the prior distribution is that the tail of the distribution on the model parameter cannot be too light [16, 17], which motivates [16, 17] to use the independent Laplace prior on the parameter. Though the prior distribution leads to optimal posterior contraction rates in Gaussian sequence model [17], it requires some excessive assumptions on the design matrix when it is applied to sparse linear regression [16]. The proposal in this paper is an elliptical Laplace prior. With this choice, not only are we able to weaken the assumptions in [16], but we can also solve a more general class of problems in a unified way. To compensate the influence of the normalizing constant of the elliptical Laplace distribution, a correction factor on the prior mass is added in the model selection step. Without this correction factor, the posterior contraction rate would be suboptimal.

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 a rate optimal posterior oracle inequality and a posterior rate of contraction. The main results are illustrated by ten examples ranging from nonparametric estimation to 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 and the supplement [23].

We close this section by introducing some notation. Given an integer d, we use $[d]$ to denote the set $\{1, 2, \ldots, d\}$, and $[d]^n$ to denote $\{(i_1, \ldots, i_n) \in \mathbb{R}^n : i_1, \ldots, i_n \in [d]\}$. 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 $a, b \in \mathbb{R}$, $a \vee b = \max(a, b)$ 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. Consider the following structured linear model:

$$Y = \mathscr{X}_Z(Q) + W \in \mathbb{R}^N, \tag{1}$$

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 $\mathscr{X}_Z(\cdot)$. In the example of sparse linear regression $Y = X\beta + W$ with a sparse regression coefficient $\beta = (\beta_S^T, 0_{S^c}^T)^T$ for some subset S, we have $Q = \beta_S$, $Z = S$ and $\mathscr{X}_Z(\cdot) = X_{*S}$. This gives the representation $X\beta = X_{*S}\beta_S = \mathscr{X}_Z(Q)$. In the general theory, the structure Z is in some discrete space $\mathcal{Z}_\tau$, which is further indexed by $\tau \in \mathcal{T}$ for some finite set $\mathcal{T}$. For sparse linear regression, $\mathcal{Z}_\tau$ is the set of models of size τ. We introduce a function $\ell(\mathcal{Z}_\tau)$ to denote 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 intrinsic dimension of the structured linear model. The complexity of the model is defined by the quantity

$$\ell(\mathcal{Z}_\tau) + \log|\mathcal{Z}_\tau|, \tag{2}$$

the sum of the intrinsic dimension and the logarithmic cardinality of the structure space. The definition of (2) has a frequentist root (see, e.g., [5, 10, 62]). As we are going to show later, (2) will be the posterior contraction rate that we target at. Moreover, in all the examples considered in the paper, (2) will be the minimax rate under the prediction loss. We require linearity of the operator $\mathscr{X}_Z(\cdot)$. That is, given any $Z \in \mathcal{Z}_\tau$ with any $\tau \in \mathcal{T}$, we have

$$\mathscr{X}_Z(Q_1 + Q_2) = \mathscr{X}_Z(Q_1) + \mathscr{X}_Z(Q_2), \quad \text{for all } Q_1, Q_2 \in \mathbb{R}^{\ell(\mathcal{Z}_\tau)}. \tag{3}$$

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 only 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 the graph nodes. Moreover, it is easy to see that the parameter Q is of dimension k^2, when we do not impose symmetry for Q. 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 subset $S \subset [p]$. 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. *Multiple linear regression with group sparsity.* It refers to the model XB with $B \in \mathbb{R}^{p\times m}$ being a coefficient matrix with nonzero rows in some subset $S \subset [p]$. It can be represented in a general form similarly as the sparse linear regression with $\ell(\mathcal{Z}_s) = ms$.

5. *Multitask learning.* Similar to the last example, multitask learning is the collection of m regression problems. We consider XB for some $B \in \mathbb{R}^{p\times m}$. The jth column of B can be 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 $\tau = (p,s)$, $\mathcal{T} = \{(p,s) \in [n \wedge d] \times [n] : s \le p\}$, $\mathcal{Z}_{p,s} = \{Z \in \{-1,0,1\}^{p\times d} : \max_{j\in[d]} |\operatorname{supp}(Z_{*j})| \le s\}$ and $\ell(\mathcal{Z}_{p,s}) = np$.

In several examples above, Q or $\mathscr{X}_Z(Q)$ can be a matrix instead of a vector. Alternative definitions of these examples in the general framework are available by vectorization and Kronecker products. For example, in dictionary learning, the linear operator $\mathscr{X}_Z : Q \mapsto QZ$ is from matrix to matrix. By the formula $\operatorname{vec}(QZ) = (Z^T \otimes I_n)\operatorname{vec}(Q)$, the linear operator $\mathscr{X}_Z$ can be identified with the matrix $Z^T \otimes I_n \in \mathbb{R}^{nd\times np}$, which is also a linear operator from $\mathbb{R}^{np}$ to $\mathbb{R}^{nd}$. Similar rearrangements apply to other examples as well.

In addition to the six examples above, we have four more examples that can be well approximated by the general structured linear models.

7. *Nonparametric graphon estimation.* For an undirected graph, the distribution of its adjacency matrix $\{A_{ij}\} \in \{0,1\}^{n\times n}$ is determined by $A_{ij}|(\xi_i,\xi_j) \sim \text{Bernoulli}(f(\xi_i,\xi_j))$, where $\{\xi_i\}$ are latent variables with some joint distribution $\mathbb{P}_\xi$. The symmetric nonparametric function f is called graphon. It governs the underlying data generating process of a random graph. When f is assumed to be in a Hölder class, it can be approximated by the stochastic block model.

8. *Aggregation.* Consider a nonparametric regression problem $Y_i = f(x_i) + W_i$ for $i \in [n]$. Given a collection of functions $\{f_1,\dots,f_p\}$ and a subset $\Theta \subset \mathbb{R}^p$, the goal of aggregation is to approximate f with some estimator so that the error is comparable to what is given by the best among the class $\{f_\beta = \sum_{j=1}^p \beta_j f_j : \beta \in \Theta\}$. In Section 5.8, we show that the regression function f can be approximated by the general structured linear model.

9. *Linear regression with approximate sparsity.* For the linear regression problem $Y = X\beta + W$, assume that β is in an ℓ_q ball so that it has an approximate sparse pattern. Then $X\beta$ can be approximated by the structured linear model with an exact sparse pattern.

10. *Wavelet estimation under Besov space.* Consider the Gaussian sequence model $Y_{jk} = \theta_{jk} + n^{-1/2}W_{jk}$ for $k = 1,\dots,2^j$ and $j = 0,1,2,\dots$. The signal θ belongs to a Besov ball $\Theta^\alpha_{p,q}(L)$. Then we can use the general structured linear model to approximate the signal at each resolution separately. This strategy leads to a minimax optimal procedure for a large collection of Besov balls.

3. The prior distribution. In this section, we introduce a prior distribution on the structured linear model (1). 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 an elliptical Laplace distribution with density function proportion to $\exp(-\lambda\|\mathscr{X}_Z(Q)\|)$. By direct calculations of its normalizing constant, the density function is

$$f_{\ell(\mathcal{Z}_\tau),\mathscr{X}_Z,\lambda}(Q) = \frac{\sqrt{\det(\mathscr{X}_Z^T\mathscr{X}_Z)}}{2}\left(\frac{\lambda}{\sqrt{\pi}}\right)^{\ell(\mathcal{Z}_\tau)}\frac{\Gamma(\ell(\mathcal{Z}_\tau)/2)}{\Gamma(\ell(\mathcal{Z}_\tau))}\exp\bigl(-\lambda\|\mathscr{X}_Z(Q)\|\bigr). \tag{4}$$

A derivation of the normalizing constant with details is given in Section A of the supplement. Note that (4) is well defined when $\det(\mathscr{X}_Z^T\mathscr{X}_Z) > 0$. 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 [21] with scatter matrix proportional to $(\mathscr{X}_Z^T\mathscr{X}_Z)^{-1}$. Compared with a product measure on Q, the density function (4) 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

$$\epsilon(\mathcal{Z}_\tau) \geq \ell(\mathcal{Z}_\tau) + \log|\mathcal{Z}_\tau|. \tag{5}$$

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 $\pi(\tau) \propto \frac{\Gamma(\ell(\mathcal{Z}_\tau))}{\Gamma(\ell(\mathcal{Z}_\tau)/2)} \exp(-D\epsilon(\mathcal{Z}_\tau))$;
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 (5). For all the examples considered in the paper, $\epsilon(\mathcal{Z}_\tau)$ is chosen to be at the same order of the model complexity (2). The quantity $\frac{\Gamma(\ell(\mathcal{Z}_\tau))}{\Gamma(\ell(\mathcal{Z}_\tau)/2)}$ is a 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, $\exp(-D\epsilon(\mathcal{Z}_\tau))$ is the complexity prior used by [16, 17] in Gaussian sequence model and sparse linear regression. Since the support $\mathcal{T}$ is a finite set, π is 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 (4), and thus Step 3 is well defined.

After defining the prior, we need to specify the likelihood function. The examples in Section 2 have different distributions. For example, the 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 $Y|(Z, Q) \sim N(\mathscr{X}_Z(Q), I_N)$ throughout the paper. Then the posterior distribution is

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

The summation over an empty set is understood to be zero in the posterior formula above. We remark that there exists at least one $\tau \in \mathcal{T}$ such that $\bar{\mathcal{Z}}_\tau$ is not empty (see Theorem 4.1), and thus the posterior formula is well defined. We also note 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 data is generated from the same model. Instead, we allow data to be generated by an arbitrary signal with sub-Gaussian noise. That is,

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

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

$$\mathbb{P}\big(|\langle W, K\rangle| > t\big) \leq e^{-\rho t^2/2} \quad \text{for all } \|K\| = 1. \tag{6}$$

The sub-Gaussianity number $\rho > 0$ is assumed to be a constant throughout the paper. It worths noting that $1/\rho$ is a bound on the noise level. We also assume a mild condition on the function $\epsilon(\mathcal{Z}_\tau)$,

$$\big|\{\tau \in \mathcal{T} : t - 1 < \epsilon(\mathcal{Z}_\tau) \le t\}\big| \le t \quad \text{for all } t \in \mathbb{N}. \tag{7}$$

The condition (7) is satisfied for all examples considered in the paper. The main result of the paper is stated in the following theorem. Recall that λ and D are parameters of the prior distribution Π.

THEOREM 4.1. *Assume* (5), (6) *and* (7). *Given any* $\theta^* \in \mathbb{R}^N$, $\tau^* \in \mathcal{T}$, $Z^* \in \bar{\mathcal{Z}}_{\tau^*}$, $Q^* \in \mathbb{R}^{\ell(\mathcal{Z}_{\tau^*})}$, *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*

$$\begin{aligned}&\mathbb{E}_{\theta^*}\Pi\big(\epsilon(\mathcal{Z}_\tau) > (1+\delta_1)\epsilon(\mathcal{Z}_{\tau^*}) + \delta_1\big\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\big\|^2 | Y\big)\\ &\quad\le \exp\big(-C'\big(\epsilon(\mathcal{Z}_{\tau^*}) + \big\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\big\|^2\big)\big),\end{aligned} \tag{8}$$

$$\begin{aligned}&\mathbb{E}_{\theta^*}\Pi\big(\big\|\mathscr{X}_Z(Q) - \theta^*\big\|^2 > (1+\delta_2)\big\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\big\|^2 + M\epsilon(\mathcal{Z}_{\tau^*}) | Y\big)\\ &\quad\le \exp\big(-C''\big(\epsilon(\mathcal{Z}_{\tau^*}) + \big\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\big\|^2\big)\big)\end{aligned} \tag{9}$$

and

$$\begin{aligned}&\mathbb{E}_{\theta^*}\big\|\mathbb{E}_\Pi\big(\mathscr{X}_Z(Q)|Y\big) - \theta^*\big\|^2\\ &\quad\le (1+\delta_2)\big\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\big\|^2 + M\epsilon(\mathcal{Z}_{\tau^*})\\ &\qquad + \exp\big(-C'''\big(\epsilon(\mathcal{Z}_{\tau^*}) + \big\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\big\|^2\big)\big),\end{aligned} \tag{10}$$

for any constant $D > D_{\lambda,\delta,\rho}$ *with* $\delta_1 = \delta$, $\delta_2 = 8\sqrt{14\delta/\rho}$ *and some constants* M, C', C'', C''' *only depending on* λ, δ, ρ, D.

Theorem 4.1 contains three results of an oracle type. The object $\mathscr{X}_{Z^*}(Q^*)$ can be chosen with arbitrary Q^* and Z^*, but is usually taken as the oracle model that best approximates the true signal θ^* in many applications. The first result (8) 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 (9) 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 $\epsilon(\mathcal{Z}_{\tau^*})$. The third result is an oracle inequality for the posterior mean $\mathbb{E}_\Pi(\mathscr{X}_Z(Q)|Y)$. It is worth noting that $\exp(-C'''(\epsilon(\mathcal{Z}_{\tau^*}) + \|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2))$ is negligible compared with $(1+\delta_2)\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2 + M\epsilon(\mathcal{Z}_{\tau^*})$ in all the examples considered in the paper.

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* (5), (6) *and* (7). *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*

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

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

and

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

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

REMARK 4.1. The above results hold for all $\epsilon(\mathcal{Z}_\tau)$ satisfying (5). By choosing $\epsilon(\mathcal{Z}_\tau)$ at the same order of (2), we obtain the contraction rate $\ell(\mathcal{Z}_{\tau^*}) + \log|\mathcal{Z}_{\tau^*}|$ for the posterior distribution. As we are going to show in the next section, this rate is minimax optimal for all the examples considered in the paper. From now on, we refer to both (2) and $\epsilon(\mathcal{Z}_\tau)$ as the complexity function.

REMARK 4.2. By carefully examining the proof, the assumption (7) 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 condition (7) is simpler and is sufficient for all the examples considered in the paper. For example, $|\{k \in [n] : t - 1 < k^2 + n\log k \le t\}| \le 1$ for stochastic block model, and $|s \in [p] : t - 1 < 2s\log\frac{ep}{s} \le t| \le 1$ for sparse linear regression.

REMARK 4.3. It is worth noting that the constant $(1 + \delta_2)$ in (9) 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 may be impossible, which is suggested by a counterexample in [49] for sparse linear regression.

REMARK 4.4. Note that we do not impose any assumption on the operator $\mathcal{X}_Z(\cdot)$ besides its linearity (3). In the regression model, this means the results are assumption-free for the design matrix, as those in the frequentist literature [59].

5. Applications.

5.1. *Stochastic block model.* The stochastic block model was proposed by [30] 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-loop in the sense that $A_{ii} = 0$ for all $i \in [n]$, the stochastic block model assumes $\{A_{ij}\}_{i>j}$ are independent Bernoulli random variables with mean $\theta_{ij} = Q_{z(i)z(j)} \in [0, 1]$ for 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 some attention. The minimax rate of estimating θ under the Frobenius norm was established by [22]. However, the upper bound in [22] was achieved by a procedure assuming the knowledge of the true number of community k^*, which 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 $\theta_{ij} = [\mathcal{X}_Z(Q)]_{ij} = Q_{z(i)z(j)}$ for all $i \ne j$. 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 (5) is satisfied with $\epsilon(\mathcal{Z}_k) = k^2 + n\log k$. The general prior distribution Π can be specialized to this case as follows:

1. Sample $k \sim \pi$ from $[n]$, where $\pi(k) \propto \frac{\Gamma(k^2)}{\Gamma(k^2/2)}\exp(-D(k^2 + n\log k))$;
2. Conditioning on k, sample z uniformly from the set $\{z \in [k]^n : \min_{u\in[k]}|\{i \in [n] : z(i) = u\}| > 0\}$;

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, $\bar{\mathcal{Z}}_k = \{z \in [k]^n : \min_{u\in[k]} |\{i \in [n] : z(i) = u\}| > 0\}$. In other words, $\bar{\mathcal{Z}}_k$ is the set of label assignments that induce k clusters. For each $u \in [k]$, $|\{i \in [n] : z(i) = u\}|$ is the size of the uth cluster. If for some $u \in [k]$, $|\{i \in [n] : z(i) = u\}| = 0$, then there must exists some $k_1 < k$ such that $z \in \bar{\mathcal{Z}}_{k_1}$. Moreover, it is easy to see that for any $z \in \bar{\mathcal{Z}}_k$, $(Q_1)_{z(i)z(j)} = (Q_2)_{z(i)z(j)}$ for all $i \neq j$ implies $Q_1 = Q_2$. This indicates that the corresponding linear operator $\mathscr{X}_Z(\cdot)$ is not degenerate. To help understand the density function $f_{k,z,\lambda}$ in Step 3, consider the case of equal community sizes, that is, $|\{i \in [n] : z(i) = u\}| = n/k$ for all $u \in [k]$. Then $f_{k,z,\lambda}(Q) \propto e^{-\frac{n\lambda}{k}\|Q\|_{\mathrm{F}}}$, if we include the diagonal entries and treat θ_{ii} as $Q_{z(i)z(i)}$.

To study the posterior distribution, we 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 $i \in [n]$, where $z^* \in \bar{\mathcal{Z}}_{k^*}$ for some $k^* \in [n]$. It can be shown that the noise $W = A - \theta^*$ satisfies (6) for some constant $\rho > 0$ by Hoeffding's inequality, and the complexity function $\epsilon(\mathcal{Z}_\tau) = k^2 + n\log k$ satisfies (7). Hence, Corollary 4.1 can be specialized for the stochastic block model as follows.

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*

$$\mathbb{E}_{\theta^*}\Pi\big(k^2 + n\log k > (1+\delta)\big((k^*)^2 + n\log k^*\big)|A\big) \le \exp\big(-C'\big((k^*)^2 + n\log k^*\big)\big)$$

and

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

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

A previous result on Bayes estimation for the stochastic block model by [46] assumes the knowledge of k^*, and the rate is suboptimal. To the best of our knowledge, our result is the first adaptive Bayes estimator for the stochastic block model with a posterior contraction rate $(k^*)^2 + n\log k^*$, which is optimal according to [22]. 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. Corollary 4.1 also implies that the posterior mean achieves the minimax rate $(k^*)^2 + n\log k^*$.

While our result uses a prior distribution that does not impose symmetry on the mean matrix θ, it may be more desirable to incorporate symmetry from a practical point of view. This can be achieved within our framework of structured linear models. To be specific, we can consider the object $\mathscr{X}_Z(Q)$ to be a triangle array with entries $\{[\mathscr{X}_Z(Q)]_{ij} : 1 \le i < j \le n\}$. Then Q is also a triangle array, but it is of dimension $k(k+1)/2$ and has entries $\{Q_{ij} : 1 \le i \le j \le k\}$. The linear operator $\mathscr{X}_Z(\cdot)$ that maps from Q to $\mathscr{X}_Z(Q)$ is specified by $\mathscr{X}_Z(Q) = Q_{z(i)z(j)}$. In other words, the symmetric SBM is also a special case of our structured linear models with $Z = z$, $\tau = k$, $\mathcal{T} = [n]$, $\mathcal{Z}_k = [k]^n$, $\ell(\mathcal{Z}_k) = k(k+1)/2$ and $N = n(n-1)/2$.

To close this section, we also mention an important problem of community detection, which is equivalent to estimating the structure Z in our general framework. The posterior distribution of Bayesian community detection was recently analyzed by [56].

5.2. *Biclustering*. The biclustering model, originated in [28], can be viewed as a precursor and an asymmetric extension of the stochastic block model. The data matrix $Y \in \mathbb{R}^{n\times m}$ is generated by a signal matrix $\theta = (\theta_{ij})$ with $\theta_{ij} = Q_{z_1(i)z_2(j)}$ for some label vectors $z_1 \in [k]^n$ and $z_2 \in [l]^m$, that is, the rows of θ have k clusters and the columns of θ have l clusters, and the values of (θ_{ij}) that belong to the same row-cluster and the same column-cluster are identical. The goal is to recover the true signal matrix θ^* from the observation Y.

To put the biclustering model in our general framework, we have $Z = (z_1, z_2)$, $\tau = (k, l)$, $\mathcal{T} = [n] \times [m]$, $\mathcal{Z}_{k,l} = [k]^n \times [l]^m$ 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 (5) and (7). The general prior Π can be specialized to this case as follows:

1. Sample $(k, l) \sim \pi$ from $[n] \times [m]$, where $\pi(k, l) \propto \frac{\Gamma(kl)}{\Gamma(kl/2)} \exp(-D(kl + n\log k + m\log l))$;
2. Conditioning on (k, l), sample (z_1, z_2) uniformly from $\bar{\mathcal{Z}}_{k,l}$;
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,

$$\bar{\mathcal{Z}}_{k,l} = \Big\{(z_1, z_2) \in [k]^n \times [l]^m : \min_{u\in[k]} \big|\{i \in [n] : z_1(i) = u\}\big| > 0,$$
$$\min_{v\in[l]} \big|\{j \in [m] : z_2(j) = v\}\big| > 0\Big\}.$$

In other words, for any $(z_1, z_2) \in \bar{\mathcal{Z}}_{k,l}$, z_1 and z_2 induce row and column clustering structures with numbers of clusters being k and l, respectively.

To analyze the posterior distribution, assume $Y = \theta^* + W$, where $\theta^*_{ij} = Q^*_{z_1^*(i)z_2^*(j)}$ for $Q^* \in \mathbb{R}^{k^*\times l^*}$ and $(z_1^*, z_2^*) \in [k^*]^n \times [l^*]^m$, and the noise W is assumed to satisfy (6).

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*

$$\mathbb{E}_{\theta^*}\Pi\big(kl + n\log k + m\log l > (1+\delta)\big(k^*l^* + n\log k^* + m\log l^*\big)|Y\big)$$
$$\le \exp\big(-C'\big(k^*l^* + n\log k^* + m\log l^*\big)\big)$$

and

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

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 [22]. To the best of our knowledge, this is the first adaptive estimation result for biclustering with an 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 [16]. However, the results of [16] imposed a

stronger assumption that is used for the Lasso estimator [9]. In this section, we show that the prior distribution that we propose in Section 3 leads to optimal posterior contraction rates with minimal assumptions.

The sparse linear regression model is a special case of the general structured linear model (1) with $Z = S$, $\tau = s$, $\mathcal{T} = [p]$, $\mathcal{Z}_s = \{S \subset [p] : |S| = s\}$, $\ell(\mathcal{Z}_s) = s$ and $Q = \beta_S$. Then we have the representation $\mathcal{X}_Z(Q) = X_{*S}\beta_S = X\beta$. 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 (5). It can be shown that $\epsilon(\mathcal{Z}_\tau)$ satisfies (7). We specialize the general prior Π in Section 3 as follows:

1. Sample $s \sim \pi$ from $[p]$, where $\pi(s) \propto \frac{\Gamma(s)}{\Gamma(s/2)}\exp(-2Ds\log\frac{ep}{s})$;
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 $\beta_{S^c} = 0$.

In Step 1, we set $\epsilon(\mathcal{Z}_s) = 2s\log\frac{ep}{s}$ 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 $\tilde{\mathcal{Z}}_s = \{S \subset [p] : |S| = s, \det(X_{*S}^T X_{*S}) > 0\}$ instead of $\mathcal{Z}_s$ such that 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 have $\tilde{\mathcal{Z}}_s = \varnothing$. Note that the exponent on the density of β_S is $-\lambda\|X_{*S}\beta_S\|$, different from $-\lambda\|\beta_S\|_1$ in [16]. We allow the prior to depend on the design matrix X to obtain an assumption-free optimal posterior prediction rate. The idea of design-dependent prior was also employed by [42] in an empirical pseudo-Bayes framework. Since $e^{-\lambda\|X_{*S}\beta_S\|}$ has an exponential tail, it is capable of modeling a large regression coefficient. We expect that an elliptical distribution with heavier tails than Laplace also works here.

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, $\exp(-2Ds\log\frac{ep}{s})$ is the common prior distribution on the model dimension used in [16, 17, 24, 42, 48]. Since $\exp(-2Ds\log\frac{ep}{s})$ is a decreasing function of s, it gives less weights for more complex models. However, with the correction factor, $\pi(s) \propto \frac{\Gamma(s)}{\Gamma(s/2)}\exp(-2Ds\log\frac{ep}{s})$ is not necessarily a decreasing function of s. For a large $D > 0$, it can be shown that $\pi(\sqrt{p}) < \pi(p)$, which leads to a counterintuitive prior modeling strategy. Nevertheless, it is worth noting that the π in Step 1 is only part of the prior Π. The elliptical Laplace distribution used later also contributes to the prior modeling on the dimension. The combination of the two gives a correct prior weight on the model dimension.

Let $Y = X\beta^* + W$ for some β^* with support S^* and sparsity $|S^*| = s^*$, where the noise vector W is assumed to be sub-Gaussian (6). Without loss of generality, we may assume $S^* \in \tilde{\mathcal{Z}}_{s^*}$. 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^*, $S^* \in \tilde{\mathcal{Z}}_{s^*}$ 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*

$$\mathbb{E}_{X\beta^*}\Pi\big(s > (1+\delta)s^*|Y\big) \le \exp\left(-C's^*\log\frac{ep}{s^*}\right) \tag{11}$$

and

$$\mathbb{E}_{X\beta^*}\Pi\left(\|X\beta - X\beta^*\|^2 > Ms^*\log\frac{ep}{s^*}\middle|Y\right) \le \exp\left(-C''s^*\log\frac{ep}{s^*}\right) \tag{12}$$

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

The result (11) is a consequence of (8) since $s\log\frac{ep}{s} > (1+\delta_1)s^*\log\frac{ep}{s^*}$ is equivalent to $s > (1+\delta)s^*$. It improves the corresponding bounds in [16, 17] at a constant level. The result (12) achieves the rate $s^*\log\frac{ep}{s^*}$ with no assumption on the design matrix X, which is comparable to the frequentist result in [10]. A slight improvement of (12) will be discussed in Section 5.8.

Besides the optimal prediction rate, we are ready to obtain optimal estimation rates given (11) and (12). Define

$$\kappa_2 = \min_{\{b\neq 0:\|b\|_0\le(2+\delta)s^*\}} \frac{\|Xb\|}{\sqrt{n}\|b\|} \quad \text{and} \quad \kappa_1 = \min_{\{b\neq 0:\|b\|_0\le(2+\delta)s^*\}} \frac{\sqrt{s^*}\|Xb\|}{\sqrt{n}\|b\|_1}. \tag{13}$$

Note that κ_2 is the restricted eigenvalue constant [9, 14] and κ_1 is the compatibility constant [12].

COROLLARY 5.4. *Under the setting of Corollary 5.3, we have*

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

and

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

for the same constants M, C', C'' in Corollary 5.3.

We note that the dependence on the quantities κ_2 and κ_1 are optimal [47], compared with the Lasso estimator and the spike and slab prior [16]. When $\kappa \asymp \kappa_1 \asymp \kappa_2$, the rates of the Lasso estimator are $\frac{s^*\log p}{n\kappa^4}$ and $\frac{(s^*)^2\log p}{n\kappa^4}$ for the loss $\|\cdot\|^2$ [9] and the loss $\|\cdot\|_1^2$ [55], respectively, and the rates of the spike and slab prior are $\frac{s^*\log\frac{ep}{s^*}}{n\kappa^6}$ and $\frac{(s^*)^2\log\frac{ep}{s^*}}{n\kappa^8}$ for the loss $\|\cdot\|^2$ and $\|\cdot\|_1^2$ [16], respectively.

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

To close this section, we briefly discuss the computational issue of the proposed prior distribution. A recent theoretical result by [64] shows that the mixing-time of a simple MCMC algorithm is polynomial in the setting of Bayesian sparse linear regression. They also use a two-step model selection prior, but the distribution on model parameters is $e^{-\lambda\|X_{*S}\beta_S\|^2}$, compared with our $e^{-\lambda\|X_{*S}\beta_S\|}$. Given the similarity between the two prior distributions, it is conceivable that similar results in [64] can also be established in our setting. More interestingly, whether a general theory of computation can be established under our framework of structured linear models will be an important topic to study in the future.

5.4. *Multiple linear regression with group sparsity.* Let us consider a multiple regression setting 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. There is some $S \subset [p]$ such that $B_{S^c*} = 0$, that is, S is the nonzero rows of B. The concept of group sparsity was proposed by [3, 65], and frequentist statistical properties were analyzed by [38].

To put the problem in the general framework, we have $Z = S$, $\tau = s$, $\mathcal{T} = [p]$, $\mathcal{Z} = \{S \subset [p] : |S| = s\}$, $\ell(\mathcal{Z}_s) = ms$ and $Q = B_{S*}$. Then $\mathcal{X}_Z(Q) = X_{*S}B_{S*} = XB$. The choice $\epsilon(\mathcal{Z}_s) = s(m + \log\frac{ep}{s})$ satisfies the conditions (5) and (7). The prior distribution Π is similar to that used in Section 5.3:

1. Sample $s \sim \pi$ from $[p]$, where $\pi(s) \propto \frac{\Gamma(s)}{\Gamma(s/2)} \exp(-Ds(m + \log \frac{ep}{s}))$;
2. Conditioning on s, sample S uniformly from $\bar{\mathcal{Z}}_s = \{S \subset [p] : |S| = s, \det(X_{*S}^T X_{*S}) > 0\}$;
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 sample S from $\bar{\mathcal{Z}}_s$ in Step 2 as for sparse linear regression. Assume the data is generated by $Y = XB^* + W$ for some matrix B^* with support S^* and sparsity s^*. 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 (6).

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*

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

and

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

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 $s^*(m + \log \frac{ep}{s^*})$, which is optimal according to [38, 41]. Posterior contraction for various estimation loss functions can also be derived in a similar way as in Section 5.3.

5.5. *Multitask learning.* Multitask learning is another name for multiple linear regression in the form of XB with $X \in \mathbb{R}^{n \times p}$ and $B \in \mathbb{R}^{p \times m}$. Compared with 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 consider another special but important class of multi-task learning problems. 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 multitask learning problem in the general framework. 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 $[\mathcal{X}_Z(Q)]_{*j} = XQ_{*z(j)}$. The complexity function $\epsilon(\mathcal{Z}_\tau) = pk + m \log k$ satisfies the conditions (5) and (7). We consider a full rank design matrix with $\det(X^T X) > 0$. The general prior distribution Π in Section 3 can be specialized to this case:

1. Sample $k \sim \pi$ from $[p]$, where $\pi(k) \propto \frac{\Gamma(pk)}{\Gamma(pk/2)} \exp(-D(pk + m \log k))$;
2. Conditioning on k, sample z uniformly from the set $\{z \in [k]^m : \min_{u \in [k]} |\{j \in [m] : z(j) = u\}| > 0\}$;
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 have $\bar{\mathcal{Z}}_k = \{z \in [k]^m : \min_{u \in [k]} |\{j \in [m] : z(j) = u\}| > 0\}$, which is due to $\det(X^T X) > 0$. The full rankness of the design matrix implicitly implies $p \le n$. In fact,

there is no loss to assume $\det(X^TX) > 0$, because whenever $\det(X^TX) = 0$, one can simply use a subset of the variables that are linearly independent without affecting the prediction error.

We 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 (6).

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*

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

and

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

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 multitask learning is $pk^* + m\log k^*$. According to [35], the rate $pk^* + m\log k^*$ is minimax optimal when $pk^* + m\log k^* \le pm$. The minimax rate for the case $pk^* + m\log k^* > pm$ is simply pm, the dimension of B. In that case, even the ordinary least-squares estimator $\hat{B} = \operatorname{argmin}_B \|Y - XB\|_{\mathrm{F}}^2$ can achieve the rate.

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. The problem is also referred to as sparse coding [45]. Recently, the minimax rate of dictionary learning has been established by [35] 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 $\tau = (p,s)$, $\mathcal{T} = \{(p,s) \in [n\wedge d]\times[n] : s \le p\}$, $\mathcal{Z}_{p,s} = \{Z \in \{-1,0,1\}^{p\times d} : \max_{j\in[d]} |\operatorname{supp}(Z_{*j})| \le s\}$ and $\ell(\mathcal{Z}_{p,s}) = np$. Moreover, we have the representation $\mathscr{X}_Z(Q) = QZ$. The complexity function is $\ell(\mathcal{Z}_{p,s}) + \log|\mathcal{Z}_{p,s}| = np + d(\log\binom{p}{s} + 3\log s)$. With $\epsilon(\mathcal{Z}_{p,s}) = 3(np + ds\log\frac{ep}{s})$, (5) and (7) 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 $\pi(p,s) \propto \frac{\Gamma(np)}{\Gamma(np/2)}\exp(-3D(np + ds\log\frac{ep}{s}))$;
2. Given (p,s), sample Z uniformly from $\bar{\mathcal{Z}}_{p,s} = \{Z \in \mathcal{Z}_{p,s} : \det(ZZ^T) > 0\}$;
3. Given (p,s,Z), sample $Q \sim f_{p,s,Z,\lambda}$ with $f_{p,s,Z,\lambda}(Q) \propto e^{-\lambda\|QZ\|_{\mathrm{F}}}$;
4. Set $\theta = QZ$.

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

We assume that the data is generated by $Y = \theta^* + W$ for some noise matrix W satisfying (6) and $\theta^* = Q^*Z^*$. Without loss of generality, we assume the matrix Z^* belongs to the set $\bar{\mathcal{Z}}_{p^*,s^*}$. If $\det(Z^*(Z^*)^T) = 0$, 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*

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

and

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

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 when $np^* + ds^*\log\frac{ep^*}{s^*} \le nd$ according to [35]. When $np^* + ds^*\log\frac{ep^*}{s^*} > nd$, the minimax rate is just nd, the dimension of θ. It can be achieved by the naive estimator $\hat\theta = Y$, and thus this is not an interesting case to us. The term $ds^*\log\frac{ep^*}{s^*}$ in the rate is the error for recovering d sparse regression coefficient vectors, and np^* is the price to pay for not knowing the design matrix Q^*. 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 an 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

$$(\xi_1,\dots,\xi_n) \sim \mathbb{P}_\xi, \quad A_{ij}|(\xi_i,\xi_j) \sim \mathrm{Bernoulli}(\theta^*_{ij}), \quad \text{where } \theta^*_{ij} = f^*(\xi_i,\xi_j). \tag{14}$$

For $i \in [n]$, $A_{ii} = \theta^*_{ii} = 0$. Conditioning on $(\xi_1,\dots,\xi_n)$, $A_{ij} = A_{ji}$ 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 [19, 31, 39, 40] and the studies of exchangeable arrays [2, 33]. 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

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

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

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

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 [22] given the knowledge of α. In this section, we solve the adaptive graphon estimation problem via a Bayes procedure.

As shown in [22], it is sufficient to approximate a graphon with Hölder smoothness by a blockwise constant function. In the random graph setting, a blockwise constant function is the

stochastic block model. 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 $\alpha, L > 0$ defined above and any constant $\lambda > 0$, there exists some constant $D_\lambda > 0$ only depending on λ such that*

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

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 (14) over both $\{A_{ij}\}$ and $\{\xi_i\}$. Moreover, we do not need any assumption on the distribution on $\{\xi_i\}$, 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 for the worst-case design according to [22]. 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 an optimal convergence rate.

5.8. *Aggregation.* Aggregation in nonparametric regression has been considered by [18, 36, 43, 53, 60, 61] among others. Let us start with the nonparametric regression setting with fixed design. The data is generated by

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

where the noise vector $W = \{W_i\}$ is assumed to satisfy (6). The goal of nonparametric regression is to estimate the true regression function f^* by some estimator $\hat{f}$ under the loss

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

where $\|\cdot\|_n$ stands for the empirical ℓ_2 norm. Assume we are given a collection of functions $\{f_1, \ldots, f_p\}$, called the dictionary. 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 estimator $\hat{f}$ to satisfy the following oracle inequality:

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

with high probability for some arbitrarily small constant $\delta \in (0, 1)$ and some optimal rate function $\Delta_{n,p}(\Theta)$ determined by the class Θ. The right-hand side of (16) is also called the index of resolvability of f^* [8, 59]. Various types of aggregation problems include linear,

convex, model selection aggregation, etc., which are 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 $f_\beta = (f_\beta(x_1), \dots, f_\beta(x_n))$ can be represented as $X\beta$ with $X_{ij} = f_j(x_i)$ for all $(i, j) \in [n] \times [p]$, the aggregation problem can be recast as a linear regression problem. Define $r = \operatorname{rank}(X)$. Without loss of generality, we assume the first r columns of X span the column space of X, that is, $\operatorname{span}(\{X_{*j}\}_{j\in[r]}) = \operatorname{span}(\{X_{*j}\}_{j\in[p]})$. 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 $\pi(s) = \mathcal{N}\frac{\Gamma(s)}{\Gamma(s/2)}\exp(-Ds\log\frac{ep}{s})$ 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 [48, 49]. 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 5.9. *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*

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

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 5.9 has a better convergence rate than Corollary 5.3. The corresponding frequentist results [48, 49] have leading constant 1 instead of the $(1 + \delta)$ in Corollary 5.9. Since our prior and posterior have a subset selection step, the result in [49] suggests that the extra constant δ may be necessary.

Let us specialize Corollary 5.9 to various types of aggregation problems. Following the notation in [54], 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

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

COROLLARY 5.10. *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*

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

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

Corollary 5.10 provides a universal aggregation result with a single posterior distribution. The rate is minimax optimal according to [48, 58]. Bayes aggregation was recently studied by [63] under the model misspecification framework [34]. Corollary 5.10 is a stronger result of posterior oracle inequality under weaker assumptions compared with that of [63]. Other types of aggregation results such as ℓ_q aggregation can also be derived directly from Corollary 5.9.

5.9. *Linear regression under ℓ_q ball.* Section 5.3 studied high dimensional linear regression under exact sparsity. In this section, we assume that regression coefficients are approximately sparse. Theorem 4.1 allows us to derive optimal posterior rates of contraction via a bias variance tradeoff argument. 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 sub-Gaussian noise vector W satisfying (6). We assume β^* is approximately sparse,

$$\beta^* \in \mathcal{B}_q(k) = \left\{\beta\in\mathbb{R}^p : \sum_{j=1}^p |\beta_j|^q \le k\right\}$$

with some $q \in (0,1]$. For $q = 0$,

$$\mathcal{B}_0(k) = \left\{\beta\in\mathbb{R}^p : \sum_{j=1}^p \mathbb{I}\{\beta_j \ne 0\} \le k\right\},$$

which is reduced to the case of exact sparsity. To facilitate the presentation, we define the effective sparsity by $s^* = \lceil x^*\rceil$, where

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

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

COROLLARY 5.11. *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*

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

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 5.3. In the special case when $k \le p^{1-\eta}(\frac{\log p}{n})^{q/2}$ for some constant $\eta \in (0,1)$, the rate has an explicit formula in terms of k, which is $\frac{s^*\log(ep/s^*)}{n} \asymp k(\frac{\log p}{n})^{1-q/2}$. When X is an identity matrix, Corollary 5.11 reduces to the results for sparse Gaussian sequence model in [17]. We also refer to [58] for a general study of minimax estimation and aggregation under ℓ_q sparsity. Bayesian ℓ_q aggregation is also possible by applying the more general Corollary 5.9. Besides the prediction error, estimation error under approximate sparsity can be derived in the same way as Corollary 5.4. Finally, we remark that in practice, the assumption $\max_{j\in[p]} n^{-1/2}\|X_{*j}\| \le L$ can be met by column normalization of the design matrix.

5.10. *Wavelet estimation in 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

$$Y_{jk} = \theta^*_{jk} + \frac{1}{\sqrt{n}}W_{jk}, \quad k = 1,\dots,2^j;\ j = 0,1,2,\dots, \tag{17}$$

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 [32], and it is closely related to nonparametric regression and density estimation [11, 44]. 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

$$\Theta^\alpha_{p,q}(L) = \left\{\theta : \sum_j 2^{ajq}\|\theta_{j*}\|_p^q \le L^q\right\} \tag{18}$$

for some $p, q, \alpha, L > 0$ and $a = \alpha + \frac{1}{2} - \frac{1}{p}$. The Besov ball (18) naturally induces a multiresolution structure of the signal. This inspires us 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 $\pi(s_j) \propto \frac{\Gamma(s_j)}{\Gamma(s_j/2)}\exp(-Ds_j\log\frac{e2^j}{s_j})$;
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 $j \ge \log_2 n$, 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 5.12. *For any constants p, q, α satisfying $0 < p, q \le \infty$, $L > 0$ and $\alpha \ge \frac{1}{p}$ and any constant $\lambda > 0$, there exists some constant D_λ only depending on λ such that*

$$\sup_{\theta^* \in \Theta^\alpha_{p,q}(L)} \mathbb{E}_{\theta^*} \Pi\big(\|\theta - \theta^*\|^2 > M n^{-\frac{2\alpha}{2\alpha+1}} | Y\big) \le \exp\big(-C' n^{\frac{1}{2\alpha+1}} / \log n\big)$$

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

The result of Corollary 5.12 can be regarded as a Bayes version of Theorem 12.1 of [32] 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 (17) over the class $\Theta^\alpha_{p,q}(L)$ has been investigated by [25, 29, 50, 57] only for a restricted configuration of (p, q, α). In comparison, Corollary 5.12 obtains adaptive optimal posterior contraction rates to all possible combinations of (p, q, α) considered in the frequentist literature [32].

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 (17) into the sequence model with single index. That is, consider data generated by

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

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

$$\mathcal{S}_\alpha(L) = \Big\{\theta : \sum_j a_j^2 \theta_j^2 \le L^2\Big\},$$

for some sequence $a_j \asymp j^\alpha$. 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 $\theta_{[k]} = (\theta_1, \ldots, \theta_k) \sim f_{k,\lambda}$ 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, since $\mathcal{Z}_k = \{[k]\}$ is a singleton set and we do not need to perform a further model selection. Specializing Theorem 4.1 to this case, we obtain the following result.

COROLLARY 5.13. *For any constants $\alpha, L > 0$ and any constant $\lambda > 0$, there exists some constant D_λ only depending on λ such that*

$$\sup_{\theta^* \in \mathcal{S}_\alpha(L)} \mathbb{E}_{\theta^*} \Pi\big(\|\theta - \theta^*\|^2 > M n^{-\frac{2\alpha}{2\alpha+1}} | Y\big) \le \exp\big(-C' n^{\frac{1}{2\alpha+1}}\big)$$

for any $D > D_\lambda$ 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.

To close this section, we remark that the prior distributions used in this section depend on n. This is a consequence of writing the Gaussian sequence model in the form of structured linear models. For adaptive priors that do not depend on n but still achieve optimal posterior contraction rates, we refer the readers to [25].

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 [20, 37], we assume the mutual coherence condition:

$$n^{-1}X_{*j}^T X_{*j} = 1 \quad \text{for all } j \in [p] \quad \text{and} \quad \max_{j\neq k} n^{-1}X_{*j}^T X_{*k} \leq \tau. \tag{19}$$

Assuming that data is generated by $Y = X\beta^* + W$ for some regression coefficient β^* with sparsity s^* and some noise vector W satisfying (6), the posterior contraction under the ℓ_∞ norm for 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*

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

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 [13, 16, 20, 37]. It implies the restricted eigenvalue κ_2 defined in (13) to be bounded away from 0 [66]. 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 model selection consistency 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 not stated in literature. We provide a Bayes solution to this problem. For simplicity of presentation, we consider the case of identity design $Y = B + W \in \mathbb{R}^{p\times m}$. 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 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

$$\pi(T) \propto \frac{\Gamma(|T|)}{\Gamma(|T|/2)} \exp\Big(-D\Big(m|r(T)| + |r(T)|\log\frac{ep}{|r(T)|} + |T|\log\frac{em|r(T)|}{|T|}\Big)\Big); \tag{20}$$

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 with 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 (20) can be viewed as the product of $e^{-D|S|(m+\log\frac{ep}{|S|})}$ and $e^{-D|T|\log\frac{em|S|}{|T|}}$ with $S = r(T)$ denoting the row support. Therefore, (20) can be understood as first picking a row support S, and then further selecting 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 in the form of $S \times [m]$ for some S, (20) can give an arbitrary support T, which is critical to obtain optimal convergence rate under the ℓ_∞ loss. Assume that the data is generated from $Y = B^* + W$ for some B^* with row support S^* and noise matrix W satisfying (6). 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*

$$\mathbb{E}_{B^*}\Pi\big(|r(T)| > (1+\delta)s^*|Y\big) \le \exp\bigg(-C's^*\bigg(m + \log\frac{ep}{s^*}\bigg)\bigg), \tag{21}$$

$$\mathbb{E}_{B^*}\Pi\bigg(\big\|B - B^*\big\|_{\mathrm{F}}^2 > Ms^*\bigg(m + \log\frac{ep}{s^*}\bigg)\bigg|Y\bigg) \le \exp\bigg(-C''s^*\bigg(m + \log\frac{ep}{s^*}\bigg)\bigg) \tag{22}$$

and

$$\mathbb{E}_{B^*}\Pi\big(\big\|B - B^*\big\|_{\infty} > M\sqrt{\log(p+m)}|Y\big) \le (pm)^{-C'''} \tag{23}$$

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 in a group sparse signal recovery problem. The $e^{-D|S|(m+\log\frac{ep}{|S|})}$ part in (20) preserves the group sparse structure and results in the optimal ℓ_2 result (22). The $e^{-D|T|\log\frac{em|S|}{|T|}}$ part in (20) does a further model selection in a finer resolution, thus giving optimal rate for each coordinate in (23). 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 would be sufficient for the task.

7. Proof of Theorem 4.1. Let us first introduce some notation and give the outline of the proof. Define the following two sets:

$$\mathcal{A}(t) = \big\{\epsilon(\mathcal{Z}_\tau) > (1+\delta_1)\epsilon(\mathcal{Z}_{\tau^*}) + \delta_1\big\|\mathscr{X}_{\mathcal{Z}^*}(Q^*) - \theta^*\big\|^2 + ct\big\},$$

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

We will specify the numbers δ_1, δ_2, c later. The goal of the proof is to derive bounds for both $\mathbb{E}\Pi(\tau \in \mathcal{A}(t)|Y)$ and $\mathbb{E}\Pi(\mathscr{X}_Z(Q) \in U(t)|Y)$ with any $t \ge 0$. Then the conclusions (8) and (9) are deduced by setting $t = 0$. The conclusion (10) is then obtained by integrating out the tail bound of $\mathbb{E}\Pi(\mathscr{X}_Z(Q) \in U(t)|Y)$ over $t \ge 0$.

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

$$\Pi\big(\mathscr{X}_Z(Q) \in U(t)|Y\big) = \frac{\sum_{\tau\in\mathcal{T}}\exp(-D\epsilon(\mathcal{Z}_\tau))\frac{1}{|\bar{\mathcal{Z}}_\tau|}\sum_{Z\in\bar{\mathcal{Z}}_\tau}R(Z, U(t))}{\sum_{\tau\in\mathcal{T}}\exp(-D\epsilon(\mathcal{Z}_\tau))\frac{1}{|\bar{\mathcal{Z}}_\tau|}\sum_{Z\in\bar{\mathcal{Z}}_\tau}R(Z)}, \tag{24}$$

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

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

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

$$\Pi\big(\tau \in \mathcal{A}(t)|Y\big) = \frac{\sum_{\tau\in\mathcal{A}(t)} \exp(-D\epsilon(\mathcal{Z}_\tau)) \frac{1}{|\bar{\mathcal{Z}}_\tau|} \sum_{Z\in\bar{\mathcal{Z}}_\tau} R(Z)}{\sum_{\tau\in\mathcal{T}} \exp(-D\epsilon(\mathcal{Z}_\tau)) \frac{1}{|\bar{\mathcal{Z}}_\tau|} \sum_{Z\in\bar{\mathcal{Z}}_\tau} R(Z)}. \tag{25}$$

We are going to work with the formulas (25) and (24) to prove (8) and (9), respectively. The main strategy is to lower bound $R(Z^*)$ in the denominator and upper bound $R(Z)$ or $R(Z, U(t))$ in the numerator given some events holding with high probability. For each $Z \in \bar{\mathcal{Z}}_\tau$ and $t \geq 0$, consider the following events:

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

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 $C_1 > 1$, $C_2 > 0$ and $t \geq 0$, the conditions (5) and (6) imply*

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

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

LEMMA 7.2. *For any $\beta \geq 2$ and $\alpha \geq 1$, the condition (7) implies*

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

The proofs of Lemma 7.1 and Lemma 7.2 are given in Section D of the supplement [23].

Lower bounding $R(Z^*)$. We first introduce some extra notation. For the matrix $\mathscr{X}_{Z^*} \in \mathbb{R}^{N\times\ell(\mathcal{Z}_{\tau^*})}$, its singular value decomposition is $\mathscr{X}_{Z^*} = \mathcal{U}\Lambda\mathcal{V}^T$, with $\mathcal{U} \in \mathbb{R}^{N\times\ell(\mathcal{Z}_{\tau^*})}$ and $\mathcal{V} \in \mathbb{R}^{\ell(\mathcal{Z}_{\tau^*})\times\ell(\mathcal{Z}_{\tau^*})}$ being orthonormal matrices, and Λ is an $\ell(\mathcal{Z}_{\tau^*}) \times \ell(\mathcal{Z}_{\tau^*})$ diagonal matrix with positive entries on the diagonal.

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

$$\begin{aligned} &\left(\frac{\sqrt{\pi}}{\lambda}\right)^{\ell(\mathcal{Z}_{\tau^*})} R(Z^*) \\ &\quad= \sqrt{\det\big(\mathscr{X}_{Z^*}^T\mathscr{X}_{Z^*}\big)} \\ &\qquad\times \int e^{-\frac{1}{2}\|\mathscr{X}_{Z^*}(Q) - \mathscr{X}_{Z^*}(Q^*)\|^2 + \langle Y - \mathscr{X}_{Z^*}(Q^*), \mathscr{X}_{Z^*}(Q) - \mathscr{X}_{Z^*}(Q^*)\rangle - \lambda\|\mathscr{X}_{Z^*}(Q)\|}\, dQ \end{aligned}$$

$$= \sqrt{\det\big(\mathscr{X}_{Z^*}^T \mathscr{X}_{Z^*}\big)}$$

$$\times \int e^{-\frac{1}{2}\|\mathscr{X}_{Z^*}(Q)\|^2 + \langle Y - \mathscr{X}_{Z^*}(Q^*), \mathscr{X}_{Z^*}(Q)\rangle - \lambda\|\mathscr{X}_{Z^*}(Q) + \mathscr{X}_{Z^*}(Q^*)\|} \, dQ \tag{26}$$

$$\geq e^{-\lambda\|\mathscr{X}_{Z^*}(Q^*)\|}\sqrt{\det\big(\mathscr{X}_{Z^*}^T \mathscr{X}_{Z^*}\big)}$$

$$\times \int e^{-\frac{1}{2}\|\mathscr{X}_{Z^*}(Q)\|^2 + \langle Y - \mathscr{X}_{Z^*}(Q^*), \mathscr{X}_{Z^*}(Q)\rangle - \lambda\|\mathscr{X}_{Z^*}(Q)\|} \, dQ \tag{27}$$

$$= e^{-\lambda\|\mathscr{X}_{Z^*}(Q^*)\|}\sqrt{\det\big(\mathcal{V}\Lambda^2\mathcal{V}^T\big)}$$

$$\times \int e^{-\frac{1}{2}\|\Lambda\mathcal{V}^T Q\|^2 + \langle \mathcal{U}^T(Y - \mathscr{X}_{Z^*}(Q^*)), \Lambda\mathcal{V}^T Q\rangle - \lambda\|\Lambda\mathcal{V}^T Q\|} \, dQ \tag{28}$$

$$= e^{-\lambda\|\mathscr{X}_{Z^*}(Q^*)\|} \int e^{-\frac{1}{2}\|b\|^2 + \langle \mathcal{U}^T(Y - \mathscr{X}_{Z^*}(Q^*)), b\rangle - \lambda\|b\|} \, db \tag{29}$$

$$\geq e^{-\lambda\|\mathscr{X}_{Z^*}(Q^*)\|} \int e^{-\frac{1}{2}\|b\|^2 - \lambda\|b\|} \, db$$

$$\times \exp\bigg(\int \langle \mathcal{U}^T(Y - \mathscr{X}_{Z^*}(Q^*)), b\rangle \frac{e^{-\frac{1}{2}\|b\|^2 - \lambda\|b\|}}{\int e^{-\frac{1}{2}\|b\|^2 - \lambda\|b\|} \, db} \, db\bigg) \tag{30}$$

$$= e^{-\lambda\|\mathscr{X}_{Z^*}(Q^*)\|} \int e^{-\frac{1}{2}\|b\|^2 - \lambda\|b\|} \, db. \tag{31}$$

The equalities (26) and (29) are due to changes of variables and the linearity (3), and we use the orthonormal property of $\mathcal{U}$ to get $\det(\mathscr{X}_{Z^*}^T \mathscr{X}_{Z^*}) = \det(\mathcal{V}\Lambda^2\mathcal{V}^T)$ and $\|\mathscr{X}_{Z^*}(Q)\| = \|\Lambda\mathcal{V}^T Q\|$. We use triangle inequality and Jensen's inequality to derive (27) and (30), respectively. The last equality (31) 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. Let us continue to 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 (31), we reach the conclusion

$$R(Z^*) \geq e^{-\lambda\|\mathscr{X}_{Z^*}(Q^*)\|} \exp\bigg(-\frac{1}{2}\ell(\mathcal{Z}_{\tau^*}) - \lambda\sqrt{\ell(\mathcal{Z}_{\tau^*})} + \ell(\mathcal{Z}_{\tau^*})\log\lambda - \log\ell(\mathcal{Z}_{\tau^*})\bigg)$$

$$\geq e^{-\lambda\|\mathscr{X}_{Z^*}(Q^*)\|} \exp\big(-\ell(\mathcal{Z}_{\tau^*}) - \lambda\sqrt{\ell(\mathcal{Z}_{\tau^*})} + \ell(\mathcal{Z}_{\tau^*})\log\lambda\big) \tag{32}$$

$$\geq e^{-\lambda\|\mathscr{X}_{Z^*}(Q^*)\| - (1+\lambda+\lambda^{-1})\ell(\mathcal{Z}_{\tau^*})}. \tag{33}$$

The inequality (32) is by $-\log\ell(\mathcal{Z}_{\tau^*}) \geq -\frac{1}{2}\ell(\mathcal{Z}_{\tau^*})$ given the fact that $\ell(\mathcal{Z}_{\tau^*})$ is an integer. To obtain (33), we discuss two cases. When $\lambda \geq 1$,

$$-\lambda\sqrt{\ell(\mathcal{Z}_{\tau^*})} + \ell(\mathcal{Z}_{\tau^*})\log\lambda \geq -\lambda\sqrt{\ell(\mathcal{Z}_{\tau^*})} \geq -\lambda\ell(\mathcal{Z}_{\tau^*}) \geq -\big(\lambda + \lambda^{-1}\big)\ell(\mathcal{Z}_{\tau^*}).$$

When $\lambda < 1$,

$$\begin{aligned}-\lambda\sqrt{\ell(\mathcal{Z}_{\tau^*})} + \ell(\mathcal{Z}_{\tau^*})\log\lambda &\geq -\ell(\mathcal{Z}_{\tau^*}) + \ell(\mathcal{Z}_{\tau^*})\log\lambda \\ &\geq -\lambda^{-1}\ell(\mathcal{Z}_{\tau^*}) \geq -\big(\lambda + \lambda^{-1}\big)\ell(\mathcal{Z}_{\tau^*}).\end{aligned}$$

Note that (33) is a deterministic lower bound for the denominator $R(Z^*)$. The arguments we have used to derive (33) are greatly inspired by the corresponding ones in [16, 17].

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

$$\bar{Q}_Z = \operatorname*{argmin}_{Q\in\mathbb{R}^{\ell(\mathcal{Z}_\tau)}} \big\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\big\|^2. \tag{34}$$

The property of least squares implies the following Pythagorean identity:

$$\big\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\big\|^2 = \big\|\mathscr{X}_Z(Q) - \mathscr{X}_Z(\bar{Q}_Z)\big\|^2 + \big\|\mathscr{X}_Z(\bar{Q}_Z) - \mathscr{X}_{Z^*}(Q^*)\big\|^2. \tag{35}$$

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

$$\begin{aligned}&-\frac{1}{2}\big\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\big\|^2 + \big\langle Y - \mathscr{X}_{Z^*}(Q^*), \mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\big\rangle - \lambda\big\|\mathscr{X}_Z(Q)\big\| \\ &= -\frac{1}{2}\big\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\big\|^2 + \big\langle W, \mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\big\rangle \\ &\quad + \big\langle \theta^* - \mathscr{X}_{Z^*}(Q^*), \mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\big\rangle - \lambda\big\|\mathscr{X}_Z(Q)\big\| \\ &\leq -\frac{1}{2}\big\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\big\|^2 + \big(\sqrt{\epsilon^*(\mathcal{Z}_\tau) + t} + \lambda\big)\big\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\big\| \qquad (36) \\ &\quad + \big\|\theta^* - \mathscr{X}_{Z^*}(Q^*)\big\|\big\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\big\| \\ &\quad - \lambda\big\|\mathscr{X}_Z(Q)\big\| - \lambda\big\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\big\| \\ &\leq 2\big(\sqrt{\epsilon^*(\mathcal{Z}_\tau) + t} + \lambda\big)^2 - \left(\frac{1}{2} - \frac{1}{8}\right)\big\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\big\|^2 \qquad (37) \\ &\quad + 2\big\|\theta^* - \mathscr{X}_{Z^*}(Q^*)\big\|^2 + \frac{1}{8}\big\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\big\|^2 - \lambda\big\|\mathscr{X}_{Z^*}(Q^*)\big\| \\ &\leq (4 + 2/C_2)\epsilon^*(\mathcal{Z}_\tau) + 4t + 4\lambda^2 - \frac{1}{4}\big\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\big\|^2 - \lambda\big\|\mathscr{X}_{Z^*}(Q^*)\big\| \qquad (38) \\ &\leq (4 + 2/C_2)\epsilon^*(\mathcal{Z}_\tau) + 4t + 4\lambda^2 - \frac{1}{4}\big\|\mathscr{X}_Z(Q) - \mathscr{X}_Z(\bar{Q}_Z)\big\|^2 - \lambda\big\|\mathscr{X}_{Z^*}(Q^*)\big\|. \qquad (39)\end{aligned}$$

We have used Cauchy–Schwarz inequality and the event $E_Z(t)$ to get (36). The inequality (37) 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 obtain (38). Finally, the inequality (39) is due to the identity (35). The above upper bound implies

$$\begin{aligned}R(Z)\mathbb{I}_{E_Z(t)} &\leq \left(\frac{\lambda}{\sqrt{\pi}}\right)^{\ell(\mathcal{Z}_\tau)} e^{(4+2/C_2)\epsilon^*(\mathcal{Z}_\tau) + 4t + 4\lambda^2 - \lambda\|\mathscr{X}_{Z^*}(Q^*)\|} \\ &\quad \times \sqrt{\det\big(\mathscr{X}_Z^T\mathscr{X}_Z\big)}\int e^{-\frac{1}{4}\|\mathscr{X}_Z(Q) - \mathscr{X}_Z(\bar{Q}_Z)\|^2}\,dQ \\ &= \left(\frac{\lambda}{\sqrt{\pi}}\right)^{\ell(\mathcal{Z}_\tau)} e^{(4+2/C_2)\epsilon^*(\mathcal{Z}_\tau) + 4t + 4\lambda^2 - \lambda\|\mathscr{X}_{Z^*}(Q^*)\|}\int e^{-\frac{1}{4}\|b\|^2}\,db \\ &= (2\lambda)^{\ell(\mathcal{Z}_\tau)} e^{(4+2/C_2)\epsilon^*(\mathcal{Z}_\tau) + 4t + 4\lambda^2 - \lambda\|\mathscr{X}_{Z^*}(Q^*)\|}.\end{aligned} \tag{40}$$

The change of variable in (40) uses the same argument in (26) and (29). Using the fact that $\ell(\mathcal{Z}_\tau) \le \epsilon^*(\mathcal{Z}_\tau)$ by (5), we reach the conclusion

$$R(Z)\mathbb{I}_{E_Z(t)} \le e^{(4+2/C_2+|\log(2\lambda)|)\epsilon^*(\mathcal{Z}_\tau)+4t+4\lambda^2-\lambda\|\mathscr{X}_{Z^*}(Q^*)\|}. \tag{41}$$

Upper bounding $R(Z, U(t))\mathbb{I}_{F_Z(t)}$. We require $\delta_2 \in (0, 1/4)$ throughout the proof. Let $\xi \in (0, 1/4)$ be a constant to be specified later. When both $F_Z(t)$ and $U(t)$ hold, the exponent in the definition of $R(Z, U(t))$ 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 \\
&\quad + \langle \theta^* - \mathscr{X}_{Z^*}(Q^*), \mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\rangle \\
&\quad - \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^*}) + t} + \lambda)\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\| \qquad (42) \\
&\quad + \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 \\
&\quad - \lambda\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\| - \lambda\|\mathscr{X}_Z(Q)\| \\
&\le \xi^{-1}(\sqrt{\epsilon^*(\mathcal{Z}_{\tau^*}) + t} + \lambda)^2 - \frac{1}{4}\xi\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\|^2 \qquad (43) \\
&\quad - \frac{1}{2}(1-\xi)\|\mathscr{X}_Z(Q) - \theta^*\|^2 + \frac{1}{2}(1+\xi)\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2 \\
&\quad + \xi\langle \mathscr{X}_Z(Q) - \theta^*, \theta^* - \mathscr{X}_{Z^*}(Q^*)\rangle \\
&\quad - \lambda\|\mathscr{X}_{Z^*}(Q^*)\| \\
&\le \xi^{-1}(\sqrt{\epsilon^*(\mathcal{Z}_{\tau^*}) + t} + \lambda)^2 - \frac{1}{4}\xi\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\|^2 - \lambda\|\mathscr{X}_{Z^*}(Q^*)\| \qquad (44) \\
&\quad - \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 16\delta_2^{-1}\lambda^2 - \frac{1}{8}M\epsilon(\mathcal{Z}_{\tau^*}) - \frac{1}{8}Mt - \frac{1}{16}\delta_2\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2 \qquad (45) \\
&\quad - \frac{1}{32}\delta_2\|\mathscr{X}_Z(Q) - \mathscr{X}_Z(\bar{Q}_Z)\|^2 - \lambda\|\mathscr{X}_{Z^*}(Q^*)\|.
\end{aligned}$$

We have used the event F_Z to get (42). Now we explain the inequality (43). Due to the fact that $ab \le \xi^{-1}a^2 + \xi b^2/4$, we have

$$\begin{aligned}
&-\frac{1}{2}\xi\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\|^2 + (\sqrt{\epsilon^*(\mathcal{Z}_{\tau^*}) + t} + \lambda)\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\| \\
&\le \xi^{-1}(\sqrt{\epsilon^*(\mathcal{Z}_{\tau^*}) + t} + \lambda)^2 - \frac{1}{4}\xi\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\|^2.
\end{aligned}$$

It is easy to check the following equality:

$$\begin{aligned}&\langle \theta^* - \mathscr{X}_{Z^*}(Q^*), \mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\rangle - \frac{1}{2}(1-\xi)\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\|^2 \\ &\quad = -\frac{1}{2}(1-\xi)\|\mathscr{X}_Z(Q) - \theta^*\|^2 + \frac{1}{2}(1+\xi)\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2 \\ &\qquad + \xi\langle \mathscr{X}_Z(Q) - \theta^*, \theta^* - \mathscr{X}_{Z^*}(Q^*)\rangle.\end{aligned}$$

Finally, by triangle inequality, we get

$$-\lambda\|\mathscr{X}_Z(Q) - \mathscr{X}_{Z^*}(Q^*)\| - \lambda\|\mathscr{X}_Z(Q)\| \le -\lambda\|\mathscr{X}_{Z^*}(Q^*)\|.$$

Then, (44) is by rearranging (43) together with the inequality

$$\langle \mathscr{X}_Z(Q) - \theta^*, \theta^* - \mathscr{X}_{Z^*}(Q^*)\rangle \le \frac{1}{2}\|\mathscr{X}_Z(Q) - \theta^*\|^2 + \frac{1}{2}\|\theta^* - \mathscr{X}_{Z^*}(Q^*)\|^2.$$

Finally, we have set

$$\xi = \frac{1}{8}\delta_2 \quad \text{and} \quad C_2 = \frac{1}{128}\delta_2^2 \tag{46}$$

and used (35) to obtain (45) on the event $U(t)$ for all $M > \max\{128\delta_2^{-1}C_1, 128\delta_2^{-1}\}$. Note that we require $\delta_2 \in (0, 1/4)$ for the inequality (45). Using the above bound, we have

$$\begin{aligned}R(Z, U(t))\mathbb{I}_{F_Z(t)} &\le \left(\frac{\lambda}{\sqrt{\pi}}\right)^{\ell(\mathcal{Z}_\tau)} e^{-\lambda\|\mathscr{X}_{Z^*}(Q^*)\| + 16\delta_2^{-1}\lambda^2 - \frac{1}{8}M\epsilon(\mathcal{Z}_{\tau^*}) - \frac{1}{8}Mt - \frac{1}{16}\delta_2\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2} \\ &\quad \times \sqrt{\det(\mathscr{X}_Z^T\mathscr{X}_Z)}\int e^{-\frac{1}{32}\delta_2\|\mathscr{X}_Z(Q) - \mathscr{X}_Z(\bar{Q}_Z)\|^2}\,dQ \\ &= \left(\frac{4\lambda}{\sqrt{\delta_2/2}}\right)^{\ell(\mathcal{Z}_\tau)} e^{-\lambda\|\mathscr{X}_{Z^*}(Q^*)\| + 16\delta_2^{-1}\lambda^2 - \frac{1}{8}M\epsilon(\mathcal{Z}_{\tau^*}) - \frac{1}{8}Mt - \frac{1}{16}\delta_2\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2}\end{aligned}$$

by the same argument in deriving (41). By $\ell(\mathcal{Z}_\tau) \le \epsilon^*(\mathcal{Z}_\tau)$ from (5), we reach the conclusion

$$R(Z, U(t))\mathbb{I}_{F_Z(t)} \le e^{-\lambda\|\mathscr{X}_{Z^*}(Q^*)\| - \frac{1}{16}M\epsilon(\mathcal{Z}_{\tau^*}) - \frac{1}{8}Mt - \frac{1}{16}\delta_2\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2}, \tag{47}$$

for all $M > \max\{128\delta_2^{-1}(C_1+1), 16\log(4\lambda/\sqrt{\delta_2/2}) + 256\delta_2^{-1}\lambda^2\}$.

After obtaining the bounds (33), (41) and (47), we are ready to prove the main results.

PROOF OF (8). First, we use (33) and (41) to bound the ratio $R(Z)\mathbb{I}_{E_Z(t)}/R(Z^*)$,

$$\begin{aligned}&|\bar{\mathcal{Z}}_{\tau^*}|\frac{R(Z)\mathbb{I}_{E_Z(t)}}{R(Z^*)} \\ &\quad \le e^{4\lambda^2}|\mathcal{Z}_{\tau^*}|\frac{e^{[4C_1 + 2C_1/C_2 + C_1|\log(2\lambda)|](\epsilon(\mathcal{Z}_\tau)) + [4C_2 + 2 + C_2|\log(2\lambda)|]\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2 + 4t}}{e^{-(1+\lambda+\lambda^{-1})\ell(\mathcal{Z}_{\tau^*})}} \\ &\quad \le e^{4\lambda^2}\exp\big((1+\lambda+\lambda^{-1})\epsilon(\mathcal{Z}_{\tau^*}) + C_1'\epsilon(\mathcal{Z}_\tau) + C_2'\|\mathscr{X}_{Z^*}(Q^*) - \theta^*\|^2 + 4t\big),\end{aligned}$$

where $C_1' = 4C_1 + 2C_1/C_2 + C_1|\log(2\lambda)|$ and $C_2' = 4C_2 + 2 + C_2|\log(2\lambda)|$. Consider (25) with $\mathcal{A}(t)$. Here, we require that $\delta_1 \in (0, 1/3)$. By $Z^* \in \bar{\mathcal{Z}}_{\tau^*}$, we have

$$\mathbb{E}\Pi(\tau \in \mathcal{A}(t)|Y) \le \sum_{\tau\in\mathcal{A}(t)} \frac{\exp(-D\epsilon(\mathcal{Z}_\tau))}{\exp(-D\epsilon(\mathcal{Z}_{\tau^*}))}\frac{|\bar{\mathcal{Z}}_{\tau^*}|}{|\bar{\mathcal{Z}}_\tau|}\sum_{Z\in\bar{\mathcal{Z}}_\tau}\mathbb{E}\frac{R(Z)\mathbb{I}_{E_Z(t)}}{R(Z^*)} \tag{48}$$

$$+ \sum_{\tau\in\mathcal{A}(t)}\sum_{Z\in\bar{\mathcal{Z}}_\tau}\mathbb{P}(E_Z(t)^c). \tag{49}$$

According to previous calculations, (48) can be bounded by

$$\begin{aligned}&\exp\bigl(4\lambda^2+(D+\lambda+\lambda^{-1}+1)\epsilon(\mathcal{Z}_{\tau^*})+C_2'\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2+4t\bigr)\\&\quad\times\sum_{\tau\in\mathcal{A}(t)}\exp\bigl(-(D-C_1')\epsilon(\mathcal{Z}_\tau)\bigr).\end{aligned}\tag{50}$$

Then we can bound the sum in the above display by Lemma 7.2. We take $\alpha=(1+\delta_1)\epsilon(\mathcal{Z}_{\tau^*})+\delta_1\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2+ct$ and $\beta=D-C_1'$. Then Lemma 7.2 gives

$$\begin{aligned}&\sum_{\tau\in\mathcal{A}(t)}\exp\bigl(-(D-C_1')\epsilon(\mathcal{Z}_\tau)\bigr)\\&\quad\le 4\alpha\exp\bigl(-\beta\lfloor\alpha\rfloor\bigr)\\&\quad\le 4e^{\beta}\exp\bigl(-(\beta-1)\alpha\bigr)\\&\quad\le 4e^{D}\exp\bigl(-(D-C_1'-1)(1+\delta_1)\epsilon(\mathcal{Z}_{\tau^*})-(D-C_1'-1)\delta_1\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2\\&\qquad-(D-C_1'-1)ct\bigr).\end{aligned}$$

This leads to a bound for (50) as

$$\begin{aligned}&4e^{D+4\lambda^2}\exp\bigl(-\bigl((D-C_1'-1)\delta_1-C_2'\bigr)\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2\bigr)\\&\qquad\times\exp\bigl(-\bigl((D-C_1'-1)(1+\delta_1)-\bigl(D+\lambda+\lambda^{-1}+1\bigr)\bigr)\epsilon(\mathcal{Z}_{\tau^*})\\&\qquad-\bigl((D-C_1'-1)c-4\bigr)t\bigr)\\&\quad\le 4e^{D+4\lambda^2}\exp\biggl(-\frac{\delta_1 D}{2}\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2-\frac{\delta_1 D}{2}\epsilon(\mathcal{Z}_{\tau^*})-\frac{Dc}{2}t\biggr),\end{aligned}$$

for $D>\max\{\frac{\lambda+\lambda^{-1}+1+2(C_1'+1)}{\delta_1/2},2(C_1'+1)+\frac{2C_2'}{\delta_1},\frac{8}{c}+2(C_1'+1)\}$. Using Lemma 7.1, Lemma 7.2 and (5), we bound the second term (49) by

$$2\exp\bigl(-\rho C_2\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2/16-\rho t/16\bigr)\sum_{\tau\in\mathcal{A}(t)}\exp\bigl(-(\rho C_1/16-6)\epsilon(\mathcal{Z}_\tau)\bigr).\tag{51}$$

Again, we will bound the sum in the above display by Lemma 7.2 with $\alpha=(1+\delta_1)\epsilon(\mathcal{Z}_{\tau^*})+\delta_1\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2+ct$ and $\beta=\rho C_1/16-6$. That is,

$$\begin{aligned}&\sum_{\tau\in\mathcal{A}(t)}\exp\bigl(-(\rho C_1/16-6)\epsilon(\mathcal{Z}_\tau)\bigr)\\&\quad\le 4\alpha\exp\bigl(-\beta\lfloor\alpha\rfloor\bigr)\\&\quad\le 4e^{\beta}\exp\bigl(-(\beta-1)\alpha\bigr)\\&\quad\le 4e^{\rho C_1/16}\exp\bigl(-(\rho C_1/16-7)(1+\delta_1)\epsilon(\mathcal{Z}_{\tau^*})-(\rho C_1/16-7)\delta_1\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2\\&\qquad-(\rho C_1/16-7)ct\bigr).\end{aligned}$$

Therefore, (51) can be bounded by

$$\begin{aligned}&8e^{\rho C_1/16}\exp\bigl(-(\rho C_1/16-7)(1+\delta_1)\epsilon(\mathcal{Z}_{\tau^*})\bigr)\\&\quad\times\exp\bigl(-(\rho C_1/16+\rho C_2/16-7)\delta_1\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2-(\rho(C_1+1)/16-7)ct\bigr)\\&\le 8\exp\bigl(-(\rho C_1/16-8)(1+\delta_1)\epsilon(\mathcal{Z}_{\tau^*})\bigr)\\&\quad\times\exp\bigl(-(\rho C_1/16+\rho C_2/16-7)\delta_1\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2-(\rho(C_1+1)/16-7)ct\bigr)\\&\le 8\exp\bigl(-7\delta_1\|\mathscr{X}_{Z^*}(Q^*)-\theta^*\|^2-6\epsilon(\mathcal{Z}_{\tau^*})-7ct\bigr),\end{aligned}$$

for $C_1 = \max\{1, 224/\rho\}$. We obtain the desired result by combining the bounds of (48) and (49) and setting $t = 0$. $\square$

PROOF OF (9). Let us first use (33) and (47) to bound the ratio $R(Z, U(t))\mathbb{I}_{F_Z(t)}/R(Z^*)$, that is,

$$\begin{aligned}
&\frac{R(Z, U(t))\mathbb{I}_{F_Z(t)}}{R(Z^*)} \\
&\quad \le \exp\biggl(-\bigl(M/16 - \bigl(1 + \lambda + \lambda^{-1}\bigr)\bigr)\epsilon(\mathcal{Z}_{\tau^*}) - \frac{1}{8}Mt - \frac{1}{16}\delta_2 \bigl\| \mathcal{X}_{Z^*}(Q^*) - \theta^* \bigr\|^2\biggr) \\
&\quad \le \exp\biggl(-\frac{M}{32}\epsilon(\mathcal{Z}_{\tau^*}) - \frac{1}{8}Mt - \frac{1}{16}\delta_2 \bigl\| \mathcal{X}_{Z^*}(Q^*) - \theta^* \bigr\|^2\biggr),
\end{aligned}$$

for $M > \max\{128\delta_2^{-1}(C_1 + 1), 16\log(4\lambda/\sqrt{\delta_2/2}) + 256\delta_2^{-1}\lambda^2, 32(1 + \lambda + \lambda^{-1})\}$. By (24), we have

$$\begin{aligned}
\mathbb{E}\Pi\bigl(\mathcal{X}_Z(Q) \in U(t) | Y\bigr) &\le \sum_{\tau \in \mathcal{T} \cap \mathcal{A}(t)^c} \frac{\exp(-D\epsilon(\mathcal{Z}_\tau))}{\exp(-D\epsilon(\mathcal{Z}_{\tau^*}))} \frac{|\bar{\mathcal{Z}}_{\tau^*}|}{|\bar{\mathcal{Z}}_\tau|} \sum_{Z \in \bar{\mathcal{Z}}_\tau} \mathbb{E}\frac{R(Z, U(t))\mathbb{I}_{F_Z}}{R(Z^*)} && (52) \\
&\quad + \sum_{\tau \in \mathcal{T} \cap \mathcal{A}(t)^c} \sum_{Z \in \bar{\mathcal{Z}}_\tau} \mathbb{P}\bigl(F_Z(t)^c\bigr) && (53) \\
&\quad + \mathbb{E}\Pi\bigl(\tau \in \mathcal{A}(t) | Y\bigr). && (54)
\end{aligned}$$

The bound for (54) has been derived in the proof of (8). Using Lemma 7.2, we bound (52) by

$$\begin{aligned}
&\exp\biggl(-\biggl(\frac{M}{32} - D - 1\biggr)\epsilon(\mathcal{Z}_{\tau^*}) - \frac{1}{8}Mt - \frac{1}{16}\delta_2 \bigl\| \mathcal{X}_{Z^*}(Q^*) - \theta^* \bigr\|^2\biggr) \\
&\qquad \times \sum_{\tau \in \mathcal{T} \cap \mathcal{A}(t)^c} \exp\bigl(-D\epsilon(\mathcal{Z}_\tau)\bigr) \\
&\quad \le 6\exp\biggl(-\frac{M}{64}\epsilon(\mathcal{Z}_{\tau^*}) - \frac{1}{8}Mt - \frac{1}{16}\delta_2 \bigl\| \mathcal{X}_{Z^*}(Q^*) - \theta^* \bigr\|^2\biggr),
\end{aligned}$$

for $M > \max\{128\delta_2^{-1}(C_1 + 1), 16\log(4\lambda/\sqrt{\delta_2/2}) + 256\delta_2^{-1}\lambda^2, 32(1 + \lambda + \lambda^{-1}), 64(D + 1)\}$. Using Lemma 7.1 and (5), the term (53) is bounded by

$$2\exp\biggl(-\rho C_1\epsilon(\mathcal{Z}_{\tau^*})/16 - \rho C_2 \bigl\| \mathcal{X}_{Z^*}(Q^*) - \theta^* \bigr\|^2 - \frac{\rho t}{16}\biggr) \sum_{\tau \in \mathcal{T} \cap \mathcal{A}(t)^c} \exp\bigl(5\epsilon(\mathcal{Z}_\tau)\bigr).$$

We use Lemma 7.2 to bound the sum in the above display with $\alpha = (1 + \delta_1)\epsilon(\mathcal{Z}_{\tau^*}) + \delta_1 \| \mathcal{X}_{Z^*}(Q^*) - \theta^* \|^2 + ct$ and $\beta = 5$.

$$\begin{aligned}
\sum_{\tau \in \mathcal{T} \cap \mathcal{A}(t)^c} \exp\bigl(5\epsilon(\mathcal{Z}_\tau)\bigr) &\le 4(\alpha + 1)\exp\bigl(\beta(\alpha + 1)\bigr) \\
&\le 4e^{\beta+1}\exp\bigl((\beta + 1)\alpha\bigr) \\
&= 4e^6 \exp\bigl(6(1 + \delta_1)\epsilon(\mathcal{Z}_{\tau^*}) + 6\delta_1 \bigl\| \mathcal{X}_{Z^*}(Q^*) - \theta^* \bigr\|^2 + 6ct\bigr).
\end{aligned}$$

Therefore, we can bound (53) by

$$\begin{aligned}
&8e^6 \exp\biggl(-\biggl(\frac{\rho C_1}{16} - 8\biggr)\epsilon(\mathcal{Z}_{\tau^*}) - (\rho C_2 - 6\delta_1) \bigl\| \mathcal{X}_{Z^*}(Q^*) - \theta^* \bigr\|^2 - \biggl(\frac{\rho}{16} - 6c\biggr)t\biggr) \\
&\quad \le 8e^6 \exp\biggl(-6\epsilon(\mathcal{Z}_{\tau^*}) - \delta_1 \bigl\| \mathcal{X}_{Z^*}(Q^*) - \theta^* \bigr\|^2 - \frac{\rho}{32}t\biggr),
\end{aligned}$$

where we set $C_2 = \delta_2^2/128$, $C_1 = \max\{1, 224/\rho\}$, $\delta_2 = 8\sqrt{14\delta_1/\rho} = 8\sqrt{14\delta/\rho}$, and $c = \rho/192$. The proof is complete by combining the bounds of (52), (53) and (54) and setting $t = 0$. □

PROOF OF (10). In the proof of (9), we obtain a general bound for $\mathbb{E}\Pi(\mathscr{X}_Z(Q) \in U(t)|Y)$ for any $t \geq 0$. The result of (10) can be obtained by integrating out the tail probability $\mathbb{E}\Pi(\mathscr{X}_Z(Q) \in U(t)|Y)$. The details of the argument is given in Section B in the supplement. □

Acknowledgment. This work was done during the first author's visit in Leiden University in 2015. Many ideas were originated from the weekly problem sessions with Johannes Schmidt-Hieber and Kolyan Ray. Johannes Schmidt-Hieber suggested using ℓ_2 norm in the exponent of the prior. Ismaël Castillo pointed out that the rate in Corollary 5.3 can be improved, which leads to Corollary 5.9. The authors are grateful for the insightful feedbacks from three referees and an associate editor.

The research of C. Gao was supported in part by NSF Grant DMS-1712957 and NSF CAREER Award DMS-1847590. The research of A. W. van der Vaart has received funding from the European Research Council under ERC Grant Agreement 320637 and from a Spinoza grant by the Netherlands Organisation of Scientific Research NWO. The research of H. H. Zhou was supported in part by NSF Grants DMS-1811740, DMS-1918925 and NIH Grant 1P50MH115716.

SUPPLEMENTARY MATERIAL

Supplement to "A general framework for Bayes structured linear models" (DOI: 10.1214/19-AOS1909SUPP; .pdf). The supplement [23] presents additional proofs.

REFERENCES

[1] AGARWAL, A., ANANDKUMAR, A. and NETRAPALLI, P. (2013). Exact recovery of sparsely used overcomplete dictionaries. ArXiv preprint. Available at arXiv:1309.1952.

[2] ALDOUS, D. J. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.* **11** 581–598. MR0637937 https://doi.org/10.1016/0047-259X(81)90099-3

[3] BAKIN, S. (1999). Adaptive regression and model selection in data mining problems.

[4] BANERJEE, S. and GHOSAL, S. (2014). Posterior convergence rates for estimating large precision matrices using graphical models. *Electron. J. Stat.* **8** 2111–2137. MR3273620 https://doi.org/10.1214/14-EJS945

[5] BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413. MR1679028 https://doi.org/10.1007/s004400050210

[6] BARRON, A., SCHERVISH, M. J. and WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561. MR1714718 https://doi.org/10.1214/aos/1018031206

[7] BARRON, A. R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Univ. of Illinois.

[8] BARRON, A. R. and COVER, T. M. (1991). Minimum complexity density estimation. *IEEE Trans. Inform. Theory* **37** 1034–1054. MR1111806 https://doi.org/10.1109/18.86996

[9] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

[10] BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* (*JEMS*) **3** 203–268. MR1848946 https://doi.org/10.1007/s100970100031

[11] BROWN, L. D. and LOW, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398. MR1425958 https://doi.org/10.1214/aos/1032181159

[12] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

[13] BUNEA, F. (2008). 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*. *Inst. Math. Stat.* (*IMS*) *Collect.* **3** 122–137. IMS, Beachwood, OH. MR2459221 https://doi.org/10.1214/074921708000000101

[14] CANDES, E. J. and TAO, T. (2005). Decoding by linear programming. *IEEE Trans. Inform. Theory* **51** 4203–4215. MR2243152 https://doi.org/10.1109/TIT.2005.858979

[15] CASTILLO, I. (2014). On Bayesian supremum norm contraction rates. *Ann. Statist.* **42** 2058–2091. MR3262477 https://doi.org/10.1214/14-AOS1253

[16] CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. MR3375874 https://doi.org/10.1214/15-AOS1334

[17] CASTILLO, I. and VAN DER VAART, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101. MR3059077 https://doi.org/10.1214/12-AOS1029

[18] CATONI, O. (2004). *Statistical Learning Theory and Stochastic Optimization*. *Lecture Notes in Math.* **1851**. Springer, Berlin. MR2163920 https://doi.org/10.1007/b99352

[19] DIACONIS, P. and JANSON, S. (2008). Graph limits and exchangeable random graphs. *Rend. Mat. Appl.* (7) **28** 33–61. MR2463439

[20] DONOHO, D. L., ELAD, M. and TEMLYAKOV, V. N. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory* **52** 6–18. MR2237332 https://doi.org/10.1109/TIT.2005.860430

[21] FANG, K. T., KOTZ, S. and NG, K. W. (1990). *Symmetric Multivariate and Related Distributions*. *Monographs on Statistics and Applied Probability* **36**. CRC Press, London. MR1071174 https://doi.org/10.1007/978-1-4899-2937-2

[22] GAO, C., LU, Y. and ZHOU, H. H. (2015). Rate-optimal graphon estimation. *Ann. Statist.* **43** 2624–2652. MR3405606 https://doi.org/10.1214/15-AOS1354

[23] GAO, C., VAN DER VAART, A. W and ZHOU, H. H (2020). Supplement to "A General Framework for Bayes Structured Linear Models." https://doi.org/10.1214/19-AOS1909SUPP.

[24] GAO, C. and ZHOU, H. H. (2015). Rate-optimal posterior contraction for sparse PCA. *Ann. Statist.* **43** 785–818. MR3325710 https://doi.org/10.1214/14-AOS1268

[25] GAO, C. and ZHOU, H. H. (2016). Rate exact Bayesian adaptation with modified block priors. *Ann. Statist.* **44** 318–345. MR3449770 https://doi.org/10.1214/15-AOS1368

[26] GHOSAL, S., GHOSH, J. K. and RAMAMOORTHI, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27** 143–158. MR1701105 https://doi.org/10.1214/aos/1018031105

[27] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007 https://doi.org/10.1214/aos/1016218228

[28] HARTIGAN, J. A. (1972). Direct clustering of a data matrix. *J. Amer. Statist. Assoc.* **67** 123–129.

[29] HOFFMANN, M., ROUSSEAU, J. and SCHMIDT-HIEBER, J. (2015). On adaptive posterior concentration rates. *Ann. Statist.* **43** 2259–2295. MR3396985 https://doi.org/10.1214/15-AOS1341

[30] HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088 https://doi.org/10.1016/0378-8733(83)90021-7

[31] HOOVER, D. N. (1979). Relations on probability spaces and arrays of random variables 2. Institute for Advanced Study, Princeton, NJ. Preprint.

[32] JOHNSTONE, I. M. (2017). Gaussian estimation: Sequence and wavelet models.

[33] KALLENBERG, O. (1989). On the representation theorem for exchangeable arrays. *J. Multivariate Anal.* **30** 137–154. MR1003713 https://doi.org/10.1016/0047-259X(89)90092-4

[34] KLEIJN, B. J. K. and VAN DER VAART, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.* **34** 837–877. MR2283395 https://doi.org/10.1214/009053606000000029

[35] KLOPP, O., LU, Y., TSYBAKOV, A. B. and ZHOU, H. H. (2019). Structured matrix estimation and completion. *Bernoulli* **25** 3883–3911. MR4010976 https://doi.org/10.3150/19-bej1114

[36] LEUNG, G. and BARRON, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory* **52** 3396–3410. MR2242356 https://doi.org/10.1109/TIT.2006.878172

[37] LOUNICI, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.* **2** 90–102. MR2386087 https://doi.org/10.1214/08-EJS177

[38] LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204. MR2893865 https://doi.org/10.1214/11-AOS896

[39] LOVÁSZ, L. (2012). *Large Networks and Graph Limits*. *American Mathematical Society Colloquium Publications* **60**. Amer. Math. Soc., Providence, RI. MR3012035 https://doi.org/10.1090/coll/060

[40] LOVÁSZ, L. and SZEGEDY, B. (2006). Limits of dense graph sequences. *J. Combin. Theory Ser. B* **96** 933–957. MR2274085 https://doi.org/10.1016/j.jctb.2006.05.002

[41] MA, Z. and WU, Y. (2015). Volume ratio, sparsity, and minimaxity under unitarily invariant norms. *IEEE Trans. Inform. Theory* **61** 6939–6956. MR3430731 https://doi.org/10.1109/TIT.2015.2487541

[42] MARTIN, R., MESS, R. and WALKER, S. G. (2017). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli* **23** 1822–1847. MR3624879 https://doi.org/10.3150/15-BEJ797

[43] NEMIROVSKI, A. (2000). Topics in non-parametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1998). *Lecture Notes in Math*. **1738** 85–277. Springer, Berlin. MR1775640

[44] NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist*. **24** 2399–2430. MR1425959 https://doi.org/10.1214/aos/1032181160

[45] OLSHAUSEN, B. A. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381** 607–609.

[46] PATI, D. and BHATTACHARYA, A. (2015). Optimal Bayesian estimation in stochastic block models. ArXiv preprint. Available at arXiv:1505.06794.

[47] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **57** 6976–6994. MR2882274 https://doi.org/10.1109/TIT.2011.2165799

[48] RIGOLLET, P. and TSYBAKOV, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist*. **39** 731–771. MR2816337 https://doi.org/10.1214/10-AOS854

[49] RIGOLLET, P. and TSYBAKOV, A. B. (2012). Sparse estimation by exponential weighting. *Statist. Sci*. **27** 558–575. MR3025134 https://doi.org/10.1214/12-STS393

[50] RIVOIRARD, V. and ROUSSEAU, J. (2012). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Anal*. **7** 311–333. MR2934953 https://doi.org/10.1214/12-BA710

[51] SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist*. **29** 687–714. MR1865337 https://doi.org/10.1214/aos/1009210686

[52] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[53] TSYBAKOV, A. B. (2003). Optimal rates of aggregation. In *Learning Theory and Kernel Machines* 303–313. Springer, Berlin.

[54] TSYBAKOV, A. B. (2014). Aggregation and minimax optimality in high-dimensional estimation. In *Proceedings of the International Congress of Mathematicians—Seoul* 2014. *Vol. IV* 225–246. Kyung Moon Sa, Seoul. MR3727610

[55] VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat*. **3** 1360–1392. MR2576316 https://doi.org/10.1214/09-EJS506

[56] VAN DER PAS, S. L. and VAN DER VAART, A. W. (2018). Bayesian community detection. *Bayesian Anal*. **13** 767–796. MR3807866 https://doi.org/10.1214/17-BA1078

[57] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist*. **36** 1435–1463. MR2418663 https://doi.org/10.1214/009053607000000613

[58] WANG, Z., PATERLINI, S., GAO, F. and YANG, Y. (2014). Adaptive minimax regression estimation over sparse ℓ_q-hulls. *J. Mach. Learn. Res*. **15** 1675–1711. MR3225246

[59] YANG, Y. (1999). Model selection for nonparametric regression. *Statist. Sinica* **9** 475–499. MR1707850

[60] YANG, Y. (2000). Combining different procedures for adaptive regression. *J. Multivariate Anal*. **74** 135–161. MR1790617 https://doi.org/10.1006/jmva.1999.1884

[61] YANG, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli* **10** 25–47. MR2044592 https://doi.org/10.3150/bj/1077544602

[62] YANG, Y. and BARRON, A. R. (1998). An asymptotic property of model selection criteria. *IEEE Trans. Inform. Theory* **44** 95–116. MR1486651 https://doi.org/10.1109/18.650993

[63] YANG, Y. and DUNSON, D. B. (2014). Minimax optimal bayesian aggregation. ArXiv preprint. Available at arXiv:1403.1345.

[64] YANG, Y., WAINWRIGHT, M. J. and JORDAN, M. I. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *Ann. Statist*. **44** 2497–2532. MR3576552 https://doi.org/10.1214/15-AOS1417

[65] YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **68** 49–67. MR2212574 https://doi.org/10.1111/j.1467-9868.2005.00532.x

[66] ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist*. **36** 1567–1594. MR2435448 https://doi.org/10.1214/07-AOS520