BAYESIAN MODEL AVERAGING WITH EXPONENTIATED LEAST SQUARE LOSS

BY DONG DAI

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Statistics

Written under the direction of

Tong Zhang

and approved by

New Brunswick, New Jersey

October, 2013

ABSTRACT OF THE DISSERTATION

Bayesian Model Averaging with Exponentiated Least Square Loss

by Dong Dai

Dissertation Director: Tong Zhang

Given a finite family of functions, the goal of model averaging is to construct a procedure that mimics the function from this family that is the closest to an unknown regression function. More precisely, we consider a general regression model with fixed design and measure the distance between functions by mean squared error (MSE) at the design points. In this thesis, we propose a new method *Bayesian model averaging with exponentiated least square loss (BMAX)* to solve the model averaging problem optimally in a minimax sense.

Acknowledgements

I would like to express my gratitude to my advisor, Professor Tong Zhang for for his tremendous support, invaluable guidance, and constant encouragement during my years of research. He opened the door of statistical learning for me and broadened my view.

I wish to thank Professor Philippe Rigollet, Q-aggregation is a joint work with him and Tong. The application of Q-aggregation to affine estimators is a joint work with Philippe Rigollet, Tong Zhang and Lucy Xia, whom I would like thank as well.

My thanks also go to the Department of Statistics and Biostatistics at Rutgers University for providing me support and a great learning and research environment. I would like to thank Baiyang Liu, Dungang Liu, Kezhen Liu, Wenhua Jiang, Yun Pu, Jun Tan, Hong Yang and other friends, I deeply appreciate their friendship and help.

At last, I would like to thank my wife Shanshan and my parents for their love and support.

Dedication

To my hometown

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

Model Selection refers to the problem of using the data to select one model from a list of candidate models, while *model averaging* averages all the candidate models. Earlier studies have already shown that model averaging techniques provide better predictive performances than any single selected model in the presence of model uncertainty (Raftery et al., 1997). Bayesian model averaging (BMA) is a strong approach of model averaging. In this chapter, we will firstly give an introduction on Bayesian model averaging approaches, then the optimal regret of the model averaging problem is defined with an exact oracle inequality.

1.1 Bayesian Model Averaging

Given data vector $\boldsymbol{Y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$, it is often routine to consider many possible models, say $\mathcal{M}_1, \ldots, \mathcal{M}_K$, and denote the dictionary of all models as $\mathcal{H} = \{\mathcal{M}_1, \ldots, \mathcal{M}_K\}$. Each model $\mathcal{M}_j$ consists a family of distributions $\{p(\boldsymbol{Y}|\boldsymbol{\theta}_j, \mathcal{M}_j)\}$, indexed by $\boldsymbol{\theta}_j \in \Omega_j \subset \mathbb{R}^{d_j}$. The comprehensive Bayesian approach for multiple model setups proceeds by assigning a prior probability distribution $p(\boldsymbol{\theta}_j|\mathcal{M}_j)$ to the parameters of each model, and a prior probability $p(\mathcal{M}_j)$ to each model.

Under this framework, the data are realized in three stages: first the model $\mathcal{M}_j$ is generated from $p(\mathcal{M}_1), \ldots, p(\mathcal{M}_k)$; second the parameter $\boldsymbol{\theta}_j$ is generated from $p(\boldsymbol{\theta}_j|\mathcal{M}_j)$; third the data $\boldsymbol{Y}$ are generated from $p(\boldsymbol{Y}|\boldsymbol{\theta}_j, \mathcal{M}_j)$.

If Δ is the quantity of interest, such as mean in regression model, then its posterior distribution given data $\boldsymbol{Y}$ is

$$p(\Delta|\boldsymbol{Y}) = \sum_{j=1}^{K} p(\Delta|\mathcal{M}_j, \boldsymbol{Y}) p(\mathcal{M}_j|\boldsymbol{Y}). \tag{1.1}$$

This is an average of the posterior distributions under each of the models considered, weighted by their posterior model probability. The posterior probability for model $\mathcal{M}_j$ is given by

$$p(\mathcal{M}_j|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|\mathcal{M}_j)p(\mathcal{M}_j)}{\sum_{l=1}^{M} p(\boldsymbol{Y}|\mathcal{M}_l)p(\mathcal{M}_l)} \ , \tag{1.2}$$

where

$$p(\boldsymbol{Y}|\mathcal{M}_j) = \int p(\boldsymbol{Y}|\boldsymbol{\theta}_j, \mathcal{M}_j) p(\boldsymbol{\theta}_j|\mathcal{M}_j) \, d\boldsymbol{\theta}_j \tag{1.3}$$

is the marginal likelihood of $\mathcal{M}_j$.

Let $\hat{\boldsymbol{\theta}}_j \in \mathbb{R}^{d_j}$ denote the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}_j$, specifically,

$$\hat{\boldsymbol{\theta}}_j = \operatorname*{argmin}_{\boldsymbol{\theta}_j \in \Omega_j} p(\boldsymbol{Y}|\boldsymbol{\theta}_j, \mathcal{M}_j) \ ,$$

and let $\ell_j(\boldsymbol{\theta}_j) = \log p(\boldsymbol{Y}|\boldsymbol{\theta}_j, \mathcal{M}_j)$ be the log-likelihood.

The posterior mean and variance of Δ are as follows:

$$\mathbb{E}(\Delta|\boldsymbol{Y}) = \sum_{j=1}^{K} \mathbb{E}(\Delta|\mathcal{M}_j, \boldsymbol{Y}) p(\mathcal{M}_j|\boldsymbol{Y}) \ , \tag{1.4}$$

and

$$\mathrm{Var}(\Delta|\boldsymbol{Y}) = \sum_{j=1}^{K} \left[\mathrm{Var}(\Delta|\mathcal{M}_j, \boldsymbol{Y}) + (\mathbb{E}(\Delta|\mathcal{M}_j, \boldsymbol{Y}) - \mathbb{E}(\Delta|\boldsymbol{Y}))^2\right] p(\mathcal{M}_j|\boldsymbol{Y}) \ . \tag{1.5}$$

Such model averaging or mixing procedures have been developed and advocated by LEAMER (1978); GEISSER (1993); DRAPER (1995); Raftery et al. (1996); CLYDE et al. (1996).

While Bayesian model averaging (BMA) is an intuitively attractive solution to be used to overcome the problem of model uncertainty, its implementation has required careful attention to prior specification and posterior calculation.

The size of model classes often makes the summation in equation (1.2) computationally infeasible. To overcome the problem of exploding model space size in the presence of large numbers of regressors, two approaches are common:

(1) The first method is to apply the Occam's window algorithm presented in Madigan and Raftery (1994) to average over a set of parsimonious, data-supported models, selected based on model posterior probability. This algorithm discards all the models that predict the data far less accurately than the models which provide the best predictions. This means all models belonging to

$$\mathcal{A} = \left\{ \mathcal{M}_j : \frac{\max_l p(\mathcal{M}_l|\boldsymbol{Y})}{p(\mathcal{M}_j|\boldsymbol{Y})} > C \right\}$$

will be excluded where C is the data analyst's choice and $\max_l p(\mathcal{M}_l|\boldsymbol{Y})$ represents the model with the highest posterior model probability. This algorithm also excludes all complex models that receive less support from the data than their simpler counterparts. Mathematically, models that are in

$$\mathcal{B} = \{\mathcal{M}_j : \exists \mathcal{M}_l \in \mathcal{A}^c,\ \mathcal{M}_l \subset \mathcal{M}_j,\ p(\mathcal{M}_j|\boldsymbol{Y}) < p(\mathcal{M}_l|\boldsymbol{Y})\}$$

will not be considered. These two principles remarkably reduce the model space size.

(2) An alternative to Occam's window approach is to use Markov Chain Monte Carlo (MCMC) sampler that is the most common method used to obtain samples from the posterior distributions. MCMC methods sample from probability distributions based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. The state of the chain after a large number of steps is then used as a sample of the desired distribution. The quality of the sample improves as the number of steps increases. Madigan and York (1995) uses Markov chain Monte Carlo model composition (MC^3) to directly approximate (1.1). Specifically, one can

construct a Markov chain $\{\mathcal{M}(t)\},\ t = 1, 2, \ldots,$ with state space $\mathcal{H}$ and equilibrium distribution $p(\mathcal{M}_i|\boldsymbol{Y})$ and simulate this Markov chain to obtain observations $\mathcal{M}(1), ..., \mathcal{M}(N)$. Then for any function $g(\mathcal{M}_i)$ defined on $\mathcal{H}$, the average

$$\hat{G} = \frac{1}{N}\sum_{t=1}^{N} g(\mathcal{M}(t))$$

is an estimate of $\mathbb{E}(g(\mathcal{M}))$. Applying standard Markov chain Monte Carlo results,

$$\hat{G} \to \mathbb{E}(g(\mathcal{M})) \text{ a.s. as } N \to \infty$$

(see, e.g., Smith and Roberts, 1993). To compute (1.1) in this fashion set $g(\mathcal{M}) = p(\Delta|\mathcal{M}, Y)$. To construct the Markov chain, define a neighborhood $nbd(\mathcal{M})$ for each $\mathcal{M} \in \mathcal{H}$. Define a transition probability function q by setting $q(\mathcal{M} \to \mathcal{M}') = 0\ \forall\ \mathcal{M}' \notin nbd(\mathcal{M})$ and $q(\mathcal{M} \to \mathcal{M}')$ nonzero for all $\mathcal{M}' \in nbd(\mathcal{M})$. If the chain is currently in state $\mathcal{M}$, proceed by drawing $\mathcal{M}'$ from $q(\mathcal{M} \to \mathcal{M}')$. $\mathcal{M}'$ is accepted with probability

$$\min\left\{1, \frac{p(\mathcal{M}'|\boldsymbol{Y})q(\mathcal{M}' \to \mathcal{M})}{p(\mathcal{M}|\boldsymbol{Y})q(\mathcal{M} \to \mathcal{M}')}\right\} .$$

For a basic introduction to Metropolis-Hastings algorithm, see e.g. Marin and Robert (2007).

There are also two practical problems to be solved in (1.3). First, we have to choose the priors $p(\boldsymbol{\theta}_j|\mathcal{M}_j)$ and second, we have to compute the integrals. For certain interesting classes of models such as discrete graphical models (see, e.g., Madigan and York, 1995) and linear regression (see, e.g., Raftery et al., 1997), closed form integrals for the marginal likelihood, (1.3) are available.

Denote $m_j = p(\boldsymbol{Y}|\mathcal{M}_j)$, the *Bayes factor* for $\mathcal{M}_i$ versus $\mathcal{M}_j$ is defined to be

$$B_{ij} = \frac{p(\boldsymbol{Y}|\mathcal{M}_i)}{p(\boldsymbol{Y}|\mathcal{M}_j)} = \frac{\int p(\boldsymbol{Y}|\boldsymbol{\theta}_i, \mathcal{M}_i)p(\boldsymbol{\theta}_i|\mathcal{M}_i)\, d\boldsymbol{\theta}_i}{\int p(\boldsymbol{Y}|\boldsymbol{\theta}_j, \mathcal{M}_j)p(\boldsymbol{\theta}_j|\mathcal{M}_j)\, d\boldsymbol{\theta}_j} = \frac{m_i}{m_j}\ , \tag{1.6}$$

which gives a measure of the evidence for model $\mathcal{M}_i$ versus model $\mathcal{M}_j$, and with Bayes factors, (1.2) can be written as

$$p(\mathcal{M}_j|\boldsymbol{Y}) = \frac{p(\mathcal{M}_j)B_{j1}}{\sum_{l=1}^{M} p(\mathcal{M}_l)B_{l1}}$$

where B_{j1} is the Bayes factor for $\mathcal{M}_j$ versus $\mathcal{M}_1$.

If we try to use a noninformative prior for $p(\boldsymbol{\theta}_j|\mathcal{M}_j)$ we run into a problem. Recall that noninformative priors are often improper and that improper priors are only defined up to a constant. So if $p(\boldsymbol{\theta}_j|\mathcal{M}_j)$ is an improper prior for $\boldsymbol{\theta}_j$ and c_j is an arbitrary positive constant, then $q(\boldsymbol{\theta}_j|\mathcal{M}_j) = c_j p(\boldsymbol{\theta}_j|\mathcal{M}_j)$ could also be used as a prior. But now the Bayes factor becomes

$$B_{ij} = \frac{c_i}{c_j} \frac{\int p(\boldsymbol{Y}|\boldsymbol{\theta}_i, \mathcal{M}_i) p(\boldsymbol{\theta}_i|\mathcal{M}_i)\, d\boldsymbol{\theta}_i}{\int p(\boldsymbol{Y}|\boldsymbol{\theta}_j, \mathcal{M}_j) p(\boldsymbol{\theta}_j|\mathcal{M}_j)\, d\boldsymbol{\theta}_j} ,$$

so the Bayes factors and the posterior probabilities are ill-defined since there are arbitrary constants floating around in the equations.

An excellent approximation to $p(\boldsymbol{Y}|\mathcal{M}_j)$ can be provided by the Laplace method (see, e.g., Tierney and Kadane, 1986).

Let $\hat{\ell}_j = \ell_j(\hat{\boldsymbol{\theta}}_j)$ and $O_P(1)$ denotes bounded in probability. It can be shown (Kass and Wasserman, 1995) that $m_j = \hat{m}_j(1 + O_P(1))$ where

$$\log \hat{m}_j = \hat{\ell}_j - \frac{d_j}{2} \log n , \tag{1.7}$$

which is known as Bayesian information criterion (BIC) or the Schwarz criterion (Schwarz, 1978; Kass and Raftery, 1995) where for each model $\mathcal{M}_j$ the BIC formula is defined as

$$BIC_j = -2\hat{\ell}_j + d_j \log n ,$$

to be used a criteria for model selection. (1.7) means that m_j can be approximated by $\hat{m}_j$, which requires no integration and does not depend on the prior. The catch is that the error $O_P(1)$ does not go to 0 as n gets large. But it is worth pointing out that, first, quantities like m_j typically tend to ∞ as sample size increases. Hence, the error of the approximation relative to the quantity we are estimating does tend to 0. In other words, $|\hat{m}_j - m_j|/|m_j| \to 0$ in probability. Second, there are certain priors for which the approximation (1.7) has an error of size $O_P(n^{-1/2})$. One example of such a prior is a *unit information prior* (UIP) which is discussed in Kass and Wasserman (1995). A second prior that justifies the smaller error term is Jeffreys' prior $p(\boldsymbol{\theta}_j|\mathcal{M}_j) \propto |I_{\boldsymbol{\theta}_j}|^{1/2}$ where $I_{\boldsymbol{\theta}_j}$ is the Fisher information. Jeffreys' prior is usually improper and thus is plagued by the arbitrary constant. But if define the arbitrary constant in front of Jeffreys' prior is defined to be $c_j = (2\pi)^{-d_j/2}$, then it turns out that, again, the error in (1.7) is $O_P(n^{-1/2})$. In short, if we adopt the noninformative prior then $\hat{m}_j$ is a fairly accurate approximation of m_j. Thus using BIC is approximately equivalent to using Jeffreys' prior with this particular choice for the constant c_j. If we use the approximation (1.7) and let $p(\mathcal{M}_j) = \pi_j$, then

$$p(\mathcal{M}_j|\boldsymbol{Y}) \approx \frac{\hat{m}_j \pi_j}{\sum_{l=1}^{M} \hat{m}_l \pi_l} \ .$$

A more exact method to calculate $m_j = p(\boldsymbol{Y}|\mathcal{M}_j)$ is by simulation. The idea is this: we draw a random sample $\boldsymbol{\theta}_j^1, \ldots, \boldsymbol{\theta}_j^N$ from the posterior $p(\boldsymbol{\theta}|\boldsymbol{Y}, \mathcal{M}_j)$, then try to find a way to use the sample to estimate m_j.

Recall that from Bayes' theorem we have

$$p(\boldsymbol{\theta}_j|\boldsymbol{Y}, \mathcal{M}_j) = \frac{p(\boldsymbol{Y}|\boldsymbol{\theta}_j, \mathcal{M}_j)p(\boldsymbol{\theta}_j|\mathcal{M}_j)}{p(\boldsymbol{Y}|\mathcal{M}_j)} \ ,$$

it follows that

$$m_j = p(\boldsymbol{Y}|\mathcal{M}_j) = \frac{p(\boldsymbol{Y}|\boldsymbol{\theta}_j, \mathcal{M}_j)p(\boldsymbol{\theta}_j|\mathcal{M}_j)}{p(\boldsymbol{\theta}_j|\boldsymbol{Y}, \mathcal{M}_j)} \ .$$

This equation holds for all values of $\boldsymbol{\theta}_j \in \Omega_j$. Pick any value $\tilde{\boldsymbol{\theta}}_j$ of $\boldsymbol{\theta}_j$ and calculate $m_j = \frac{p(\boldsymbol{Y}|\tilde{\boldsymbol{\theta}}_j, \mathcal{M}_j)p(\tilde{\boldsymbol{\theta}}_j|\mathcal{M}_j)}{p(\tilde{\boldsymbol{\theta}}_j|\boldsymbol{Y}, \mathcal{M}_j)}$ where $p(\boldsymbol{Y}|\tilde{\boldsymbol{\theta}}_j, \mathcal{M}_j)$ and $p(\tilde{\boldsymbol{\theta}}_j|\mathcal{M}_j)$ are easy to evaluate since they are given functions. The remaining is to evaluate $p(\tilde{\boldsymbol{\theta}}_j|\boldsymbol{Y}, \mathcal{M}_j)$, the value at point $\tilde{\boldsymbol{\theta}}_j$ for function $p(\boldsymbol{\theta}_j|\boldsymbol{Y}, \mathcal{M}_j)$. Notice that

$$p(\boldsymbol{\theta}_j|\boldsymbol{Y}, \mathcal{M}_j) = \frac{p(\boldsymbol{Y}|\boldsymbol{\theta}_j, \mathcal{M}_j)p(\boldsymbol{\theta}_j|\mathcal{M}_j)}{p(\boldsymbol{Y}|\mathcal{M}_j)} \propto p(\boldsymbol{Y}|\boldsymbol{\theta}_j, \mathcal{M}_j)p(\boldsymbol{\theta}_j|\mathcal{M}_j) \ ,$$

then we can use MCMC. Specifically, pick a starting point $\boldsymbol{\theta}_j^0$, draw a candidate value ψ from some distribution $q(\boldsymbol{\theta}_j^0 \to \psi)$ (a usual choice is Gaussian distribution centered at $\boldsymbol{\theta}_j^0$), then $\boldsymbol{\theta}_j^1 = \psi$ with probability

$$\min \left\{ 1, \frac{p(\boldsymbol{Y}|\psi, \mathcal{M}_j) p(\psi|\mathcal{M}_j) q(\psi \to \boldsymbol{\theta}_j^0)}{p(\boldsymbol{Y}|\boldsymbol{\theta}_j^0, \mathcal{M}_j) p(\boldsymbol{\theta}_j^0|\mathcal{M}_j) q(\boldsymbol{\theta}_j^0 \to \psi)} \right\} ,$$

otherwise $\boldsymbol{\theta}_j^1 = \boldsymbol{\theta}_j^0$. Now draw a candidate from $q(\boldsymbol{\theta}_j^1 \to \psi)$ and so on. Continue the process until we have N draws, $\boldsymbol{\theta}_j^1, \ldots, \boldsymbol{\theta}_j^N$. Then we apply any density estimation technique (see, e.g., Silverman, 1986) to use the sample $\boldsymbol{\theta}_j^1, \ldots, \boldsymbol{\theta}_j^N$ to estimate $\hat{p}(\tilde{\boldsymbol{\theta}}_j|\boldsymbol{Y}, \mathcal{M}_j)$ of $p(\tilde{\boldsymbol{\theta}}_j|\boldsymbol{Y}, \mathcal{M}_j)$. Then our estimate of m_j is

$$\hat{m}_j = \frac{p(\boldsymbol{Y}|\tilde{\boldsymbol{\theta}}_j, \mathcal{M}_j) p(\tilde{\boldsymbol{\theta}}_j|\mathcal{M}_j)}{\hat{p}(\tilde{\boldsymbol{\theta}}_j|\boldsymbol{Y}, \mathcal{M}_j)} .$$

This process is repeated for each model to get estimates $\hat{m}_1, \ldots, \hat{m}_K$.

An alternative theory, that might be effective in these more delicate problems is the theory of *intrinsic Bayes factors* by Berger and Pericchi (1996). Briefly, suppose that we are comparing two models $\mathcal{M}_j : \{p(y|\boldsymbol{\theta}_j), \pi_j(\boldsymbol{\theta}_j)\}, \ j = 1, 2$, where $\pi_j(\boldsymbol{\theta}_j)$ are the conventional priors. (The extension to several models is straightforward.) We start with improper noninformative priors $p_j(\boldsymbol{\theta}_j) = c_j h_j(\boldsymbol{\theta}_j)$ for each model $\mathcal{M}_j$, where $h_j(\boldsymbol{\theta}_j)$ is a non-integrable function and c_j is an arbitrary constant which can not be determined. A small subset S of the data $\boldsymbol{Y} = (y_1, \ldots, y_n)^\top$ (thus denote $\boldsymbol{Y} = S \cup S^c$) is used as the training set to update the prior by Bayes' theorem. Mathematically, denote this posterior by $p(\boldsymbol{\theta}_j|S)$ which is calculated as (for $j = 1, 2$)

$$p(\boldsymbol{\theta}_j|S) = \frac{p(S|\boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j)}{m_j(S)}$$

where $m_j(S) = \int p(S|\boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j) \ d\boldsymbol{\theta}_j$ and S is such that $m_j(S) \in (0, \infty)$. With the remainder of the data S^c, the Bayes factor is computed using $p(\boldsymbol{\theta}_j|S)$ as the prior. This gives the partial Bayes factor,

$$\begin{aligned} B_{21}^{S} &= \frac{\int p(S^c|\boldsymbol{\theta}_2)p(\boldsymbol{\theta}_2|S)\ d\boldsymbol{\theta}_2}{\int p(S^c|\boldsymbol{\theta}_1)p(\boldsymbol{\theta}_1|S)\ d\boldsymbol{\theta}_1} \\ &= B_{21}B_{12}(S) \end{aligned}$$

where $B_{12}(S) = \frac{p(S|\mathcal{M}_1)}{p(S|\mathcal{M}_2)}$. Thus the partial Bayes factor B_{21}^{S} corrects B_{21} with a term $B_{12}(S)$, and the arbitrary constants c_1 and c_2 cancel out.

It should be noted that for a given sample $\boldsymbol{Y}$, we can consider different training samples S, and hence there exists a multiplicity of partial Bayes factors, one for each training sample. To avoid dependence on a particular training sample, Berger and Pericchi (1996) first suggested considering all possible subsamples S for which there is no proper subsample satisfying the inequalities $m_j(S) \in (0, \infty)$ for any c_j. They termed this subsample a *minimal training sample.* Second, they considered the arithmetic mean of B_{21}^{S} for all minimal training samples. This produces the so-called "arithmetic intrinsic Bayes factor", defined as

$$B_{21}^{AI} = B_{21} \sum_{\ell=1}^{L} B_{12}(S_\ell)\ ,$$

where L is the number of minimal training samples contained in the sample.

Other ways of "averaging" B_{21}^{S} are possible, but whereas the arithmetic mean produces priors for model selection, other methods may not necessarily do the same. The intrinsic methodology is still being developed (see, e.g., Casella and Moreno, 2006; Casella et al., 2009; Moreno et al., 2010), along with other related technique such as the *Fractional Bayes Factors* that firstly discussed in O'Hagan (1995) and De Santis and Spezzaferri (1997) and *Expected Posterior* prior (Pérez and Berger, 2002). All of these methods are within a more general topic of *Objective Bayesian* methods, see BERGER and PERICCHI (2001); Clyde and George (2004); Berger (2006) for introduction and review.

Prior density choice for BMA analysis is not limited to model priors. Priors on the parameter space also need to be specified in linear models. Most of BMA studies assume a conditionally normal distribution as the choice of prior structures for the coefficients with zero mean and a variance structure proposed by Zellner (1986).

Mathematically, given response vector $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_n)$ and design matrix $\boldsymbol{X} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_d) \in \mathbb{R}^{n \times d}$, and assume $\boldsymbol{\mu} = \mathbb{E}\boldsymbol{Y} \in \mathbb{R}^n$ in the space spanned by $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_d\}$, the columns of $\boldsymbol{X}$.

The model choice problem involves selecting a subset of predictor variables that places additional restrictions on the subspace that contains the mean. We index the model space by $\gamma \in \{0,1\}^d \subset \mathbb{R}^d$, a vector of indicators with $\gamma_j = 1$, meaning that $\boldsymbol{f}_j$ is included in the set of predictor variables, and with $\gamma_j = 0$, meaning that $\boldsymbol{f}_j$ is excluded.

Under each model $\mathcal{M}_\gamma$, $\boldsymbol{\mu}$ may be expressed in vector form as

$$\boldsymbol{\mu} | \mathcal{M}_\gamma = \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma \ ,$$

where $\boldsymbol{X}_\gamma \in \mathbb{R}^{n \times d_\gamma}$ represents the design matrix under model $\mathcal{M}_\gamma$ and $\boldsymbol{\beta}_j \in \mathbb{R}^{d_\gamma}$ is the vector of regression coefficients. Or we can write

$$\boldsymbol{Y} | \mathcal{M}_\gamma \sim N(\boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma, \sigma^2 \boldsymbol{I}_n) \ . \tag{1.8}$$

Zellner (1986)'s *g prior* for $\boldsymbol{\beta}_\gamma$ is defined as

$$\boldsymbol{\beta}_\gamma | \mathcal{M}_\gamma, g \sim N(0, g\sigma^2 (\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma)^{-1}) \ , \tag{1.9}$$

which has been widely adopted because of its computational efficiency in evaluating marginal likelihoods and model search and, perhaps most important, because of its simple, understandable interpretation as arising from the analysis of a conceptual sample generated using the same design matrix $\boldsymbol{X}$ as employed in the current sample.

The choice of g effectively controls model selection. With large g typically concentrating the prior on parsimonious models with a few large coefficients, whereas small g tends to concentrate the prior on saturated models with small coefficients (George and Foster, 2000). Recommendations for g have included the following:

- Unit information prior (UIP). Kass and Wasserman (1995) recommended choosing priors with the amount of information about the parameter equal to the amount of information contained in one observation. For regular parametric families, the "amount of information" is defined through Fisher information. In the normal regression case, the unit information prior corresponds to taking $g = n$, leading to Bayes factors that behave like the BIC. Eicher et al. (2011) conducted a thorough study of different prior structures and show that the combination of the UIP, on the parameter space and the uniform distribution on the model space is superior to any other possible combinations of priors proposed in the BMA literature.

- Risk inflation criterion. Foster and George (1994) calibrated priors for model selection based on the RIC and recommended the use of $g = d^2$ from a minimax perspective.

- Benchmark prior. Fernández et al. (2001) did a thorough study on various choices of g with dependence on the sample size n or the model dimension d and concluded with the recommendation to take $g = \max(n, d^2)$. We refer to their "benchmark prior" specification as "BRIC" as it bridges BIC and RIC.

- Empirical Bayes (EB). George and Foster (2000) proposed and developed empirical Bayes methods using either global or local estimate of g. In addition to assumption of linear model (1.8) and g priors (1.9), they also assume the hierarchical Bernoulli(w) priors of sparsity pattern γ for each model $\mathcal{M}_\gamma$, more specifically,

$$p(\mathcal{M}_\gamma|w) = w^{d_\gamma}(1-w)^{d-d_\gamma} \quad , w \in [0,1] \; , \tag{1.10}$$

where $d_\gamma = \|\gamma\|_0$.

It can be easily shown that

$$p(\mathcal{M}_\gamma|\boldsymbol{Y}, g, w) \propto \exp\left\{\frac{g}{2(1+g)}[SS_\gamma/\sigma^2 - F(g,w)d_\gamma]\right\}$$

where $SS_\gamma = \|\boldsymbol{Y} - \boldsymbol{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma\|_2^2$ with $\hat{\boldsymbol{\beta}}_\gamma$ as least square estimator for $\boldsymbol{\beta}_\gamma$, and

$$F(g,w) = \frac{g}{1+g}\left\{2\log\frac{1-w}{w} + \log(1+g)\right\} \ .$$

Thus for given $\boldsymbol{Y}$, g and w, $p(\mathcal{M}_\gamma|\boldsymbol{Y}, g, w)$ is increasing in $SS_\gamma/\sigma^2 - F(g,w)d_\gamma$, which means maximizing

$$SS_\gamma/\sigma^2 - F(g,w)d_\gamma \tag{1.11}$$

is equivalent to selecting the highest posterior model. Suitable choice of g and w will recover us the well known criteria such as AIC, BIC, RIC.

Rather than using fixed pre-specified values of g and w, George and Foster (2000) considered estimating them from the data via empirical Bayes. For this purpose, they proposed two approaches, marginal maximum likelihood (MML) and conditional maximum likelihood (CML). MML entails finding $g = \hat{g}$ and $w = \hat{w}$ that maximize the overall marginal likelihood

$$\begin{aligned} p(g,w|\boldsymbol{Y}) &\propto \sum_\gamma p(\mathcal{M}_\gamma|w)p(\boldsymbol{Y}|\mathcal{M}_\gamma, g) \\ &\propto \sum_\gamma w^{d_\gamma}(1-w)^{d-d_\gamma}(1+g)^{-d_\gamma/2}\exp\{\frac{gSS_\gamma}{2\sigma^2(1+g)}\} \ , \end{aligned} \tag{1.12}$$

and inserting them into (1.11) to obtain

$$C_{MML} = SS_\gamma/\sigma^2 - F(\hat{g}, \hat{w})d_\gamma \ . \tag{1.13}$$

Note that the penalty $F(\hat{g}, \hat{w})$ adapts to the data through the estimates of g and w. George and Foster (2000) showed via simulations that, as opposed to fixed penalty criteria, the performance of C_{MML} is nearly as good as the best possible fixed penalty criterion over a broad range of model specifications.

A drawback of C_{MML} is that it can be computationally overwhelming especially when $\boldsymbol{X}$ is nonorthogonal because maximizing (1.12) involves averaging $\mathcal{M}_\gamma$ over the whole model space. To mitigate this difficulty George and Foster (2000) also proposed C_{CML}, an easily computable alternative. C_{CML} entails choosing the model $\mathcal{M}_\gamma$ for which the conditional likelihood

$$\begin{aligned} p^*(g, w, \mathcal{M}_\gamma|\boldsymbol{Y}) &\propto p(\mathcal{M}_\gamma|w)p(\boldsymbol{Y}|\mathcal{M}_\gamma, g) \\ &\propto w^{d_\gamma}(1-w)^{d-d_\gamma}(1+g)^{-d_\gamma/2}\exp\{\frac{gSS_\gamma}{2\sigma^2(1+g)}\} \ , \end{aligned} \tag{1.14}$$

is maximized over g, w and $\mathcal{M}_\gamma$. Although its performance was not quite as good as that of C_{MML}, George and Foster (2000) showed that C_{CML} offered similar adaptive improvements over fixed penalty criteria.

- Fully Bayes (FB). Rather than using a plug-in estimate to eliminate g, a natural alternative is FB with the integrated marginal likelihood under a proper prior on g. Consequently, a prior on g leads to a mixture of g priors for the coefficients $\boldsymbol{\beta}_\gamma$, which typically provides more robust inference. And in many statistical decision problems, the admissible estimators are either Bayes or limits of Bayes procedures (Berger, 1985), one might anticipate that such FB procedures would improve over EB which are neither Bayes nor limits of Bayes procedures. Although Zellner and Siow (1980) did not explicitly use a g-prior formulation with a prior on g, their recommendation of a multivariate Cauchy form for $p(\boldsymbol{\beta}_\gamma|\sigma^2)$ implicitly corresponds to using a g-prior with an Inv-Gamma(1/2,n/2) prior on g, namely,

$$p(\boldsymbol{\beta}_\gamma|\sigma^2) \propto \int N\left(\boldsymbol{\beta}_\gamma \,|\, 0, g\sigma^2(\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma)^{-1}\right)\pi(g)\, dg \ ,$$

with

$$\pi(g) = \frac{(n/2)^{1/2}}{\Gamma(1/2)} g^{-3/2} e^{-n/(2g)} \ . \tag{1.15}$$

Besides the above Zellner-Siow prior (1.15) on g, both Liang et al. (2008)

and Cui and George (2008) investigated another prior on g in the form of

$$\pi(g) = \frac{a-2}{2}(1+g)^{-a/2} \ , \quad g > 0 \ , \tag{1.16}$$

which is a proper distribution for $a > 2$. This family of priors includes priors used by Strawderman (1971) to provide improved mean square risk over ordinary maximum likelihood estimates in the normal means problem. Liang et al. (2008) also modified the Strawderman prior to

$$\pi(g) = \frac{a-2}{2n}(1+\frac{g}{n})^{-a/2} \ ,$$

for the consistency under the null model. Liang et al. (2008) found that all of the FB mixture g-priors do as well as the global EB with model selection, except under the null model, whereas Cui and George (2008) found that the global EB outperformed FB procedures (under the assumption of known σ^2). Liang et al. (2008) used a uniform prior on the model space (for both the EB and the FB procedures), whereas Cui and George (2008) placed independent Bernoulli(w) priors on variable inclusion and compared EB estimates of w with FB procedures that place a uniform prior on w. Thus the prior distributions over models is an important aspect and may explain some of the difference in their findings. Additionally, the simulations in Cui and George (2008) are for the $d = n$ case while Liang et al. (2008) show that FB procedures are consistent as $n \to \infty$ for fixed d, additional study of their theoretical properties is necessary for the situation when p is close to the sample size n. Maruyama and George (2011) extended FB with generalized g-prior and a beta-prime distribution as a prior on g.

Although BMA has become a mainstream tool in empirical settings with large numbers of potential regressors, it remains problematic. The set of priors on parameters within a model and the set of the prior model probabilities must be specified before calculating posterior probabilities attached to different models.

As stated in Ley and Steel (2009), the choice of prior distributions can be extremely critical for the outcome of BMA analysis. These prior probabilities must be informative with respect to the likelihood, meaning priors should be relatively high where the likelihoods are large; otherwise, the choice of priors will have a substantial effect on the posteriors. Another issue, raised by Hjort and Claeskens (2003), is the difficulty of dealing with the priors when they are in conflict with each other, stemming from mixing together many prior opinions regarding the parameters of interest.

1.2 Optimal Regret of Model Averaging under Misspecification

The seminal works of Nemirovski (2000) and Tsybakov (2003) have introduced an idealized setup to study the properties of model averaging procedures independently of the models themselves.

Let $x_1, \dots, x_n$ be n given design points in a space $\mathcal{X}$ and let $\mathcal{H} = \{f_1, \dots, f_M\}$ be a given dictionary of real valued functions on $\mathcal{X}$. The goal is to estimate an unknown regression function $\eta : \mathcal{X} \to \mathbb{R}$ at the design points based on observations

$$Y_i = \eta(x_i) + \xi_i \,,$$

where $\xi_1, \dots, \xi_n$ are i.i.d $\mathcal{N}(0, \sigma^2)$.

The performance of an estimator $\hat{\eta}$ is measured by its mean square error (MSE) defined by

$$\mathrm{MSE}(\hat{\eta}) = \frac{1}{n} \sum_{i=1}^{n} (\hat{\eta}(x_i) - \eta(x_i))^2 \,.$$

The goal is to build an estimator $\hat{\eta}$ that mimics the function f_j in the dictionary with the smallest MSE. Formally, a good estimator $\hat{\eta}$ should satisfy the following *exact oracle inequality* in a certain probabilistic sense:

$$\mathrm{MSE}(\hat{\eta}) \le \min_{j=1,\dots,M} \mathrm{MSE}(f_j) + \Delta(n, M, \sigma^2) \,, \tag{1.17}$$

where the remainder term $\Delta > 0$ should be as small as possible. Note that oracle inequality (1.17) is a truly finite sample result and the remainder term should show the interplay between the three fundamental parameters of the problem: the "dimension" M, the sample size n and the noise level σ^2.

From the early days of model averaging problem, it has been established (see, e.g., Tsybakov, 2003; Rigollet, 2012) that the smallest possible order for $\Delta(n, M, \sigma^2)$ was $\sigma^2 \log M/n$ for oracle inequalities in expectation, where "smallest possible" is understood in the following minimax sense. There exists a dictionary $\mathcal{H} = \{f_1, \dots, f_M\}$ such that the following holds. For any estimator $\hat{\eta}$, there exists a regression function η such that

$$\mathbb{E}\,\mathrm{MSE}(\hat{\eta}) \geq \min_{j=1,\dots,M} \mathrm{MSE}(f_j) + C\sigma^2 \frac{\log M}{n}\,.$$

for some positive constant C. Moreover, it follows from the same results that this lower bound holds not only in expectation but also with positive probability.

While the goal is to mimic the best model in the dictionary $\mathcal{H}$, it has been shown (see Rigollet and Tsybakov, 2012, Theorem 2.1) that there exists a dictionary $\mathcal{H}$ such that any estimator (selector) $\hat{\eta}$ restricted to be one of the elements of $\mathcal{H}$ cannot satisfy an oracle inequality such as (1.17) with a remainder term of order smaller than $\sigma\sqrt{(\log M)/n}$, in other words, model selection is suboptimal to compete with best single model from a given family.

Rather than model selection, model averaging has been successfully employed to derive oracle inequalities (1.17) in expectation (see the references in Rigollet and Tsybakov, 2012) with notable exceptions (Audibert, 2008; Lecué and Mendelson, 2009; Gaïffas and Lecué, 2011; Dai and Zhang, 2011; Rigollet, 2012; Dai et al., 2012) who produced oracle inequalities that hold in deviation (with high probability).

When the oracle inequality (1.17) holds in expectation, the remainder term $\Delta(n, M, \sigma^2)$ assesses the expected risk of $\hat{\eta}$ compared to the best single model in

$\mathcal{H}$, but it does not precise the fluctuations of risk. In several application fields of learning algorithms, these fluctuations play a key role: in finance for instance, the bigger the losses can be, the more money the bank needs to freeze in order to alleviate these possible losses. In this case, a good algorithm is an algorithm having not only low expected risk but also small deviations.

Thus below we seperate the cases of in expectation and in deviation when oracle inequality (1.17) holds. Hereafter we call an estimator $\hat{\eta}$ is *expectation optimal* (or *optimal in expectation*) if $\hat{\eta}$ satisfies exact oracle inequality with remainder term $\Delta(n, M, \sigma^2)$ of order $\sigma^2 \frac{\log M}{n}$:

$$\mathbb{E}\,\mathrm{MSE}(\hat{\eta}) \leq \min_{j=1,\dots,M} \mathrm{MSE}(f_j) + \Delta(n, M, \sigma^2) \; ; \tag{1.18}$$

and $\hat{\eta}$ is called *deviation optimal* (or *optimal in deviation*) if it satisfies the following probably approximately correct (PAC) type inequality with probability greater than $1-\delta$ with remainder $\Delta(n, M, \sigma^2, \delta)$ of order $\sigma^2 \frac{\log(M/\delta)}{n}$:

$$\mathrm{MSE}(\hat{\eta}) \leq \min_{j=1,\dots,M} \mathrm{MSE}(f_j) + \Delta(n, M, \sigma^2, \delta) \; . \tag{1.19}$$

Precisely, model averaging consists in choosing $\hat{\eta}$ as a convex combination of the f_j's with carefully chosen weights. Let Λ^M be the flat simplex of $\mathbb{R}^M$ defined by

$$\Lambda^M = \Big\{ \lambda = (\lambda_1, \dots, \lambda_M)^\top \in \mathbb{R}^M \; : \; \lambda_j \geq 0 \, , \sum_{j=1}^{M} \lambda_j = 1 \Big\} \, .$$

Given dictionary $\mathcal{H}$, each $\boldsymbol{\lambda} \in \Lambda^M$ yields a model averaging estimator $\hat{\eta} = \mathsf{f}_{\boldsymbol{\lambda}}$, where

$$\mathsf{f}_{\boldsymbol{\lambda}} = \sum_{j=1}^{M} \lambda_j f_j \, .$$

The early papers of Catoni (1999) and Yang (1999) introduced and proved optimal theoretical guarantees for a model averaging estimator called *progressive mixture* that was later studied in Audibert (2008) and Juditsky et al. (2008) from various perspectives. This estimator is based on *exponential weights*, which, since then, have been predominantly used and have led to optimal oracle inequalities in expectation.

Mathematically, let $\boldsymbol{\pi} = (\pi_1, \dots, \pi_M)^\top \in \Lambda^M$ be a given prior and $\beta > 0$ be a temperature parameter, then the jth exponential weight is given by

$$\lambda_j^{\text{EXP}} \propto \pi_j \exp\big(- n\widehat{\text{MSE}}(f_j)/\beta\big) \,, \tag{1.20}$$

where

$$\widehat{\text{MSE}}(f_j) = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - f_j(x_i)\big)^2 \,.$$

Then it is shown that (see, e.g., Rigollet and Tsybakov, 2011; Dalalyan and Salmon, 2012; Rigollet and Tsybakov, 2012)

$$\mathbb{E}\,\text{MSE}(\mathsf{f}_{\boldsymbol{\lambda}^{\text{EXP}}}) \le \min_j \left\{ \text{MSE}(f_j) + \frac{\beta}{n}\log(\pi_j^{-1}) \right\}$$

where the most common prior choice of $\boldsymbol{\pi}$ is the uniform prior $\boldsymbol{\pi} = (1/M, \dots, 1/M)^\top$ but other choices that put more or less weight on different functions of the dictionary have been successfully applied to various related problems (see, e.g., Rigollet and Tsybakov, 2011; Dalalyan and Salmon, 2012; Rigollet and Tsybakov, 2012).

The fixed design Gaussian regression was considered in Dalalyan and Tsybakov (2007, 2008) who proved an oracle inequality of the form (1.17) with optimal remainder term. This result holds only in expectation and not with high probability. While the limitation may have followed the proof technique, we actually show in next chapter that it is inherent to exponential weights. Consequently, we say that exponential weights are *deviation suboptimal* since the expectation of the resulting MSE is of the optimal order but the deviations around the expectation are not. Note also that the original paper of Dalalyan and Tsybakov (2007) made some boundedness assumption on the distance between function in the dictionary $\mathcal{H}$ and the regression function η. This assumption was lifted in their subsequent paper (Dalalyan and Tsybakov, 2008). In this thesis, we make no such assumption except for the lower bound.

For regression with random design, Audibert (2008) observed also that various progressive mixture rules are deviation suboptimal. In the same paper, he addressed this issue by proposing the STAR algorithm which is optimal both in expectation and in deviation under the uniform prior and, remarkably, does not require any parameter tuning as opposed to progressive mixture rules. Mathematically, suppose f_{k_1} is the empirical risk minimizer among functions in $\mathcal{H}$, where

$$k_1 = \operatorname*{argmin}_j \widehat{\mathrm{MSE}}(f_j) \ , \tag{1.21}$$

the STAR estimator f^* is defined as

$$f^* = (1-\alpha^*) f_{k_1} + \alpha^* f_{k_2} \ , \tag{1.22}$$

where

$$(\alpha^*, k_2) = \operatorname*{argmin}_{\alpha\in(0,1),j} \widehat{\mathrm{MSE}}\big((1-\alpha) f_{k_1} + \alpha f_j\big) \ . \tag{1.23}$$

Also for random design, Lecué and Mendelson (2009) followed by Gaïffas and Lecué (2011) proposed deviation optimal methods based on the same sample splitting idea. However, sample splitting method does not carry to fixed design.

Subsequently, Dai et al. (2012) proposed a new Q-aggregation estimator, which is quite similar to that proposed in Rigollet (2012). The Q-aggregation estimator enjoys the same theoretical properties as the STAR algorithm but for fixed design regression, with implementation of a greedy algorithm GMA-0, which is a cleaner version of Greedy Model Averaging (GMA) algorithm firstly proposed in an earlier paper by Dai and Zhang (2011), and GMA-0 and GMA both enjoy optimal deviation. Though the deviation optimality of Q-aggregation is derived with sharpest exact oracle inequality which holds both in expectation and in deviation, but there are also two limitations there. (1) Q-aggregation could be generalized for continuous candidates dictionary $\mathcal{H}$, but the greedy model averaging method GMA-0 can not be adapted to this scenario, and to solve Q-aggregation is to estimate a posterior distribution; (2) though Q-aggregation can be regarded intuitively as regression plus variance penalty, it still lacks of good interpretation.

In Chapter 2, an innovative method, Bayesian Model Averaging with Exponentiated Least Square Loss (BMAX) will be introduced. While exponential weighted model averaging estimator, can be treated as Bayes estimator (posterior mean) under least square loss, it is already shown to be optimal in expectation, yet it is deviation suboptimal. The new model averaging estimator, aggregate by BMAX, is essentially a Bayes estimator under some exponentiated least square loss, and is proven to be optimal both in expectation and in deviation. Not only the aggregate by BMAX can be approximated by greedy approach for discrete candidates dictionary $\mathcal{H}$, gradient descent algorithm can also be implemented to this strong convex optimization problem and it can be adapted to continuous candidates dictionary. Moreover, under some conditions Q-aggregation in Dai et al. (2012) is essentially a dual representation of aggregate by BMAX, which is yet more extensive and better defined, lifting the two limitations of Q-aggregation mentioned above.

BMAX is applied to linear models with Gaussian priors in Chapter 3, where naturally, the BMAX estimator is competitive to the best single linear model, and a gradient descent algorithm implemented with Metropolis-Hastings sampler is proposed. In addition, a Frequentist's aggregation of affine estimators is provided, as an extension of Q-aggregation from static models to affine estimators which are not independent of noise.

In Chapter 4 we discuss the mixture of g under the same Bayesian frame work of Chapter 3 with application of BMAX, and the BMAX estimator is shown to be competitive to that of Chapter 3 with hyper parameters of distribution on g properly chosen.

Chapter 2

Bayesian Model Averaging with Exponentiated Least Square Loss (BMAX)

2.1 Notations and Settings

Let $x_1, \ldots, x_n$ be n given design points in a space $\mathcal{X}$, let $\mathcal{H} = \{f_1, \ldots, f_M\}$ be a given dictionary of real valued functions on $\mathcal{X}$ and denote $\boldsymbol{f}_j = (f_j(x_1), \ldots, f_j(x_n))^\top \in \mathbb{R}^n$ for each j. The goal is to estimate an unknown regression function $\eta : \mathcal{X} \to \mathbb{R}$ at the design points based on observations

$$y_i = \eta(x_i) + \xi_i \ ,$$

where $\xi_1, \ldots, \xi_n$ are i.i.d $\mathcal{N}(0, \sigma^2)$.

Denote vectors as $\boldsymbol{Y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{\eta} = (\eta(x_1), \ldots, \eta(x_n))^\top$ and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$, the true model can be expressed as

$$\boldsymbol{Y} = \boldsymbol{\eta} + \boldsymbol{\xi} \ , \tag{2.1}$$

with $\boldsymbol{\xi} \sim N(0, \sigma^2 \boldsymbol{I}_n)$. Denote ℓ_2 norm as $\|\boldsymbol{Y}\|_2 = (\sum_{i=1}^n y_i^2)^{1/2}$ and inner product as $\langle \boldsymbol{\xi}, \boldsymbol{f} \rangle_2 = \boldsymbol{\xi}^\top \boldsymbol{f}$.

Let Λ^M be the flat simplex in $\mathbb{R}^M$ defined by

$$\Lambda^M = \left\{ \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_M)^\top \in \mathbb{R}^M : \lambda_j \geq 0, \sum_{j=1}^M \lambda_j = 1 \right\} \ ,$$

and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)^\top \in \Lambda^M$ be a given prior.

The Kullback-Leibler divergence for $\boldsymbol{\lambda}, \boldsymbol{\pi} \in \Lambda^M$ is defined as

$$\mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi}) = \sum_{j=1}^M \lambda_j \log(\lambda_j / \pi_j) \ ,$$

and hereafter, we use the convention $0 \cdot \log(0) = 0$.

Given any $\boldsymbol{\lambda} \in \Lambda^M$ and $\mathcal{H}$, we define the model averaging estimator for function η as

$$\mathsf{f}_{\boldsymbol{\lambda}} = \sum_{j=1}^{M} \lambda_j f_j \ ,$$

then with notation $\mathfrak{f}_{\boldsymbol{\lambda}} = (\mathsf{f}_{\boldsymbol{\lambda}}(x_1), \ldots, \mathsf{f}_{\boldsymbol{\lambda}}(x_n))^\top$ we have

$$\mathfrak{f}_{\boldsymbol{\lambda}} = \sum_{j=1}^{M} \lambda_j \boldsymbol{f}_j \ ,$$

which will be used as estimator for $\boldsymbol{\eta} \in \mathbb{R}^n$.

Define $V(\boldsymbol{\lambda})$ as

$$V(\boldsymbol{\lambda}) = \sum_{j=1}^{M} \lambda_j \|\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2 \ , \tag{2.2}$$

the variance of aggregation by $\boldsymbol{\lambda} \in \Lambda^M$ given $\mathcal{H}$ on design points.

Given $\nu \in (0, 1)$, define $P(\boldsymbol{\lambda})$ as

$$P(\boldsymbol{\lambda}) = (1 - \nu)\|\mathfrak{f}_{\boldsymbol{\lambda}} - \boldsymbol{\eta}\|_2^2 + \nu \sum_{j=1}^{M} \lambda_j \|\boldsymbol{f}_j - \boldsymbol{\eta}\|_2^2 \ , \tag{2.3}$$

and it is easy to see that $P(\boldsymbol{\lambda}) = \|\mathfrak{f}_{\boldsymbol{\lambda}} - \boldsymbol{\eta}\|_2^2 + \nu V(\boldsymbol{\lambda})$.

2.2 Deviation suboptimality of two commonly used estimators

2.2.1 Aggregate by exponential weights

The exponential weights $\boldsymbol{\lambda}^{\text{EXP}} = (\lambda_1^{\text{EXP}}, \ldots, \lambda_M^{\text{EXP}})^\top \in \Lambda^M$ are defined as

$$\lambda_j^{\text{EXP}} \propto \pi_j \exp\left(-\frac{\|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2}{2\omega^2}\right), \ \forall\, j \in \{1, \ldots, M\}. \tag{2.4}$$

where $\omega > 0$ is a temperature parameter.

It is well known (see, e.g., Rigollet and Tsybakov, 2012) that the exponential weights $\boldsymbol{\lambda}^{\text{EXP}}$ defined in (2.4) are the solution to a minimization problem:

$$\boldsymbol{\lambda}^{\text{EXP}} \in \underset{\boldsymbol{\lambda} \in \Lambda^M}{\operatorname{argmin}} \left\{ \sum_{j=1}^{M} \lambda_j \|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2 + 2\omega^2 \mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi}) \right\} . \tag{2.5}$$

It was shown in Dalalyan and Tsybakov (2007, 2008) that for $\omega^2 \geq 2\sigma^2$, it holds that

$$\mathbb{E}\|\mathsf{f}_{\boldsymbol{\lambda}^{\text{EXP}}} - \boldsymbol{\eta}\|_2^2 \leq \min_{j=1,\dots,M} \left\{ \|\boldsymbol{f}_j - \boldsymbol{\eta}\|_2^2 + 2\omega^2 \log(\pi_j^{-1}) \right\} . \tag{2.6}$$

The proof of this result relies heavily on the fact that the oracle inequality holds in expectation and whether the result also holds with high probability arises as a natural question. While the paper of Audibert (2008) does not cover the fixed design Gaussian regression framework of our paper and concerns exponential weights with an extra averaging step, it contributed to the common belief that exponential weights would be suboptimal in deviation. In particular, Lecué and Mendelson (2012) derived lower bounds for the performance of exponential weights in expectation when ω^2 is chosen below a certain constant threshold in the case of regression with random design. Moreover, they proved deviation suboptimality of exponential weights when ω^2 is less than $\sqrt{n}/(\log n)$. However, these lower bounds rely heavily on the fact that the design is random and do not extend to the fixed design case.

Now we consider the following dictionary $\mathcal{H}$. Assume that $M, n \geq 3$. Let $\mathsf{e}^{(1)} = (1, 0, \dots, 0)^\top \in \mathbb{R}^n$ and $\mathsf{e}^{(2)} = (0, 1, 0, \dots, 0)^\top \in \mathbb{R}^n$ be the first two vectors of the canonical basis of $\mathbb{R}^n$. Moreover, let $\mathsf{e}^{(3)}, \dots, \mathsf{e}^{(M)} \in \mathbb{R}^n$ be $M - 2$ unit vectors of $\mathbb{R}^n$ that are orthogonal to both $\mathsf{e}^{(1)}$ and $\mathsf{e}^{(2)}$. Let $\boldsymbol{f}_1, \dots, \boldsymbol{f}_M$ be such that

$$\boldsymbol{f}_1 = \sigma\sqrt{n}\mathsf{e}^{(1)} , \quad \boldsymbol{f}_2 = \sigma(1 + \sqrt{n})\mathsf{e}^{(2)} ,$$

and for any $3 \leq j \leq M$, $\boldsymbol{f}_j$ is defined by

$$\boldsymbol{f}_j = \boldsymbol{f}_2 + \sigma\alpha_j \mathsf{e}^{(j)} ,$$

where $\alpha_3, \dots, \alpha_M \geq 0$ are tuning parameter to be chosen later. Moreover, take the regression function $\boldsymbol{\eta} \equiv 0$ so that $\text{MSE}(\boldsymbol{f}_1) \leq \text{MSE}(\boldsymbol{f}_j)$ for any $j \geq 2$. Observe that $\|\boldsymbol{f}_j\|_2^2/n \geq \sigma^2$ so that the following lower bounds cannot be interpreted as artifacts of scaling the signal-to-noise ratio.

Assume that $M \geq 4$ and $n \geq 3$. We call parameters $\omega > 0$ as *low temperatures* when

$$\omega^2 \leq \frac{\sigma^2 \sqrt{n}}{\log(8\sqrt{n})} \,. \tag{2.7}$$

In particular the exponential weights employed in the literature on model averaging use the low temperature $\omega^2 = 2\sigma^2$ (see, e.g., (2.6) above).

Proposition 1. *Fix $M \geq 4, n \geq 3$ and assume that the noise random variables $\xi_1, \ldots, \xi_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. Let $\boldsymbol{\eta}$ and $\mathcal{H}$ be defined as above. Then, the aggregate $\mathsf{f}_{\boldsymbol{\lambda}^{\text{EXP}}}$ with exponential weights $\boldsymbol{\lambda}^{\text{EXP}}$ given by* (2.4) *satisfies*

$$\text{MSE}(\mathsf{f}_{\boldsymbol{\lambda}^{\text{EXP}}}) \geq \min_{j=1,\ldots,M} \text{MSE}(f_j) + \frac{\sigma^2}{4\sqrt{n}} \,,$$

with probability at least 0.07 *at low temperatures, for any $\alpha_3, \ldots, \alpha_M \geq 0$.*

Moreover, if $M \geq 8\sqrt{n}$ and for any $j \geq 3$, we have

$$2\sqrt{2\log(100M)} \leq \alpha_j \leq n^{1/4} \,, \tag{2.8}$$

then, the same result holds at any temperature, with probability at least 0.06*.*

Remark 1. *Proposition 1 states precisely that exponential weights are deviation suboptimal, if ω^2 is chosen small enough and in particular if ω is* any *constant with respect to M and n.*

2.2.2 Aggregate by projection

Another natural solution to solve the model averaging problem is to take the vector of weights $\boldsymbol{\lambda}^{\text{PROJ}}$ defined by

$$\boldsymbol{\lambda}^{\text{PROJ}} \in \operatorname*{argmin}_{\boldsymbol{\lambda} \in \Lambda^M} \widehat{\text{MSE}}(\mathsf{f}_{\boldsymbol{\lambda}}) \,, \tag{2.9}$$

which minimizes the empirical risk. We call $\boldsymbol{\lambda}^{\text{PROJ}}$ the vector of *projection weights* since the aggregate estimator $\mathsf{f}_{\boldsymbol{\lambda}^{\text{PROJ}}}$ is the projection of $\boldsymbol{Y}$ onto the convex hull of the $\boldsymbol{f}_j$s.

It has been established that this choice is *near*-optimal for the more difficult problem of *convex aggregation* with fixed design (see Juditsky and Nemirovski, 2000; Nemirovski, 2000; Rigollet, 2012) where the goal is to mimic the best convex combination of the $\boldsymbol{f}_j$s as opposed to simply mimicking the best single one of them. More precisely, it follows from Theorem 3.5 in Rigollet (2012) that

$$\begin{aligned}\mathbb{E}\,\mathrm{MSE}\left(\mathsf{f}_{\boldsymbol{\lambda}^{\text{PROJ}}}\right) &\le \min_{\boldsymbol{\lambda}\in\Lambda^M} \mathrm{MSE}(\mathsf{f}_{\boldsymbol{\lambda}}) + 2\sigma\sqrt{\frac{\log M}{n}}\\ &\le \min_{j=1,\dots,M} \mathrm{MSE}(f_j) + 2\sigma\sqrt{\frac{\log M}{n}}\,,\end{aligned}$$

and a similar oracle inequality also holds with high probability. The second inequality is very coarse and it is therefore natural to study whether a finer analysis of this estimator would yield an optimal oracle inequality for the aggregate $\mathfrak{f}_{\boldsymbol{\lambda}^{\text{PROJ}}}$ both in expectation and with high probability. This question was investigated by Lecué and Mendelson (2009) who proved that $\mathfrak{f}_{\boldsymbol{\lambda}^{\text{PROJ}}}$ cannot satisfy an oracle inequality of the form (1.17) with high probability and with a remainder term $\Delta(n, M, \sigma^2)$ of order smaller than $n^{-1/2}$. Their proof, however, heavily uses the fact that the design is random and we extend it to the fixed design case in Proposition 2 below.

Our lower bound for the aggregate by projection relies on a different construction of the dictionary. Let m be the smallest integer that satisfies $m^2 \ge 4n/13$ and let n, M be large enough to ensure that $m \ge 16$, $M-1 \ge 2m$. Let $\mathsf{e}^{(1)}, \dots, \mathsf{e}^{(m)} \in \mathbb{R}^n$ be the first m vectors of the canonical basis of $\mathbb{R}^n$. For any $j = 1, \dots, M$, the $\boldsymbol{f}_j$s are defined as

$$\boldsymbol{f}_j = \begin{cases} \sqrt{n}\mathsf{e}^{(j)} & \text{if} \quad 1 \le j \le m,\\ -\sqrt{n}\mathsf{e}^{(j)} & \text{if} \quad m+1 \le j \le 2m,\\ 0 & \text{if} \quad j = 2m+1\,,\\ \boldsymbol{f}_1 & \text{if} \quad j > 2m+1\,,\end{cases}$$

Moreover, define $\boldsymbol{\eta} \equiv 0$ so that $0 = \mathrm{MSE}(f_{2m+1}) \le \mathrm{MSE}(f_j)$ for all $1 \le j \le M$.

Proposition 2. *Fix $n \geq 416, M \geq \sqrt{n}$ and assume that the noise random variables $\xi_1, \ldots, \xi_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. Let $\boldsymbol{\eta}$ and $\mathcal{H}$ be defined as above. Then, the projection aggregate $\mathsf{f}_{\boldsymbol{\lambda}^{\text{PROJ}}}$ with weights $\boldsymbol{\lambda}^{\text{PROJ}}$ defined in* (2.9) *is such that*

$$\text{MSE}(\mathsf{f}_{\boldsymbol{\lambda}^{\text{PROJ}}}) \geq \min_{j=1,\ldots,M} \text{MSE}(f_j) + \frac{\sigma^2}{\sqrt{48n}} ,$$

with probability larger than $1/4$. *Moreover, the above lower bound holds with arbitrary large probability if n is chosen large enough.*

Note that we employed a different dictionary for each of the aggregates. Therefore, it may be the case that choosing the right aggregate for the right dictionary gives the correct deviation bounds. In the next section, we propose a new model averaging method, Bayesian model averaging with exponentiated least square loss (BMAX), that automatically adjusts the aggregate to the dictionary at hand.

2.3 Deviation Optimal Aggregate by Bayesian Model Averaging with Exponentiated Least Square Loss

In this section, we will show that the aggregate by Bayesian model averaging with exponentiated least square loss (BMAX) is deviation and expectation optimal.

Consider the following Bayesian framework, $\boldsymbol{Y}$ is normally distributed with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)^\top$ and covariance matrix $\omega^2 \boldsymbol{I}_n$:

$$\boldsymbol{Y} | \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \omega^2 \boldsymbol{I}_n) , \tag{2.10}$$

and for $j = 1, \ldots, M$, the prior for each model is

$$\pi(\boldsymbol{\mu} = \boldsymbol{f}_j) = \pi_j . \tag{2.11}$$

Then the posterior distribution of $\boldsymbol{\mu}$ given $\boldsymbol{Y}$ is

$$p(\boldsymbol{\mu} = \boldsymbol{f}_j | \boldsymbol{Y}) = \frac{p(\boldsymbol{Y} | \boldsymbol{\mu} = \boldsymbol{f}_j) p(\boldsymbol{\mu} = \boldsymbol{f}_j)}{\sum_{j=1}^M p(\boldsymbol{Y} | \boldsymbol{\mu} = \boldsymbol{f}_j) p(\boldsymbol{\mu} = \boldsymbol{f}_j)} = \frac{\exp\left(-\frac{\|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2}{2\omega^2}\right) \pi_j}{\sum_{j=1}^M \exp\left(-\frac{\|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2}{2\omega^2}\right) \pi_j}$$

The Bayesian framework above is to output an aggregate estimator for the model averaging problem, given $\boldsymbol{Y}$ and $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M\}$. Please note that parameters $\boldsymbol{\mu}$ and ω^2 are not necessarily equal to true mean $\boldsymbol{\eta}$ and variance σ^2 in (2.1), and $\boldsymbol{\eta}$ is not necessarily in dictionary $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M\}$. Yet the Bayesian framework and loss function $L(\boldsymbol{\psi}, \boldsymbol{\mu})$ defined and discussed later will convey interpretation of the model averaging problem. We will, avoid the question of whether the Bayesian or Frequentist approach to statistics is "philosophically correct", the focus here is simply on methodology.

The quantity of interest is $\boldsymbol{\eta} = \mathbb{E}\boldsymbol{Y}$, we consider Bayes estimator $\hat{\boldsymbol{\psi}}$, which minimizes the posterior expected loss from $\boldsymbol{\mu}$, the mean of $\boldsymbol{Y}$ in our Bayesian framework:

$$\hat{\boldsymbol{\psi}} = \operatorname*{argmin}_{\boldsymbol{\psi} \in \mathbb{R}^n} \mathbb{E}\left[L(\boldsymbol{\psi}, \boldsymbol{\mu}) | \boldsymbol{Y}\right] \ , \tag{2.12}$$

where L is some loss function.

With least square loss $L(\boldsymbol{\psi}, \boldsymbol{\mu}) = \|\boldsymbol{\psi} - \boldsymbol{\mu}\|_2^2$, the Bayes estimator is posterior mean, which is essentially Exponential Weighted Aggregation (EWA) estimator (Rigollet and Tsybakov, 2012) :

$$\boldsymbol{\psi}_{\ell_2}(\omega^2) = \frac{\sum_{j=1}^M \exp\left(-\frac{\|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2}{2\omega^2}\right) \pi_j \boldsymbol{f}_j}{\sum_{j=1}^M \exp\left(-\frac{\|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2}{2\omega^2}\right) \pi_j} \tag{2.13}$$

which is already proven to optimal in expectation (Dalalyan and Tsybakov, 2007, 2008), yet suboptimal in deviation (Proposition 1).

In this thesis, we introduce an exponentiated least square loss

$$L(\boldsymbol{\psi}, \boldsymbol{\mu}) = \exp\left(\frac{1-\nu}{2\omega^2} \|\boldsymbol{\psi} - \boldsymbol{\mu}\|_2^2\right) \ , \tag{2.14}$$

where $\nu \in (0, 1)$.

It follows that

$$\begin{aligned}\mathbb{E}\left[L(\boldsymbol{\psi},\boldsymbol{\mu})|\boldsymbol{Y}\right] &= \sum_{j=1}^{M} L(\boldsymbol{\psi},\boldsymbol{f}_j)p(\boldsymbol{\mu}=\boldsymbol{f}_j|\boldsymbol{Y}) \\ &= \sum_{j=1}^{M} \exp\left(\frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi}-\boldsymbol{f}_j\|_2^2\right)\frac{\exp\left(-\frac{\|\boldsymbol{f}_j-\boldsymbol{Y}\|_2^2}{2\omega^2}\right)\pi_j}{\sum_{j=1}^{M}\exp\left(-\frac{\|\boldsymbol{f}_j-\boldsymbol{Y}\|_2^2}{2\omega^2}\right)\pi_j}\end{aligned}$$

Thus the Bayes estimator defined in (2.12) with loss (2.14) is

$$\boldsymbol{\psi}_X(\omega^2,\nu) = \operatorname*{argmin}_{\boldsymbol{\psi}\in\mathbb{R}^n} J(\boldsymbol{\psi}) \ , \tag{2.15}$$

where

$$J(\boldsymbol{\psi}) = \sum_{j=1}^{M} \pi_j \exp\left(-\frac{1}{2\omega^2}\|\boldsymbol{f}_j-\boldsymbol{Y}\|_2^2 + \frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi}-\boldsymbol{f}_j\|_2^2\right) \ , \tag{2.16}$$

and $\boldsymbol{\psi}_X(\omega^2,\nu)$ is the aggregate by Bayesian model averaging with exponentiated least square loss (BMAX).

The below theorem shows that $\boldsymbol{\psi}_X(\omega^2,\nu)$ is optimal both in expectation and in deviation.

Theorem 1. *Assume $\nu \in (0,1)$ and if $\omega^2 \geq \frac{\sigma^2}{\min(\nu,1-\nu)}$, we have oracle inequality for any $\boldsymbol{\lambda} \in \Lambda^M$,*

$$\|\boldsymbol{\psi}_X(\omega^2,\nu)-\boldsymbol{\eta}\|_2^2 \leq \nu\sum_{j=1}^{M}\lambda_j\|\boldsymbol{f}_j-\boldsymbol{\eta}\|_2^2+(1-\nu)\,\|\mathfrak{f}_{\boldsymbol{\lambda}}-\boldsymbol{\eta}\|_2^2+2\omega^2\mathcal{K}(\boldsymbol{\lambda},\boldsymbol{\pi}\delta) \ , \tag{2.17}$$

with probability at least $1-\delta$. Moreover,

$$\mathbb{E}\|\boldsymbol{\psi}_X(\omega^2,\nu)-\boldsymbol{\eta}\|_2^2 \leq \nu\sum_{j=1}^{M}\lambda_j\|\boldsymbol{f}_j-\boldsymbol{\eta}\|_2^2+(1-\nu)\,\|\mathfrak{f}_{\boldsymbol{\lambda}}-\boldsymbol{\eta}\|_2^2+2\omega^2\mathcal{K}(\boldsymbol{\lambda},\boldsymbol{\pi}) \ . \tag{2.18}$$

Our theorem implies that $\boldsymbol{\psi}_X(\omega^2,\nu)$ can compete with an arbitrary $\mathfrak{f}_{\boldsymbol{\lambda}}$ in the convex hull with $\boldsymbol{\lambda}\in\Lambda^M$. However, we are mainly interested in competing with single models, the situation where $\boldsymbol{\lambda}$ is at a vertex of the simplex Λ^M, specifically $\|\boldsymbol{\lambda}\|_0 = 1$. With $\nu\in(0,1)$, the theorem implies that $\boldsymbol{\psi}_X(\omega^2,\nu)$ is deviation optimal unlike the aggregate with exponential weights. This is explicitly stated in the following corollary, which shows that our estimator solves optimally the problem of model averaging. Its proof follows by simply restricting the minimum over Λ^M to the minimum over its vertices in Theorem 1.

Corollary 1. *Under the assumptions of Theorem 1, $\boldsymbol{\psi}_X(\omega^2, \nu)$ satisfies*

$$\|\boldsymbol{\psi}_X(\omega^2,\nu) - \boldsymbol{\eta}\|_2^2 \leq \min_{j\in 1,\ldots,M} \left\{ \|\boldsymbol{f}_j - \boldsymbol{\eta}\|_2^2 + 2\omega^2 \log\left(\frac{1}{\pi_j \delta}\right) \right\} , \tag{2.19}$$

with probability at least $1-\delta$. *Moreover,*

$$\mathbb{E}\|\boldsymbol{\psi}_X(\omega^2,\nu) - \boldsymbol{\eta}\|_2^2 \leq \min_{j\in 1,\ldots,M} \left\{ \|\boldsymbol{f}_j - \boldsymbol{\eta}\|_2^2 + 2\omega^2 \log\left(\frac{1}{\pi_j}\right) \right\} . \tag{2.20}$$

Also it is worthy to point out that the condition $\omega^2 \geq \frac{\sigma^2}{\min(\nu,1-\nu)}$ implies ω^2 is at least greater than $2\sigma^2$ (when $\nu = 1/2$), and the author believes that the inflation of noise is a trade off for tolerance of misspecification of the true η, which is not necessarily included in the candidates dictionary $\mathcal{H}$.

In the Bayesian framework stated above, as loss function $L(\boldsymbol{\psi}, \boldsymbol{\mu})$ changes from least square loss to exponentiated least square loss (2.14), Bayes estimator changes from exponential weighted model averaging estimator which is optimal only in expectation, to BMAX estimator $\boldsymbol{\psi}_X(\omega^2, \nu)$ which is proven to be optimal both in expectation and in deviation. The possible reason for this change is that least square loss only controls the bias, while exponentiated least square loss controls bias and variance at the same time, which can be seen roughly by Taylor expansion

$$\exp\left(\frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi}-\boldsymbol{\mu}\|_2^2\right) = 1 + \frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi}-\boldsymbol{\mu}\|_2^2 + (1/2)\left(\frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi}-\boldsymbol{\mu}\|_2^2\right)^2 + \cdots .$$

It is also natural to extend Theorem 1 from discrete candidates dictionary $\mathcal{H} = \{f_1, \ldots, f_M\}$ to general parameterized dictionary $\mathcal{H}_\Omega = \{f_\gamma : \gamma \in \Omega\}$, denote $\boldsymbol{f}_\gamma = (f_\gamma(x_1), \ldots, f_\gamma(x_n))^\top$.

Corollary 2. *Assume $\nu \in (0,1)$ and Ω is the parameter space, for any distribution $\Theta(\gamma)$ over Ω such that $\mathcal{K}(\Theta,\pi)$ is finite for given prior $\pi(\gamma)$. If $\omega^2 \geq \frac{\sigma^2}{\min(\nu,1-\nu)}$ then we have oracle inequality*

$$\|\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}\|_2^2 \leq \nu \int_\Omega \|\boldsymbol{f}_\gamma-\boldsymbol{\eta}\|_2^2 \Theta(\gamma)\ d\gamma + (1-\nu) \left\| \int_\Omega \boldsymbol{f}_\gamma \Theta(\gamma)\ d\gamma - \boldsymbol{\eta} \right\|_2^2 + 2\omega^2 \mathcal{K}(\Theta, \pi\delta) \tag{2.21}$$

with probability at least $1-\delta$. Moreover,

$$\mathbb{E}\|\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}\|_2^2 \leq \nu \int_\Omega \|\boldsymbol{f}_\gamma-\boldsymbol{\eta}\|_2^2 \Theta(\gamma)\ d\gamma + (1-\nu) \left\| \int_\Omega \boldsymbol{f}_\gamma \Theta(\gamma)\ d\gamma - \boldsymbol{\eta} \right\|_2^2 + 2\omega^2 \mathcal{K}(\Theta, \pi)\ , \tag{2.22}$$

where

$$\hat{\boldsymbol{\eta}} = \underset{\boldsymbol{\psi}\in\mathbb{R}^n}{\operatorname{argmin}} \int_\Omega \exp\left(-\frac{1}{2\omega^2}\|\boldsymbol{f}_\gamma - \boldsymbol{Y}\|_2^2 + \frac{1-\nu}{2\omega^2}\|\boldsymbol{f}_\gamma - \boldsymbol{\psi}\|_2^2\right) \pi(\gamma)\ d\gamma \tag{2.23}$$

Remark 2. *Under this continuous scenario, we can not achieve result parallel to Corollary 1 for competing with single models, because restricting $\Theta(\gamma)$ to mass on a specific point will cause singularity, specifically, $\mathcal{K}(\Theta,\pi) = 1 \cdot \log(1/0) = \infty$.*

2.4 Algorithms to solve BMAX

In last section, we introduced and analyzed the BMAX estimator $\boldsymbol{\psi}_X(\omega^2,\nu)$, which is optimal both in expectation and in deviation to solve the model averaging problem. In this section, we provide two algorithms to approximate the minimizer of $\log J(\boldsymbol{\psi})$, equivalently, $\boldsymbol{\psi}_X(\omega^2,\nu)$.

We propose two algorithms, Greedy Model Averaging (GMA-BMAX) algorithm, and Gradient Descent (GD-BMAX) algorithm. The convergence rates of both algorithms will be shown. Specifically, denote k as the number of iterations in the algorithms, GMA-BMAX algorithm has a converge rate of $O(1/k)$, and GD-BMAX algorithm converges with a geographic rate of $O(q^k)$ for some

$q \in (0,1)$. Oracle inequalities will be shown for the k-th step estimators of both algorithms.

Define condition under which the ℓ_2-norm of $\boldsymbol{f}_j$ is bounded by constant $L \in \mathbb{R}$:

$$\|\boldsymbol{f}_j\|_2 \leq L \ , \quad \forall\, j = 1, \ldots, M \ . \tag{2.24}$$

Given $\nu \in (0,1)$ and $\omega > 0$, define

$$A_1 = \frac{1-\nu}{\omega^2} \ , \tag{2.25}$$

in addition, with L in (2.24), define

$$A_2 = \frac{1-\nu}{\omega^2} + \left(\frac{1-\nu}{\omega^2}\right)^2 L^2 \ , \tag{2.26}$$

and

$$D = \left(\frac{1-\nu}{\omega^2}\right) L^2 + \left(\frac{1-\nu}{\omega^2}\right)^2 L^4 \ . \tag{2.27}$$

Lemma 1 and Lemma 2 are listed below for convenience, and they describe the strong convexity of $\log J(\boldsymbol{\psi})$, similar derivation details can also be found at (e.g., Boyd and Vandenberghe, 2004, Section 9.1.2).

Denote

$$\nabla^2 \log J(\boldsymbol{\psi}) = \frac{\partial^2 \log J(\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^\top}$$

as the Hessian matrix of $\log J(\boldsymbol{\psi})$, then we have the following lemma,

Lemma 1. *For any $\boldsymbol{\psi} \in \mathbb{R}^n$ we have*

$$\nabla^2 \log J(\boldsymbol{\psi}) \geq A_1 \boldsymbol{I}_n \ , \tag{2.28}$$

and if $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M\}$ satisfies condition (2.24)*, then*

$$\nabla^2 \log J(\boldsymbol{\psi}) \leq A_2 \boldsymbol{I}_n \ , \tag{2.29}$$

where A_1 and A_2 are defined as (2.25) *and* (2.26)*.*

The strong convexity of $\log J(\boldsymbol{\psi})$ described by Lemma 1 implies the following lemma, which measure the quantity of $[\log J(\boldsymbol{\psi}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))]$.

Lemma 2. *For any $\boldsymbol{\psi} \in \mathbb{R}^n$ we have the following inequalities*

$$\log J(\boldsymbol{\psi}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu)) \leq \frac{1}{2A_1} \|\nabla \log J(\boldsymbol{\psi})\|_2^2 , \tag{2.30}$$

and if $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M\}$ satisfies condition (2.24)*, then*

$$\log J(\boldsymbol{\psi}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu)) \geq \frac{1}{2A_2} \|\nabla \log J(\boldsymbol{\psi})\|_2^2 , \tag{2.31}$$

where A_1 and A_2 are defined as (2.25) *and* (2.26)*.*

Also note that, since $\log J(\boldsymbol{\psi})$ is strongly convex, its minimizer $\boldsymbol{\psi}_X(\omega^2, \nu)$ is unique.

2.4.1 Greedy Model Averaging Algorithm (GMA-BMAX)

Optimizing convex functions over convex sets is an important topic in modern statistical computing, with many algorithms ranging from gradient descent to interior point (IP) methods (see, e.g., Boyd and Vandenberghe, 2004, for a recent overview). For simple constraints sets such as the simplex Λ^M considered here, so-called *proximal methods* (see, e.g., Beck and Teboulle, 2009) have shown very promising performance, especially when M becomes large. However, the most efficient of these methods (IP and proximal methods) do not output a sparse solution in a general case.

In the sequel, we focus on greedy algorithms introduced into the statistical literature by Jones (1992). In optimization, greedy algorithms over simplex Λ^M are known as *Frank-Wolfe* type (or reduced gradient) methods. Their name refers to the original paper of Frank and Wolfe (1956).

The GMA-BMAX algorithm below can be seen as greedy algorithm that add at most one function from the dictionary $\mathcal{H}$ at each iteration. This feature is attractive as it outputs a k-sparse solution that depends on at most k functions from the dictionary after k iterations. Similar algorithms with the purpose to

```
Algorithm 1 Greedy Model Averaging Algorithm (GMA-BMAX)
Input: Noisy observation $\boldsymbol{Y}$, dictionary $\mathcal{H} = \{f_1, \ldots, f_M\}$, prior $\boldsymbol{\pi} \in \Lambda^M$, parameters $\nu, \omega$.
Output: Aggregate estimator $\boldsymbol{\psi}^{(k)}$.
  Let $\boldsymbol{\psi}^{(0)} = 0$.
  for $k = 1, 2, \ldots$ do
    Set $\alpha_k = \frac{2}{k+1}$
    $J^{(k)} = \operatorname{argmin}_j \log J(\boldsymbol{\psi}^{(k-1)} + \alpha_k(\boldsymbol{f}_j - \boldsymbol{\psi}^{(k-1)}))$
    $\boldsymbol{\psi}^{(k)} = \boldsymbol{\psi}^{(k-1)} + \alpha_k(\boldsymbol{f}_{J^{(k)}} - \boldsymbol{\psi}^{(k-1)})$
  end for
```

solve model averaging has appeared in Dai and Zhang (2011) and Dai et al. (2012).

The following proposition follows from the standard analysis in Frank and Wolfe (1956); Jones (1992); Barron (1993). It shows that the estimator $\boldsymbol{\psi}^{(k)}$ from Algorithm 1 converges to $\boldsymbol{\psi}_X(\omega^2, \nu)$, the solution of BMAX as defined in (2.15).

Proposition 3. *For $\boldsymbol{\psi}^{(k)}$ as defined in Algorithm 1 (GMA-BMAX), if $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M\}$ satisfies condition* (2.24)*, then*

$$\log J(\boldsymbol{\psi}^{(k)}) \leq \log J(\boldsymbol{\psi}_X(\omega^2, \nu)) + \frac{8D}{k+3} \ . \tag{2.32}$$

Proposition 3 states that GMA-BMAX outputs $\boldsymbol{\psi}^{(k)}$ at the k-step, such that $\log J(\boldsymbol{\psi}^{(k)})$ converges with a rate of $O(1/k)$ to $\log J(\boldsymbol{\psi}_X(\omega^2, \nu))$, the minimum of $\log J(\boldsymbol{\psi})$.

Another proposition below describes that the upper bound in the oracle inequality of $\boldsymbol{\psi}^{(k)}$ output from GMA-BMAX converges to that of $\boldsymbol{\psi}_X(\omega^2, \nu)$ in Theorem 1, with a rate of $O(1/\sqrt{k})$.

Proposition 4. *Assume $\nu \in (0,1)$ and if $\omega^2 \geq \frac{\sigma^2}{\min(\nu,1-\nu)}$, we have oracle inequality for any $\boldsymbol{\lambda} \in \Lambda^M$,*

$$\begin{aligned}\|\boldsymbol{\psi}^{(k)} - \boldsymbol{\eta}\|_2^2 \leq \nu \sum_{j=1}^{M} \lambda_j \|\boldsymbol{f}_j - \boldsymbol{\eta}\|_2^2 + (1-\nu)\left\|\mathfrak{f}_{\boldsymbol{\lambda}} - \boldsymbol{\eta}\right\|_2^2 + 2\omega^2 \mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi}\delta) \\ + 2\sqrt{\frac{16D}{A_1(k+3)}}\|\boldsymbol{\psi}_X(\omega^2,\nu) - \boldsymbol{\eta}\|_2 + \frac{16D}{A_1(k+3)},\end{aligned} \tag{2.33}$$

with probability at least $1-\delta$. Moreover,

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\psi}^{(k)} - \boldsymbol{\eta}\|_2^2 \leq \nu \sum_{j=1}^{M} \lambda_j \|\boldsymbol{f}_j - \boldsymbol{\eta}\|_2^2 + (1-\nu)\left\|\mathfrak{f}_{\boldsymbol{\lambda}} - \boldsymbol{\eta}\right\|_2^2 + 2\omega^2 \mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi}) \\ + 2\sqrt{\frac{16D}{A_1(k+3)}}\mathbb{E}\|\boldsymbol{\psi}_X(\omega^2,\nu) - \boldsymbol{\eta}\|_2 + \frac{16D}{A_1(k+3)}.\end{aligned} \tag{2.34}$$

Remark 3. *From Proposition 4, if $\omega^2 \geq \frac{\sigma^2}{\min(\nu,1-\nu)}$, for any $j = 1, \ldots, M$ we have*

$$\mathrm{MSE}(\psi^{(k)}) \leq \mathrm{MSE}(f_j) + 2\omega^2 \log\left(\frac{1}{\pi_j \delta}\right) + O(1/\sqrt{k}) \ ,$$

with probability at least $1-\delta$ and

$$\mathbb{E}\,\mathrm{MSE}(\psi^{(k)}) \leq \mathrm{MSE}(f_j) + 2\omega^2 \log\left(\frac{1}{\pi_j}\right) + O(1/\sqrt{k}) \ .$$

When $k \to \infty$, $\boldsymbol{\psi}^{(k)}$ achieves optimal deviation bound. However, it does not imply optimal deviation bound of $\boldsymbol{\psi}^{(k)}$ for small k ($k < \infty$), while the greedy algorithms described in Dai and Zhang (2011) (GMA) and Dai et al. (2012) (GMA-0 and GMA-0_+) both achieve optimal deviation bound for small $k \geq 2$.

2.4.2 Gradient Descent Algorithm (GD-BMAX)

A Gradient Descent Algorithm (GD-BMAX) is proposed in this section to solve $\boldsymbol{\psi}_X(\omega^2,\nu)$, the unique minimizer of $J(\boldsymbol{\psi})$ in $\mathbb{R}^n$.

Notice that

$$\nabla \log J(\boldsymbol{\psi}^{(k-1)}) = \frac{\nabla J(\boldsymbol{\psi}^{(k-1)})}{J(\boldsymbol{\psi}^{(k-1)})} = \frac{1-\nu}{\omega^2}(\boldsymbol{\psi}^{(k-1)} - \mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}}) \ ,$$

Algorithm 2 Gradient Descent Algorithm (GD-BMAX)

```
Input: Noisy observation $\boldsymbol{Y}$, dictionary $\mathcal{H} = \{f_1, \ldots, f_M\}$, prior $\boldsymbol{\pi} \in \Lambda^M$, parameters $\nu, \omega^2$.
Output: Aggregate estimator $\boldsymbol{\psi}^{(k)}$.
  Let $\boldsymbol{\psi}^{(0)} = 0$.
  for $k = 1, 2, \ldots$ do
    Choose fixed step size $t_k = s \in (0, 2/A_2)$ for $k > 0$.
        $\mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}} = \sum_{j=1}^{M} \lambda_j^{(k-1)} \boldsymbol{f}_j$
    where $\boldsymbol{\lambda}^{(k-1)} \in \Lambda^M$ and
        $\lambda_j^{(k-1)} \propto \pi_j \exp\left(-\frac{1}{2\omega^2}\|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2 + \frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi}^{(k-1)} - \boldsymbol{f}_j\|_2^2\right)$    (2.35)
    and $\mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}}$ can be approximated by Algorithm 3 when $M$ is large.
        $\boldsymbol{\psi}^{(k)} = (1 - t_k\frac{1-\nu}{\omega^2})\boldsymbol{\psi}^{(k-1)} + t_k\frac{1-\nu}{\omega^2}\mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}}$
  end for
```

where $\boldsymbol{\lambda}^{(k-1)} \in \Lambda^M$ is defined as (2.35), it implies that the k-th step update is

$$\boldsymbol{\psi}^{(k)} = (1 - t_k\frac{1-\nu}{\omega^2})\boldsymbol{\psi}^{(k-1)} + t_k\frac{1-\nu}{\omega^2}\mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}} = \boldsymbol{\psi}^{(k-1)} - t_k\nabla \log J(\boldsymbol{\psi}^{(k-1)}) \ ,$$

thus Algorithm 2 is essentially a gradient decent algorithm with step size t_k.

Proposition 5. *For $\boldsymbol{\psi}^{(k)}$ as defined in Algorithm 2 and choose fixed step size $t_k = s \in (0, 2/A_2)$ for $k > 0$, if $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M\}$ satisfies condition* (2.24)*, then*

$$\begin{aligned} &\log J(\boldsymbol{\psi}^{(k)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu)) \\ &\leq [1 - 2A_1(s - (A_2/2)s^2)]^k \left(\log J(\boldsymbol{\psi}^{(0)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))\right) \ . \end{aligned} \tag{2.36}$$

Remark 4. *For the step size t_k, we may choose $t_k = s = 1/A_2$ to minimize the righthand side of* (2.36)*, it follows that*

$$\log J(\boldsymbol{\psi}^{(k)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu)) \leq (1 - A_1/A_2)^k \left(\log J(\boldsymbol{\psi}^{(0)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))\right)$$

Remark 5. *The convergence rate can be improved by the Heavy Ball Method (see, e.g., Poljak, 1987; Nesterov and Nesterov, 2004).*

Proposition 5 states that GD-BMAX outputs $\boldsymbol{\psi}^{(k)}$ at the k-step, such that $\log J(\boldsymbol{\psi}^{(k)})$ converges with a geographic rate to $\log J(\boldsymbol{\psi}_X(\omega^2,\nu))$, the minimum of $\log J(\boldsymbol{\psi})$.

The following proposition describes that the upper bound in the oracle inequality of $\boldsymbol{\psi}^{(k)}$ output from GD-BMAX converges to optimal deviation bound of $\boldsymbol{\psi}_X(\omega^2,\nu)$ in Theorem 1 as when $k\to\infty$, with a geographic rate.

Proposition 6. *Assume $\nu\in(0,1)$ and if $\omega^2\geq\frac{\sigma^2}{\min(\nu,1-\nu)}$, we have oracle inequality for any $\boldsymbol{\lambda}\in\Lambda^M$,*

$$\begin{aligned}\|\boldsymbol{\psi}^{(k)}-\boldsymbol{\eta}\|_2^2\leq\nu\sum_{j=1}^{M}\lambda_j\|\boldsymbol{f}_j-\boldsymbol{\eta}\|_2^2+(1-\nu)\,\|\mathfrak{f}_{\boldsymbol{\lambda}}-\boldsymbol{\eta}\|_2^2+2\omega^2\mathcal{K}(\boldsymbol{\lambda},\boldsymbol{\pi})+2\omega^2\log(1/\delta)\\ +2\sqrt{L^2[1-2A_1(s-(A_2/2)s^2)]^k}\|\boldsymbol{\psi}_X(\omega^2,\nu)-\boldsymbol{\eta}\|_2\\ +L^2[1-2A_1(s-(A_2/2)s^2)]^k\ ,\end{aligned}\tag{2.37}$$

with probability at least $1-\delta$. Moreover,

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\psi}^{(k)}-\boldsymbol{\eta}\|_2^2\leq\nu\sum_{j=1}^{M}\lambda_j\|\boldsymbol{f}_j-\boldsymbol{\eta}\|_2^2+(1-\nu)\,\|\mathfrak{f}_{\boldsymbol{\lambda}}-\boldsymbol{\eta}\|_2^2+2\omega^2\mathcal{K}(\boldsymbol{\lambda},\boldsymbol{\pi})\\ +2\sqrt{L^2[1-2A_1(s-(A_2/2)s^2)]^k}\mathbb{E}\|\boldsymbol{\psi}_X(\omega^2,\nu)-\boldsymbol{\eta}\|_2\\ +L^2[1-2A_1(s-(A_2/2)s^2)]^k\ .\end{aligned}\tag{2.38}$$

Remark 6. *From Proposition 6, if $\omega^2\geq\frac{\sigma^2}{\min(\nu,1-\nu)}$, for any $j=1,\ldots,M$ we have*

$$\mathrm{MSE}(\psi^{(k)})\leq\mathrm{MSE}(f_j)+2\omega^2\log\left(\frac{1}{\pi_j\delta}\right)+O(q^k)\ ,$$

with probability at least $1-\delta$ and

$$\mathbb{E}\,\mathrm{MSE}(\psi^{(k)})\leq\mathrm{MSE}(f_j)+2\omega^2\log\left(\frac{1}{\pi_j}\right)+O(q^k)\ ,$$

for some constant $q\in(0,1)$.

Though the GD-BMAX algorithm does not give sparse output as GMA-BMAX algorithm, it has a faster geographic convergence rate than $O(1/k)$ of GMA-BMAX. Like Proposition 4, the results in Proposition 6 do not imply optimal deviation bounds of $\boldsymbol{\psi}^{(k)}$ for small k ($k < \infty$). Later in Section 2.5, a greedy algorithm GMA-0 is to solve Q-aggregation (with linear entropy), GMA-0 not only outputs sparse estimator like GMA-BMAX algorithm, but also it has optimal deviation bound after small iterations ($k \geq 2$). Yet it does have limitations, which will be discussed as well.

When M is large, it is not practical to directly calculate $\boldsymbol{\lambda}^{(k-1)} \in \Lambda^M$ in GD-BMAX algorithm with formulation

$$\lambda_j^{(k-1)} \propto \pi_j \exp\left(-\frac{1}{2\omega^2}\|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2 + \frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi}^{(k-1)} - \boldsymbol{f}_j\|_2^2\right) ,$$

instead, we can apply Metropolis-Hastings (MH) sampler to approximate $\boldsymbol{\lambda}^{(k-1)}$ for the k-th iteration in Algorithm 2. For a basic introduction to Monte Carlo methods and Metropolis-Hastings sampler, see e.g. Marin and Robert (2007).

The MH algorithm stated above is an approach to approximate $\mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}}$ with $\boldsymbol{u}_T^{(k-1)}$, it results that the sequence $\{\boldsymbol{\psi}^{(k)}\}$ in the GD-BMAX algorithm will have perturbations.

Below we give a simple proposition describing how the perturbations from approximating $\mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}}$ would influence convergence of sequence $\{\log J(\boldsymbol{\psi}^{(k)})\}$ to $\log J(\boldsymbol{\psi}_X(\omega^2, \nu))$.

Proposition 7. *Given $\boldsymbol{Y} \in \mathbb{R}^n$, for all $k > 0$, we assume $\boldsymbol{u}_T^{(k-1)}$ from Algorithm 3 satisfying the following:*

$$\mathbb{E}[\boldsymbol{u}_T^{(k-1)}|\boldsymbol{\psi}^{(k-1)}] = \mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}} \tag{2.39}$$

$$\|COV[\boldsymbol{u}_T^{(k-1)}|\boldsymbol{\psi}^{(k-1)}]\|_{op} \leq s^2 \tag{2.40}$$

Algorithm 3 Metropolis-Hastings (MH) Sampler for estimating $\mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}}$ at k-th step in Algorithm 2

```
Input: Noisy observation $\boldsymbol{Y}$, dictionary $\mathcal{H} = \{f_1, \ldots, f_M\}$, prior $\boldsymbol{\pi} \in \Lambda^M$, parameters $\nu, \omega^2$, $(k-1)$-th step estimator $\boldsymbol{\psi}^{(k-1)}$.
Output: $\boldsymbol{u}_T^{(k-1)}$ as estimator of $\mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}} = \sum_{j=1}^{M} \lambda_j^{(k-1)} \boldsymbol{f}_j$.
  Initialize $j(0) = 0$.
  for $t = 1, \cdots, T_0 + T$ do
    Generate $\tilde{j} \sim q(\cdot | j(t-1))$.
    Compute
      $\rho(j(t-1), \tilde{j}) = \min\left( \frac{q(j(t-1)|\tilde{j})\theta(\tilde{j})}{q(\tilde{j}|j(t-1))\theta(j(t-1))}, 1 \right)$ ,
    where
      $\theta(j) = \pi_j \exp\left( -\frac{1}{2\omega^2} \|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2 + \frac{1-\nu}{2\omega^2} \|\boldsymbol{\psi}^{(k-1)} - \boldsymbol{f}_j\|_2^2 \right)$.
    Generate a random variable
      $j(t) = \begin{cases} \tilde{j} \ , & \text{with probability} \quad \rho(j(t-1), \tilde{j}) \\ j(t-1) \ , & \text{with probability} \quad 1 - \rho(j(t-1), \tilde{j}) \end{cases}$
  end for
  Calculate
      $\boldsymbol{u}_T^{(k-1)} = \frac{1}{T} \sum_{t=T_0+1}^{T_0+T} \boldsymbol{f}_{j(t)}$.
```

where $\|\cdot\|_{op}$ *is matrix spectral norm. Then we have*

$$\begin{aligned}&\mathbb{E}\left(\log J(\boldsymbol{\psi}^{(k)}) - \log J(\boldsymbol{\psi}_X(\omega^2,\nu))\right)\\ \leq\;& [1-2A_1(s-(A_2/2)s^2)]^k\left(\log J(\boldsymbol{\psi}^{(0)}) - \log J(\boldsymbol{\psi}_X(\omega^2,\nu))\right) + A_1 n s^2/2\;.\end{aligned}$$

2.5 Q-aggregation: A Dual Representation of Aggregate by BMAX

In this section, we will firstly introduce another model averaging method, Q-*aggregation*, which was proposed in Dai et al. (2012); then we will show that Q-aggregation (with Kullback-Leibler entropy) is essentially a dual representation of the aggregate by BMAX as defined in (2.15) and (2.16); finally GMA-0 algorithm and its deviation optimality originally proved in Dai et al. (2012) are listed for comparisons with GMA-BMAX and GD-BMAX, both in theory and via numerical experiments in the next section.

Given $\boldsymbol{Y}$ and $\{\boldsymbol{f}_1,\ldots,\boldsymbol{f}_M\}$, Q-aggregation $\mathfrak{f}_{\boldsymbol{\lambda}^Q}$ is defined as following:

$$\mathfrak{f}_{\boldsymbol{\lambda}^Q} = \sum_{j=1}^{M}\lambda_j^Q \boldsymbol{f}_j\;, \tag{2.41}$$

where $\boldsymbol{\lambda}^Q = (\lambda_1^Q,\ldots,\lambda_M^Q)^\top \in \Lambda^M$ such that

$$\boldsymbol{\lambda}^Q \in \operatorname*{argmin}_{\boldsymbol{\lambda}\in\Lambda^M} Q(\boldsymbol{\lambda})\;, \tag{2.42}$$

and

$$Q(\boldsymbol{\lambda}) = \|\boldsymbol{f}_{\boldsymbol{\lambda}} - \boldsymbol{Y}\|_2^2 + \nu\sum_{j=1}^{M}\lambda_j\|\boldsymbol{f}_j - \boldsymbol{f}_{\boldsymbol{\lambda}}\|_2^2 + 2\omega^2\mathcal{K}_\rho(\boldsymbol{\lambda},\boldsymbol{\pi})\;, \tag{2.43}$$

for some $\nu\in(0,1)$. $\mathcal{K}_\rho(\boldsymbol{\lambda},\boldsymbol{\pi})$ is defined as

$$\mathcal{K}_\rho(\boldsymbol{\lambda},\boldsymbol{\pi}) = \sum_{j=1}^{M}\lambda_j\log\left(\frac{\rho(\lambda_j)}{\pi_j}\right)\;, \tag{2.44}$$

where ρ is a real valued function on $[0,1]$ satisfying

$$\begin{aligned} \rho(t) &\geq t\ , \\ t\log\rho(t) &\text{ is convex}\ . \end{aligned} \tag{2.45}$$

When $\rho(t) = t$, $\mathcal{K}_\rho(\boldsymbol{\lambda},\boldsymbol{\pi})$ becomes $\mathcal{K}(\boldsymbol{\lambda},\boldsymbol{\pi})$, the Kullback-Leibler entropy. When $\rho(t) = 1$, $\mathcal{K}_\rho(\boldsymbol{\lambda},\boldsymbol{\pi}) = \sum_{j=1}^M \lambda_j \log(1/\pi_j)$, a linear entropy in Λ^M, especially, penalty of $\mathcal{K}_\rho(\boldsymbol{\lambda},\boldsymbol{\pi})$ in (2.43) vanishes when $\boldsymbol{\pi}$ is a flat prior.

To build duality, first define function $T : \mathbb{R}^n \to \mathbb{R}$ as

$$T(\boldsymbol{h}) = -\frac{\nu}{1-\nu}\|\boldsymbol{h}-\boldsymbol{Y}\|_2^2 - 2\omega^2 \log\left(\sum_{j=1}^M \pi_j \exp\left(-\frac{\nu}{2\omega^2}\|\boldsymbol{f}_j - \boldsymbol{h}\|_2^2\right)\right)\ , \tag{2.46}$$

and denote the maximizer of $T(\boldsymbol{h})$ as

$$\hat{\boldsymbol{h}} = \operatorname*{argmax}_{\boldsymbol{h}\in\mathbb{R}^n} T(\boldsymbol{h})\ . \tag{2.47}$$

Define function $S : \Lambda^M \times \mathbb{R}^n \to \mathbb{R}$ as

$$S(\boldsymbol{\lambda},\boldsymbol{h}) = -\frac{\nu}{1-\nu}\|\boldsymbol{h}-\boldsymbol{Y}\|_2^2 + \nu\sum_{j=1}^M \lambda_j\|\boldsymbol{f}_j - \boldsymbol{h}\|_2^2 + 2\omega^2\mathcal{K}(\boldsymbol{\lambda},\boldsymbol{\pi})\ . \tag{2.48}$$

Define two hyper planes A and B in $\Lambda^M \times \mathbb{R}^n$ as

$$\begin{aligned} A &= \left\{(\boldsymbol{\lambda},\boldsymbol{h}) \in \Lambda^M \times \mathbb{R}^n : \boldsymbol{h} = \frac{1}{\nu}\boldsymbol{Y} - \frac{1-\nu}{\nu}\mathfrak{f}_{\boldsymbol{\lambda}}\right\}\ , \\ B &= \left\{(\boldsymbol{\lambda},\boldsymbol{h}) \in \Lambda^M \times \mathbb{R}^n : \lambda_j = \frac{\exp\left(-\frac{\nu}{2\omega^2}\|\boldsymbol{f}_j - \boldsymbol{h}\|_2^2\right)\pi_j}{\sum_{i=1}^M \exp\left(-\frac{\nu}{2\omega^2}\|\boldsymbol{f}_i - \boldsymbol{h}\|_2^2\right)\pi_i}\right\}\ . \end{aligned} \tag{2.49}$$

The following lemma states the relationship between $\hat{\boldsymbol{h}}$ and $\mathfrak{f}_{\boldsymbol{\lambda}^Q}$:

Lemma 3. *When $\rho(t) = t$, with all above definitions, we have the following*

$$\min_{\boldsymbol{\lambda}\in\Lambda^M} Q(\boldsymbol{\lambda}) = \min_{\boldsymbol{\lambda}\in\Lambda^M}\max_{\boldsymbol{h}\in\mathbb{R}^n} S(\boldsymbol{\lambda},\boldsymbol{h}) = \max_{\boldsymbol{h}\in\mathbb{R}^n}\min_{\boldsymbol{\lambda}\in\Lambda^M} S(\boldsymbol{\lambda},\boldsymbol{h}) = \max_{\boldsymbol{h}\in\mathbb{R}^n} T(\boldsymbol{h}),$$

moreover, $A \cap B = \left\{(\boldsymbol{\lambda}^Q, \hat{\boldsymbol{h}})\right\}$.

Lemma 3 states that, $(\boldsymbol{\lambda}^Q, \hat{\boldsymbol{h}})$ is the joint of hyper planes A and B, and the saddle point of function $S(\boldsymbol{\lambda}, \boldsymbol{h})$ over space $\Lambda^M \times \mathbb{R}^n$.

With $T(\boldsymbol{h})$ defined as (2.46), make the transformation $\boldsymbol{h} = \frac{1}{\nu}\boldsymbol{Y} - \frac{1-\nu}{\nu}\boldsymbol{\psi}$ then it is easy to verify that

$$T(\boldsymbol{h}) = -2\omega^2 \log\left(J(\boldsymbol{\psi})\right) \ , \tag{2.50}$$

where $J(\boldsymbol{\psi})$ is defined as (2.16).

So maximizing $T(\boldsymbol{h})$ is equivalent to minimizing $J(\boldsymbol{\psi})$, thus

$$\hat{\boldsymbol{h}} = \frac{1}{\nu}\boldsymbol{Y} - \frac{1-\nu}{\nu}\boldsymbol{\psi}_X(\omega^2, \nu) \ ,$$

combine it with

$$\hat{\boldsymbol{h}} = \frac{1}{\nu}\boldsymbol{Y} - \frac{1-\nu}{\nu}\mathfrak{f}_{\boldsymbol{\lambda}^Q} \ ,$$

from Lemma 3, it follows that $\boldsymbol{\psi}_X(\omega^2, \nu) = \mathfrak{f}_{\boldsymbol{\lambda}^Q}$, thus we have

Theorem 2. *When $\rho(t) = t$,*

$$\boldsymbol{\psi}_X(\omega^2, \nu) = \mathfrak{f}_{\boldsymbol{\lambda}^Q} \ ,$$

where $\boldsymbol{\psi}_X(\omega^2, \nu)$ is defined by (2.15) *and* (2.16)*, and $\mathfrak{f}_{\boldsymbol{\lambda}^Q}$ is defined by* (2.41),(2.42) *and* (2.43)*.*

Below we also include a corollary from Lemma 3 and Theorem 2, and it describes an sufficient and necessary condition of $\boldsymbol{\lambda}^Q$,

Corollary 3. *When $\rho(t) = t$, $\tilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda}^Q$ if and only if*

$$\tilde{\lambda}_j = \frac{\exp\left(\left(-\|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2 + (1-\nu)\|\mathfrak{f}_{\tilde{\boldsymbol{\lambda}}} - \boldsymbol{f}_j\|_2^2\right)/2\omega^2\right)\pi_j}{\sum_{i=1}^M \exp\left(\left(-\|\boldsymbol{f}_i - \boldsymbol{Y}\|_2^2 + (1-\nu)\|\mathfrak{f}_{\tilde{\boldsymbol{\lambda}}} - \boldsymbol{f}_i\|_2^2\right)/2\omega^2\right)\pi_i} \ . \tag{2.51}$$

Proof. (Necessity) Since $\boldsymbol{\psi}_X(\omega^2, \nu) = \mathfrak{f}_{\boldsymbol{\lambda}^Q}$ from Theorem 2, then let $\frac{\partial J(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} = 0$, (2.51) is obtained.

(Sufficiency) If $\tilde{\boldsymbol{\lambda}}$ satisfies condition (2.51), it is solution to $\boldsymbol{\lambda}$ in $A \cap B$, which has the unique point $(\boldsymbol{\lambda}^Q, \hat{\boldsymbol{h}})$ from Lemma 3, resulting that $\tilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda}^Q$. □

Theorem 2 states that, when $\rho(t) = t$, $\mathcal{K}_\rho(\boldsymbol{\lambda}, \boldsymbol{\pi})$ becomes Kullback-Leibler entropy, Q-aggregation (with Kullback-Leibler entropy) $\mathfrak{f}_{\boldsymbol{\lambda}^Q}$ is essentially a dual representation of $\boldsymbol{\psi}_X(\omega^2, \nu)$, the aggregate by BMAX, and apparently it follows that, $\mathfrak{f}_{\boldsymbol{\lambda}^Q}$ shares the same expectation and deviation optimality as $\boldsymbol{\psi}_X(\omega^2, \nu)$ in solving the model averaging problem, and this matches the results on optimality of $\mathfrak{f}_{\boldsymbol{\lambda}^Q}$ by Theorem 3.1 in Dai et al. (2012), where it is proved with general $\mathcal{K}_\rho(\boldsymbol{\lambda}, \boldsymbol{\pi})$ with $\rho(t)$ satisfying condition (2.45). Theorem 3.1 of Dai et al. (2012) is listed below without proof for convenience.

Theorem 3. *Assume* $\nu \in (0,1)$ *and if* $\omega^2 \geq \frac{\sigma^2}{\min(\nu, 1-\nu)}$, *for* $\mathfrak{f}_{\boldsymbol{\lambda}^Q}$ *as defined in* (2.41),(2.42) *and* (2.43) *with* $\rho(t)$ *satisfying condition* (2.45), *we have oracle inequality for any* $\boldsymbol{\lambda} \in \Lambda^M$,

$$\|\mathfrak{f}_{\boldsymbol{\lambda}^Q} - \boldsymbol{\eta}\|_2^2 \leq \nu \sum_{j=1}^{M} \lambda_j \|\boldsymbol{f}_j - \boldsymbol{\eta}\|_2^2 + (1-\nu)\|\mathfrak{f}_{\boldsymbol{\lambda}} - \boldsymbol{\eta}\|_2^2 + 2\omega^2 \mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi}\delta) \,, \tag{2.52}$$

with probability at least $1 - \delta$. *Moreover,*

$$\mathbb{E}\|\mathfrak{f}_{\boldsymbol{\lambda}^Q} - \boldsymbol{\eta}\|_2^2 \leq \nu \sum_{j=1}^{M} \lambda_j \|\boldsymbol{f}_j - \boldsymbol{\eta}\|_2^2 + (1-\nu)\|\mathfrak{f}_{\boldsymbol{\lambda}} - \boldsymbol{\eta}\|_2^2 + 2\omega^2 \mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi}) \,. \tag{2.53}$$

GMA-0 is an algorithm proposed in Dai et al. (2012) as a greedy approach to solve Q-aggregation with linear entropy ($\mathcal{K}_\rho(\boldsymbol{\lambda}, \boldsymbol{\pi}) = \sum_{j=1}^{M} \lambda_j \log(1/\pi_j)$ when $\rho(t) = 1$). For comparison purpose, we list GMA-0 algorithm below.

In GMA-0 algorithm, $\mathsf{e}^{(j)}$ denotes the jth vector of the canonical basis of $\mathbb{R}^M$. Similar to GMA-BMAX, it is a greedy algorithm that add at most one function from the dictionary at each iteration. It outputs a k-sparse solution that depends on at most k functions from the dictionary after k iterations. Moreover, GMA-0 leads to sparse estimators that achieve the optimal deviation bounds for small $k \geq 2$, while the estimators from GMA-BMAX (sparse) and GD-BMAX (dense) only have such bounds when $k \to \infty$.

Below we show the convergence rate of GMA-0 algorithm is $O(1/k)$,

Algorithm 4 GMA-0 Algorithm

```
Input: Noisy observation $\boldsymbol{Y}$, dictionary $\mathcal{H} = \{f_1, \ldots, f_M\}$, prior $\boldsymbol{\pi} \in \Lambda^M$, parameters $\nu, \beta$.
Output: Aggregate estimator $\mathfrak{f}_{\boldsymbol{\lambda}^{(k)}}$.
  Let $\boldsymbol{\lambda}^{(0)} = 0$, $\mathfrak{f}_{\boldsymbol{\lambda}^{(0)}} = 0$.
  for $k = 1, 2, \ldots$ do
    Set $\alpha_k = \frac{2}{k+1}$
    $J^{(k)} = \operatorname{argmin}_j Q(\boldsymbol{\lambda}^{(k-1)} + \alpha_k(\mathsf{e}^{(j)} - \boldsymbol{\lambda}^{(k-1)}))$
    $\boldsymbol{\lambda}^{(k)} = \boldsymbol{\lambda}^{(k-1)} + \alpha_k(\mathsf{e}^{(J^{(k)})} - \boldsymbol{\lambda}^{(k-1)})$
  end for
```

Proposition 8. *When $\rho(t) = 1$, $\boldsymbol{\lambda}^{(k)}$ is output from GMA-0, for any $\boldsymbol{\lambda} \in \Lambda^M$, for $k \geq 1$ it holds that*

$$Q(\boldsymbol{\lambda}^{(k)}) \leq Q(\boldsymbol{\lambda}) + \frac{4(1-\nu)}{k+3} \sum_{j=1}^{M} \lambda_j \|\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2 \,.$$

For $k \geq 2$, $\mathfrak{f}_{\boldsymbol{\lambda}^{(k)}}$ achieves optimal deviation.

Theorem 4. *Fix $\nu \in (0,1), k \geq 2$ and $\boldsymbol{\pi} \in \Lambda^M$. Take*

$$\omega^2 \geq \sigma^2 \inf_{\theta \in (0,1]} \max \left\{ \frac{1}{\nu - \frac{4(1-\nu)(1-\theta)}{(k+3)\theta}} ; \frac{1}{(1-\theta)(1-\nu)(1-\frac{4}{k+3})} \right\} ,$$

then $\mathfrak{f}_{\boldsymbol{\lambda}^{(k)}}$ with $\boldsymbol{\lambda}^{(k)}$ output by GMA-0 satisfies

$$\|\mathfrak{f}_{\boldsymbol{\lambda}^{(k)}} - \boldsymbol{\eta}\|_2^2 \leq \min_j \left\{ \|\boldsymbol{f}_j - \boldsymbol{\eta}\|_2^2 + 2\omega^2 \log\left(\frac{1}{\pi_j \delta}\right) \right\} ,$$

with probability $1 - \delta$. Moreover,

$$\mathbb{E}\|\mathfrak{f}_{\boldsymbol{\lambda}^{(k)}} - \boldsymbol{\eta}\|_2^2 \leq \min_j \left\{ \|\boldsymbol{f}_j - \boldsymbol{\eta}\|_2^2 + 2\omega^2 \log\left(\frac{1}{\pi_j}\right) \right\} .$$

Both the proof of Proposition (8) and Theorem 4 can be found in Dai et al. (2012) (Theorem 4.1 and 4.2).

To get a better quantitative idea of the result, we illustrate the particular choice $\nu = 1/2$. In this case, it can be easily shown that the optimal θ is given by $\theta_k^\star = 2/(\sqrt{k+3}+2)$. Therefore, in this case, one may take

$$\omega^2 \geq \frac{2\sigma^2}{1 - 2/\sqrt{k+3}} \, .$$

In particular, for $k = 2$, it is sufficient to take $\omega^2 = 2(5+2\sqrt{5})\sigma^2 \geq 19\sigma^2$. Although it achieves the optimal rate for model averaging, the large constant implies that it is may still be beneficial to run the algorithm for more than two iterations.

It is worth pointing out that with flat prior, the first stage estimator $\mathsf{f}_{\boldsymbol{\lambda}^{(1)}} = f_{\hat{j}}$ is simply the empirical risk minimizer with $\hat{j} \in \operatorname{argmin}_j \widehat{\mathrm{MSE}}(f_j)$. We have already pointed out that this estimator achieves suboptimal deviation bounds; therefore the requirement of $k \geq 2$ in our analysis is natural.

Remark 7. *Theorem 4 implies deviation bounds of the optimal order for all $k \geq 2$, and the constant ω^2 decreases to $\sigma^2/\min(\nu, 1-\nu)$ as in Theorem 3 when $k \to \infty$. Such results indicate that the choice of ν is not critical and any positive constant leads to the same optimal bound. However, we can optimize the constant by choosing $\nu = 1/2$ which can be used in the simulations.*

Remark 8. *GMA-0 algorithm uses only zero order information, namely, the coordinate that minimizes the objective value $Q(\cdot)$ (which is relatively uncommon in the greedy algorithm literature), instead, the standard Frank-Wolfe procedure in the greedy algorithm literature uses first order information, namely the gradient ∇Q, to pick the best coordinate $J^{(k)}$. Mathematically,*

$$J^{(k)} = \operatorname*{argmin}_j \left(\nabla Q(\lambda^{(k)})\right)_j \, ,$$

then from the classical greedy algorithm analysis in (Frank and Wolfe, 1956; Jones, 1992; Barron, 1993) we could still get similar results as Proposition 4 of GMA-BMAX algorithm under condition (2.24), and $\lambda^{(k)}$ is well-known in the literature (also see surveys Clarkson, 2008; Jaggi, 2011).

Remark 9. *At iteration k in GMA-0 algorithm, we could take a more aggressive optimization step to update $\boldsymbol{\lambda}^{(k)}$ given $\{J^{(1)}, \ldots, J^{(k)}\}$, specifically,*

$$\boldsymbol{\lambda}^{(k)} = \operatorname*{argmin}_{\boldsymbol{\lambda} \in \Lambda^M} Q(\boldsymbol{\lambda}) \ ,$$

s.t. $\lambda_j = 0$ for $j \notin \{J^{(1)}, \ldots, J^{(k)}\}$. This kind of additional optimization is referred to as fully-corrective *step (Shalev-Shwartz et al., 2010), which is known to improve performance in practice. And apparently this full-corrective outputs also share the deviation optimality as GMA-0 in Theorem 4.*

Note that when we choose flat prior $\boldsymbol{\pi}$, the choice of $J^{(k)}$ in GMA-0 algorithm can be further simplified to

$$J^{(k)} = \operatorname*{argmin}_{j} \left\{ \|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2 - (1-\nu)(1-\alpha_k)\|\mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}} - \boldsymbol{f}_j\|_2^2 + 2\omega^2 \ln(1/\lambda_j) \right\} \ , \tag{2.54}$$

which can be interpreted as at each iteration of GMA-0 algorithm, estimator $\boldsymbol{f}_j$ is preferred to other candidates if it is closer to $\boldsymbol{Y}$ and has less correlated with current aggregate estimator $\mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}}$ (i.e. we want $\|\mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}} - \boldsymbol{f}_j\|_2^2$ be large while $\|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2$ being small).

With equation

$$\begin{aligned} & \|(1-\alpha_k)\mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}} + \alpha_k \boldsymbol{f}_j - \boldsymbol{Y}\|_2^2 \\ = \ & (1-\alpha_k)\|\mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}} - \boldsymbol{Y}\|_2^2 + \alpha_k \|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2 - \alpha_k(1-\alpha_k)\|\mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}} - \boldsymbol{f}_j\|_2^2 \ , \end{aligned}$$

reorganize equation (2.54) and $J^{(k)}$ can be equivalently chosen by

$$\begin{aligned} J^{(k)} = \operatorname*{argmin}_{j} \{ \ & \|(1-\alpha_k)\mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}} + \alpha_k \boldsymbol{f}_j - \boldsymbol{Y}\|_2^2 + \nu\alpha_k(1-\alpha_k)\|\mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}} - \boldsymbol{f}_j\|_2^2 \\ & + 2\alpha_k\omega^2 \ln(1/\lambda_j) \} \ , \end{aligned} \tag{2.55}$$

which implies that GMA-0 algorithm is a cleaner version of GMA algorithm in Dai and Zhang (2011), and they all share the deviation and expectation optimality. The advantage of GMA-0 compared to algorithms GMA-BMAX and GD-BMAX are: it not only outputs spare estimators which hold optimality for small $k \geq 2$, it also becomes parameter free if we choose flat prior as $\boldsymbol{\pi} = (1/M, \ldots, 1/M)^\top$ (penalty term $2\omega^2 \ln(1/\lambda_j)$ in (2.54) vanishes) and set $\nu = 1/2$ as default, while GMA-BMAX and GD-BMAX algorithms need to tune ω for practical applications.

In addition, with flat prior $\boldsymbol{\pi}$ when $k = 2$, GMA-0 is different from STAR algorithm (defined as (1.21), (1.22) and (1.23)) with an additional term $\nu\alpha_2(1 - \alpha_2)\|\mathfrak{f}_{\boldsymbol{\lambda}^{(1)}} - \boldsymbol{f}_j\|_2^2$ where $\alpha_2 = 1/2$, penalizing the variance comes from aggregation.

2.6 Numerical Experiments

The purpose of this section is to illustrate the advantages of using BMAX estimators by numerical examples. We focus on the average performance of different algorithms and configurations.

2.6.1 Model Setup

We identify a function $\boldsymbol{f}$ with a vector $(f(x_1), \ldots, f(x_n))^\top \in \mathbb{R}^n$. Define $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M$ so that the $n \times M$ design matrix $\boldsymbol{X} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M)$ has i.i.d standard Gaussian entries. Let $\boldsymbol{I}_n$ denote the identity matrix of $\mathbb{R}^n$ and let $\Delta \sim \mathcal{N}(0, \boldsymbol{I}_n)$ be a random vector. The regression function is defined by $\boldsymbol{\eta} = \boldsymbol{f}_1 + 0.5\Delta$. Note that typically $\boldsymbol{f}_1$ will be the closest function to $\boldsymbol{\eta}$ but not necessarily. The noise vector $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$ is drawn independently of $\boldsymbol{X}$ where $\sigma = 2$.

We define the oracle model (OM) f_{k^*}, where $k^* = \operatorname{argmin}_j \mathrm{MSE}(f_j)$. The model f_{k^*} is clearly not a valid estimator because it depends on the unobserved η, however it can be used as a performance benchmark. The performance difference between an estimator $\hat{\eta}$ and the oracle model f_{k^*} is measured by the *regret* defined as:

$$R(\hat{\eta}) = \mathrm{MSE}(\hat{\eta}) - \mathrm{MSE}(f_{k^*}) \; . \tag{2.56}$$

Since the target is $\boldsymbol{\eta} = \boldsymbol{f}_1 + 0.5\Delta$, and $\boldsymbol{f}_1$ and Δ are random Gaussian vectors, the oracle model is likely $\boldsymbol{f}_1$ (but it may not be $\boldsymbol{f}_1$ due to the misspecification vector Δ). The noise $\sigma = 2$ is relatively large, which implies a situation where the best convex aggregation does not outperform the oracle model. This is the scenario we considered here. For simplicity, all algorithms use a flat prior $\pi_j = 1/M$ for all j.

The experiment is performed with the parameters $n = 50$, $M = 200$, and $\sigma = 2$, and repeated for 500 replications.

2.6.2 Comparative Results among Different Models

GMA-BMAX and GD-BMAX algorithms are provided to solve BMAX. Q-aggregation (with Kullback-Leibler entropy) is a dual representation of the aggregate by BMAX , while GMA-0 algorithm is a greedy approach to solve Q-aggregation (with linear entropy), thus GMA-0 is included for comparison purpose and it is parameter free with flat prior and $\nu = 1/2$ fixed. From the definition of $Q(\boldsymbol{\lambda})$ (2.43), it is easy to see that, the minimizer of $Q(\boldsymbol{\lambda})$ (when $\rho(t) = 1$ with flat prior) becomes $\boldsymbol{\lambda}^{\text{PROJ}}$ in (2.9) by setting $\nu = 0$, so $\boldsymbol{\lambda}^{\text{PROJ}}$ is approximated by GMA-0 with $\nu = 0$ by running 200 iterations, and the projection algorithm is denoted by "PROJ". GMA-BMAX, GD-BMAX and GMA-0 are run for K iterations up to $K = 150$, with $\nu = 1/2$ (this choice theoretically optimize upper bound of the oracle inequality (2.17),(2.18)), parameter ω for GMA-BMAX, GD-BMAX is chosen as $\omega^2 = \sigma^2/5$, and parameter ω for exponential weighted model averaging (denoted by "EWMA") is tuned by ten fold cross validation. STAR estimator is also included. Regrets (2.56) of all algorithms are reported for comparisons.

Results are composed in two forms: table (Regrets of STAR, EWMA, PROJ, Regrets versus iterations for GMA-BMAX, GD-BMAX, GMA-0), and figure (Regrets vs iterations for GMA-BMAX, GD-BMAX, GMA-0).

Table 2.1 is a comparison of commonly used estimators (STAR, EWMA and

Table 2.1: Performance Comparison

STAR	**EWMA**	**PROJ**
0.458 ± 0.44	0.435 ± 0.5	0.425 ± 0.3

	$k = 1$	$k = 5$	$k = 15$	$k = 60$	$k = 100$	$k = 150$
GMA-BMAX	0.687 ± 0.72	0.493 ± 0.43	0.417 ± 0.38	0.376 ± 0.37	0.37 ± 0.37	0.368 ± 0.38
GD-BMAX	0.974 ± 0.23	0.873 ± 0.21	0.69 ± 0.2	0.415 ± 0.33	0.376 ± 0.36	0.368 ± 0.38
GMA-0	0.549 ± 0.78	0.395 ± 0.45	0.373 ± 0.41	0.368 ± 0.4	0.369 ± 0.41	0.368 ± 0.4

PROJ) with GMA-BMAX, GD-BMAX, GMA-0. The regrets are reported using the "mean ± standard deviation" format.

The results in Table 2.1 indicate that GMA-BMAX, GD-BMAX and GMA-0 perform better as iteration k increases, and all three algorithms beat STAR, EWMA and PROJ when k is large enough. GMA-BMAX beats when $k = 15$ and GD-BMAX beats when $k = 60$. This does not conflict with Proposition 4 and Proposition 6 which state that GD-BMAX has faster convergence rate than GMA-BMAX, because they start from different initial points, and if we calculate the total decrements after $k = 60$ iterations for them, the regret of GMA-BMAX decreases by $0.687 - 0.376 = 0.311$ while that of GD-BMAX decreases by $0.974 - 0.415 = 0.559$, we can see the performance of GD-BMAX actually improves faster than that of GMA-BMAX. GMA-0 beats STAR, EWMA and PROJ after as small as $k = 5$ iterations, which still gives a relatively sparse averaged model. This is consistent with Theorem 4 which states that GMA-0 has optimal bounds for small k ($k \geq 2$), we can also see that in order to achieve good performance, it is necessary to use more iterations than $k = 2$ (although this does not change the $O(1/n)$ rate for the regret, it can significantly reduce the constant).

Figure 2.1 compares the MSE performance of GMA-BMAX, GD-BMAX and GMA-0 with $\nu = 1/2$. GMA-0 is parameter free when set $\nu = 1/2$ with flat prior, while GMA-BMAX and GD-BMAX need to choose some proper ω^2 (in this experiment, we simply set $\omega^2 = \sigma^2/5$), after large iterations ($k = 100$), they

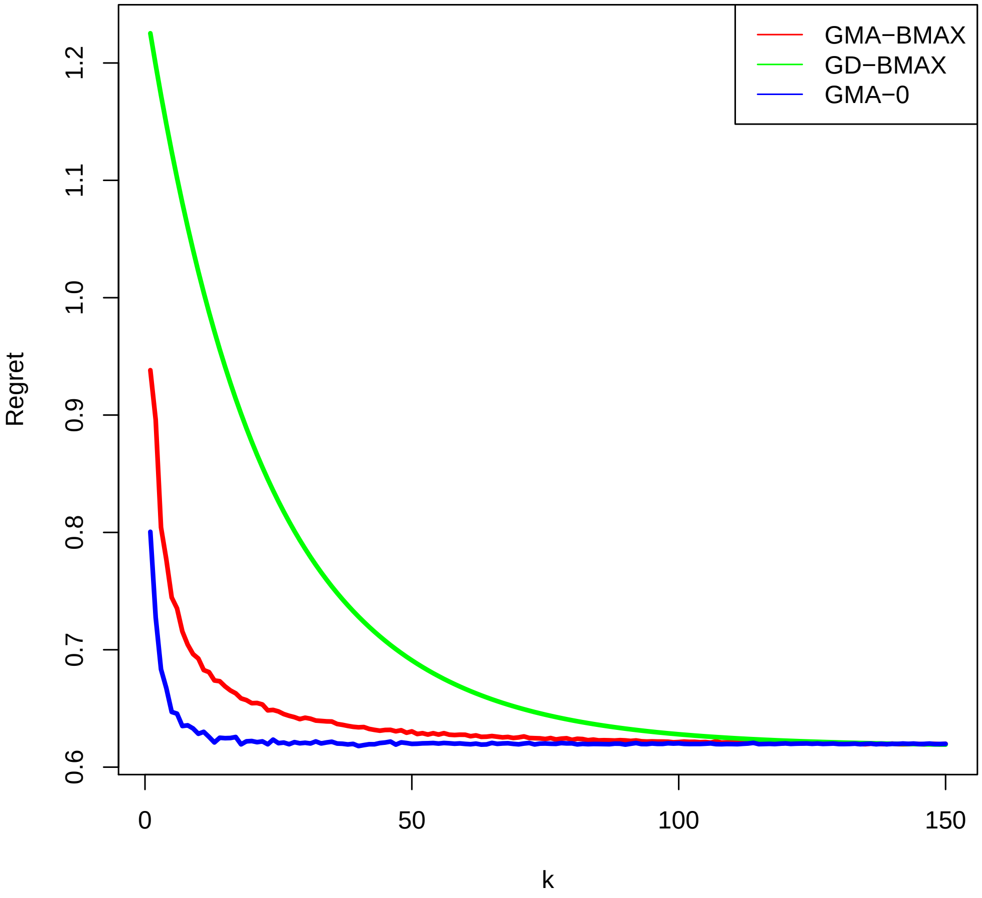

Figure 2.1: Regrets $R(\boldsymbol{\psi}^{(k)})$ versus iterations k.

behaves very similar. Notice that GMA-BMAX and GD-BMAX initialize both with $\boldsymbol{\psi}^{(0)} = 0$, but they has difference estimator even after the first iteration $(k = 1)$, GMA-BMAX selects $j \in \{1, \ldots, M\}$ that minimizes $\log J(\boldsymbol{f}_j)$ and GD-BMAX outputs a dense estimator, whiles GMA-0 selects $j \in \{1, \ldots, M\}$ that minimizes $Q(\boldsymbol{f}_j)$ and the first stage output is actually the empirical risk minimizer $\boldsymbol{f}_{k_1}$ where $k_1 = \text{argmin}_j \widehat{\text{MSE}}(f_j)$.

Chapter 3
Aggregation of Linear Models

3.1 Aggregate by BMAX for Linear Models with Gaussian priors

In the previous chapter, we introduced aggregate by Bayesian model averaging with exponentiated least square loss (BMAX), which is proved to solve the model averaging problem optimally both in deviation and in expectation. There we assume $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M$, the M candidate estimators of $\boldsymbol{\eta} = \mathbb{E}\boldsymbol{Y}$ are deterministic, independent of the noise $\boldsymbol{\xi} \sim N(0, \sigma^2 \boldsymbol{I}_n)$, where $\boldsymbol{Y} = \boldsymbol{\eta} + \boldsymbol{\xi}$. This scenario will apply for example, when data are split for training and testing, and different estimators for the testing data are learned from training data based on different models. In this chapter, we will investigate the scenario where the candidates are not independent of noise under linear model assumption.

3.1.1 Bayesian Framework of BMAX for Linear Models with Gaussian priors

Given response vector $\boldsymbol{Y} = (y_1, \ldots, y_n)^\top$ which is constructed by some unknown mean $\boldsymbol{\eta} \in \mathbb{R}^n$ corrupted by Gaussian noise $\boldsymbol{\xi} \sim N(0, \sigma^2 \boldsymbol{I}_n)$, namely,

$$\boldsymbol{Y} = \boldsymbol{\eta} + \boldsymbol{\xi} \; ,$$

and we assume that σ^2 is known. Also a set of predictor variables $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_d \in \mathbb{R}^n$ are given, and define $\boldsymbol{X} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_d) \in \mathbb{R}^{n \times d}$. Our goal is to estimate the mean vector $\boldsymbol{\eta}$, given vector $\boldsymbol{Y}$ and matrix $\boldsymbol{X}$, and usually, we are interested in the relationship between the truth $\boldsymbol{\eta}$ and predictor variables $\boldsymbol{X}$, thus the basic but fundamental linear model is often considered.

Under the linear model, we assume $\boldsymbol{\eta}$ is in the linear space spanned by $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_d \in \mathbb{R}^n$, the columns of matrix $\boldsymbol{X}$, yet this assumption, is not necessarily true. Thus we assume a Bayesian framework in which the mean of $\boldsymbol{Y}$ is denoted by $\boldsymbol{\mu} \in \mathbb{R}^n$ instead of $\boldsymbol{\eta}$ to avoid confusions. Moreover, the model choice problem involves selecting a subset of predictor variables and placing additional restrictions on the subspace that contains the mean. Specifically, let $\wp = \{0,1\}^d \subset \mathbb{R}^d$, we index the model space by $\gamma = (\gamma_1, \ldots, \gamma_d)^\top \in \wp$, a vector of indicators with $\gamma_j = 1$, meaning that $\boldsymbol{f}_j$ is included in the set of predictor variables, and with $\gamma_j = 0$, meaning that $\boldsymbol{f}_j$ is excluded. Denote the number of elements in $\wp$ as $|\wp| = 2^d$. In addition, we assume that, given sparsity pattern γ, let $\mathcal{M}_\gamma$ denote the linear model under which $\boldsymbol{\mu}$ is in the linear space spanned by the columns in $\boldsymbol{X}$ with respective to the sparsity pattern γ.

Mathematically, we assume that

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\zeta} \ ,$$

$$\boldsymbol{\mu} | \mathcal{M}_\gamma = \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma \ ,$$

where $\boldsymbol{\zeta} \sim N(0, \omega^2 \boldsymbol{I}_n)$ with some parameter $\omega > 0$, and $\boldsymbol{X}_\gamma \in \mathbb{R}^{n \times d_\gamma}$ represents the design matrix composed by columns in $\boldsymbol{X}$ respective to sparsity pattern γ and $\boldsymbol{\beta}_\gamma \in \mathbb{R}^{d_\gamma}$ is the regression coefficients vector.

Zellner (1986)'s *g prior* for $\boldsymbol{\beta}_\gamma$ is defined as

$$\boldsymbol{\beta}_\gamma | \mathcal{M}_\gamma \sim N(0, g\omega^2 (\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma)^{-1}) \ , \tag{3.1}$$

which has been widely adopted because of its computational efficiency in evaluating marginal likelihoods and model search and, perhaps most important, because of its simple, understandable interpretation as arising from the analysis of a conceptual sample generated using the same design matrix $\boldsymbol{X}$ as employed in the current sample. So in this chapter, we adapt to use the Gaussian prior for $\boldsymbol{\beta}_\gamma$,

$$\boldsymbol{\beta}_\gamma | \mathcal{M}_\gamma \sim N(\tilde{\boldsymbol{\beta}}_\gamma, g\omega^2 \boldsymbol{K}_\gamma^{-1}) \ ,$$

where $\boldsymbol{K}_\gamma \in \mathbb{R}^{d_\gamma \times d_\gamma}$ is positive definite, $\tilde{\boldsymbol{\beta}}_\gamma \in \mathbb{R}^{d_\gamma}$ and $g > 0$ are given.

In a nutshell, our Bayesian framework is

$$\boldsymbol{Y} | \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \omega^2 \boldsymbol{I}_n) \ , \tag{3.2}$$

$$\boldsymbol{\mu} | \mathcal{M}_\gamma, \boldsymbol{\beta}_\gamma = \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma \ , \tag{3.3}$$

$$\boldsymbol{\beta}_\gamma | \mathcal{M}_\gamma \sim N(\tilde{\boldsymbol{\beta}}_\gamma, g\omega^2 \boldsymbol{K}_\gamma^{-1}) \ , \tag{3.4}$$

and model prior is

$$p(\mathcal{M}_\gamma) = \pi_\gamma \ , \tag{3.5}$$

for $\gamma \in \wp$, where $\boldsymbol{K}_\gamma \in \mathbb{R}^{d_\gamma \times d_\gamma}$ is positive definite, $\tilde{\boldsymbol{\beta}}_\gamma \in \mathbb{R}^{d_\gamma}$, and $g > 0$, $\omega > 0$ are given.

Consider Bayes estimator $\hat{\boldsymbol{\psi}}$

$$\hat{\boldsymbol{\psi}} = \operatorname*{argmin}_{\boldsymbol{\psi} \in \mathbb{R}^n} \mathbb{E}\left[L(\boldsymbol{\psi}, \boldsymbol{\mu}) | \boldsymbol{Y}\right] \tag{3.6}$$

where L is some loss function.

It follows from the above Bayesian framework that

$$\begin{aligned}
& \mathbb{E}\left[L(\boldsymbol{\psi}, \boldsymbol{\mu}) | \boldsymbol{Y}\right] \\
= \ & \sum_{\gamma \in \wp} \int L(\boldsymbol{\psi}, \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma) p(\boldsymbol{\beta}_\gamma | \boldsymbol{Y}, \mathcal{M}_\gamma) p(\mathcal{M}_\gamma | \boldsymbol{Y}) \ d\boldsymbol{\beta}_\gamma \\
= \ & \frac{1}{p(\boldsymbol{Y})} \sum_{\gamma \in \wp} \int L(\boldsymbol{\psi}, \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma) p(\boldsymbol{Y} | \mathcal{M}_\gamma, \boldsymbol{\beta}_\gamma) p(\boldsymbol{\beta}_\gamma | \mathcal{M}_\gamma) p(\mathcal{M}_\gamma) \ d\boldsymbol{\beta}_\gamma \\
= \ & \frac{(2\pi\omega^2)^{-n/2}}{p(\boldsymbol{Y})} \sum_{\gamma \in \wp} \pi_\gamma \int L(\boldsymbol{\psi}, \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma) \exp\left(-\frac{1}{2\omega^2} \|\boldsymbol{Y} - \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma\|_2^2 \right) p(\boldsymbol{\beta}_\gamma | \mathcal{M}_\gamma) \ d\boldsymbol{\beta}_\gamma
\end{aligned}$$

For least square loss $L(\boldsymbol{\psi}, \boldsymbol{\mu}) = \|\boldsymbol{\psi} - \boldsymbol{\mu}\|_2^2$, the Bayes estimator is the posterior mean

$$\boldsymbol{\psi}_{\ell_2}(\omega^2) = \mathbb{E}(\boldsymbol{\mu}|\boldsymbol{Y}) = \sum_{\gamma \in \wp} p(\mathcal{M}_\gamma|\boldsymbol{Y})\boldsymbol{X}_\gamma \mathbb{E}(\boldsymbol{\beta}_\gamma|\boldsymbol{Y}, \mathcal{M}_\gamma) = \sum_{\gamma \in \wp} \lambda_\gamma \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma \ , \tag{3.7}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{|\wp|})^\top \in \Lambda^\wp$,

$$\begin{aligned}\lambda_\gamma \propto \pi_\gamma |\frac{1}{g}\boldsymbol{K}_\gamma|^{1/2} |\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \frac{1}{g}\boldsymbol{K}_\gamma|^{-1/2} \\ \cdot \exp\left(-\frac{1}{2\omega^2}\|\boldsymbol{Y} - \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma\|_2^2 - \frac{(\hat{\boldsymbol{\beta}}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)^\top \boldsymbol{K}_\gamma (\hat{\boldsymbol{\beta}}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)}{2g\omega^2}\right) \ ,\end{aligned} \tag{3.8}$$

and

$$\hat{\boldsymbol{\beta}}_\gamma = (\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \frac{1}{g}\boldsymbol{K}_\gamma)^{-1}(\boldsymbol{X}_\gamma^\top \boldsymbol{Y} + \frac{1}{g}\boldsymbol{K}_\gamma \tilde{\boldsymbol{\beta}}_\gamma) \ . \tag{3.9}$$

Proof of (3.7) can be found in Appendix.

Now we apply the exponentiated least square loss $L(\boldsymbol{\psi}, \boldsymbol{\mu}) = \exp\{\frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi} - \boldsymbol{\mu}\|_2^2\}$ that introduced in Chapter 2, the respective Bayes estimator is

$$\boldsymbol{\psi}_X(\omega^2, \nu) = \underset{\boldsymbol{\psi} \in \mathbb{R}^n}{\operatorname{argmin}} \, J(\boldsymbol{\psi}) \ , \tag{3.10}$$

where

$$\begin{aligned}J(\boldsymbol{\psi}) = \sum_{\gamma \in \wp} \pi_\gamma \int \exp\left(\frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi} - \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma\|_2^2 - \frac{1}{2\omega^2}\|\boldsymbol{Y} - \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma\|_2^2\right) \\ (2\pi g\omega^2)^{-d_\gamma/2} |\boldsymbol{K}_\gamma|^{1/2} \exp\left(-\frac{(\boldsymbol{\beta}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)^\top \boldsymbol{K}_\gamma (\boldsymbol{\beta}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)}{2g\omega^2}\right) \, d\boldsymbol{\beta}_\gamma \ .\end{aligned} \tag{3.11}$$

By integrating out $\boldsymbol{\beta}_\gamma$ in (3.11), we can rewrite $J(\boldsymbol{\psi})$ as following:

Proposition 9.

$$\begin{aligned}J(\boldsymbol{\psi}) = \sum_{\gamma \in \wp} \pi_\gamma |\frac{1}{g}\boldsymbol{K}_\gamma|^{1/2} |\nu \boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \frac{1}{g}\boldsymbol{K}_\gamma|^{-1/2} \\ \cdot \exp\left(\frac{1-\nu}{2\omega^2}(\boldsymbol{\psi} - \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma)^\top \boldsymbol{W}_\gamma (\boldsymbol{\psi} - \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma)\right) \\ \cdot \exp\left(-\frac{1}{2\omega^2}\|\boldsymbol{Y} - \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma\|_2^2 - \frac{(\hat{\boldsymbol{\beta}}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)^\top \boldsymbol{K}_\gamma (\hat{\boldsymbol{\beta}}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)}{2g\omega^2}\right) \ ,\end{aligned} \tag{3.12}$$

where

$$\boldsymbol{W}_\gamma = \boldsymbol{I}_n + (1-\nu)\boldsymbol{X}_\gamma \left(\nu \boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \frac{1}{g}\boldsymbol{K}_\gamma\right)^{-1} \boldsymbol{X}_\gamma^\top \, . \tag{3.13}$$

From the above proposition, by setting $\nabla \log J(\boldsymbol{\psi}) = 0$, it is easy to see that $\boldsymbol{\psi}_X(\omega^2, \nu)$ is essentially the aggregation of $\boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma$'s, the Maximum A Posteriori (MAP) estimator of $\boldsymbol{\mu}$ under each model $\mathcal{M}_\gamma$, though the aggregation is not simply in the form of $\sum_{\gamma\in\wp} \lambda_\gamma \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma$ where $\sum_{\gamma\in\wp} \lambda_\gamma = 1$, like ordinary Bayesian model averaging framework (e.g. (3.7)) in which λ_γ is proportional to the posterior probability $p(\boldsymbol{Y}|\mathcal{M}_\gamma)$ (or modified with some dimension penalty); but in the form of $\sum_{\gamma\in\wp} A_\gamma \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma$ where $A_\gamma \in \mathbb{R}^{n\times n}$ and $\sum_{\gamma\in\wp} A_\gamma = \boldsymbol{I}_n$, thus in this case it puts different weights for each observation in the averaging procedure, which may gives us some clues about model averaging for linear models: estimate each observation by model averaging with different weights on the predictor variables.

3.1.2 Deviation Bounds of BMAX for Linear Models with Gaussian priors

Next we propose a theorem stating BMAX estimator for linear models with Gaussian priors is competitive with any single linear model.

Theorem 5. *Consider BMAX estimator* $\boldsymbol{\psi}_X(\omega^2, \nu)$ *as defined in* (3.10) *and* (3.11)*, assume* $\nu \in (0,1)$ *and if* $\omega^2 \geq \frac{\sigma^2}{\min(\nu, 1-\nu)}$*, we have oracle inequality*

$$\|\boldsymbol{\psi}_X(\omega^2,\nu) - \boldsymbol{\eta}\|_2^2 \leq \min_{\substack{\gamma\in\wp \\ \boldsymbol{\beta}_\gamma^* \in \mathbb{R}^{d_\gamma}}} \left\{ \|\boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma^* - \boldsymbol{\eta}\|_2^2 + \frac{1}{g}(\boldsymbol{\beta}_\gamma^* - \tilde{\boldsymbol{\beta}}_\gamma)^\top \boldsymbol{K}_\gamma (\boldsymbol{\beta}_\gamma^* - \tilde{\boldsymbol{\beta}}_\gamma) \right.$$
$$\left. + 2\omega^2 \log(\frac{1}{\pi_\gamma \delta}) + \omega^2 \log(|\nu g \boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma \boldsymbol{K}_\gamma^{-1} + \boldsymbol{I}_{d_\gamma}|) \right\} \, , \tag{3.14}$$

with probability at least $1-\delta$. *Moreover,*

$$\mathbb{E}\|\boldsymbol{\psi}_X(\omega^2,\nu)-\boldsymbol{\eta}\|_2^2 \leq \min_{\substack{\gamma\in\wp\\ \boldsymbol{\beta}^*_\gamma\in\mathbb{R}^{d_\gamma}}} \left\{\|\boldsymbol{X}_\gamma\boldsymbol{\beta}^*_\gamma-\boldsymbol{\eta}\|_2^2+\frac{1}{g}(\boldsymbol{\beta}^*_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)^\top\boldsymbol{K}_\gamma(\boldsymbol{\beta}^*_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)\right.$$
$$\left.+2\omega^2\log(\frac{1}{\pi_\gamma})+\omega^2\log(|\nu g\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma\boldsymbol{K}_\gamma^{-1}+\boldsymbol{I}_{d_\gamma}|)\right\}\ . \tag{3.15}$$

Remark 10. *The term* $\|\boldsymbol{X}_\gamma\boldsymbol{\beta}^*_\gamma-\boldsymbol{\eta}\|_2^2+\frac{1}{g}(\boldsymbol{\beta}^*_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)^\top\boldsymbol{K}_\gamma(\boldsymbol{\beta}^*_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)$ *on the right hand side of the above two inequalities, can be further minimized for* $\boldsymbol{\beta}^*_\gamma$ *over* $\mathbb{R}^{d_\gamma}$*, and the minimizer is essentially the ridge regression estimator under model* $\mathcal{M}_\gamma$ *as if there is no noise.*

Remark 11. *If we use Zellner's g prior* (3.1) *for* $\boldsymbol{\beta}_\gamma|\mathcal{M}_\gamma$ *(i.e. set* $\boldsymbol{K}_\gamma=\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma$*), then* $\log(|\nu g\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma\boldsymbol{K}_\gamma^{-1}+\boldsymbol{I}_{d_\gamma}|)=d_\gamma\log(1+\nu g)$ *is of order d.*

Theorem 5 shows oracle inequalities of BMAX estimator to compete with several linear models with Gaussian priors. It is easy to see that, the two inequalities (3.14) and (3.15) still hold when setting $\pi_k=1$ for some $k\in\wp$ and $\pi_j=0$ for $j\neq k$, which means there is only one linear model with Gaussian prior to be considered. Then the following corollary is directly obtained from Theorem 5 and Proposition 9, and we will also see that $\boldsymbol{\psi}_X(\omega^2,\nu)$ turns out to be a ridge regression estimator.

Corollary 4. *For fixed sparsity pattern* $k\in\wp$*, consider BMAX estimator* $\boldsymbol{\psi}_X(\omega^2,\nu)$ *as defined in* (3.10), (3.11) *with* $\pi_k=1$*, assume* $\nu\in(0,1)$ *and if* $\omega^2\geq\frac{\sigma^2}{\min(\nu,1-\nu)}$*, for any* $\boldsymbol{\beta}^*_k\in\mathbb{R}^{d_k}$ *we have oracle inequality*

$$\|\boldsymbol{\psi}_X(\omega^2,\nu)-\boldsymbol{\eta}\|_2^2\leq\|\boldsymbol{X}_k\boldsymbol{\beta}^*_k-\boldsymbol{\eta}\|_2^2+\frac{1}{g}(\boldsymbol{\beta}^*_k-\tilde{\boldsymbol{\beta}}_k)^\top\boldsymbol{K}_k(\boldsymbol{\beta}^*_k-\tilde{\boldsymbol{\beta}}_k)$$
$$+2\omega^2\log(\frac{1}{\delta})+\omega^2\log(|\nu g\boldsymbol{X}_k^\top\boldsymbol{X}_k\boldsymbol{K}_k^{-1}+\boldsymbol{I}_{d_k}|)\ , \tag{3.16}$$

with probability at least $1-\delta$. *Moreover,*

$$\mathbb{E}\|\boldsymbol{\psi}_X(\omega^2,\nu)-\boldsymbol{\eta}\|_2^2 \le \|\boldsymbol{X}_k\boldsymbol{\beta}_k^*-\boldsymbol{\eta}\|_2^2+\frac{1}{g}(\boldsymbol{\beta}_k^*-\tilde{\boldsymbol{\beta}}_k)^\top\boldsymbol{K}_k(\boldsymbol{\beta}_k^*-\tilde{\boldsymbol{\beta}}_k) + \omega^2\log(|\nu g\boldsymbol{X}_k^\top\boldsymbol{X}_k\boldsymbol{K}_k^{-1}+\boldsymbol{I}_{d_k}|)\ . \tag{3.17}$$

Also, $\boldsymbol{\psi}_X(\omega^2,\nu)$ *can be expressed explicitly as*

$$\boldsymbol{\psi}_X(\omega^2,\nu)=\boldsymbol{X}_k\hat{\boldsymbol{\beta}}_k=\boldsymbol{X}_k(\boldsymbol{X}_k^\top\boldsymbol{X}_k+\frac{1}{g}\boldsymbol{K}_k)^{-1}(\boldsymbol{X}_k^\top\boldsymbol{Y}+\frac{1}{g}\boldsymbol{K}_k\tilde{\boldsymbol{\beta}}_k) \tag{3.18}$$

From the above corollary we can see that, the BMAX estimator for fixed sparsity $k \in \wp$ can compete with any linear predictors of sparsity pattern k. Also the BMAX estimator turns out to be a ridge regression estimator when sparsity pattern is fixed.

3.2 Gradient Descent Algorithm for Solving BMAX in Linear Models with Gaussian Priors

To solve $\boldsymbol{\psi}_X(\omega^2,\nu)$ is equivalent to solve the minimization problem of $\log J(\boldsymbol{\psi})$. In this section we propose a gradient descent algorithm to solve $\boldsymbol{\psi}_X(\omega^2,\nu)$ as defined in (3.10) and (3.11) based on expression (3.12) in Proposition 9.

Define the following condition under which the ℓ_2-norm of $\hat{\boldsymbol{f}}_\gamma=\boldsymbol{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma$ is bounded by some constant $L_2\in\mathbb{R}$:

$$\|\hat{\boldsymbol{f}}_\gamma\|_2 \le L_2 \quad \forall\,\gamma\in\wp\ . \tag{3.19}$$

Given L_2, define

$$A_3=\frac{2(1-\nu)^2(1+\nu)}{\nu^3}(L_2^2/\omega^4)+\frac{1-\nu}{\nu}(1/\omega^2)\ , \tag{3.20}$$

With L_2 and A_3 defined, the algorithm is as following.

The following lemma describes the strong convexity of $\log J(\boldsymbol{\psi})$.

Algorithm 5 Gradient Descent Algorithm to solve BMAX in Linear Models with Gaussian Priors (GD-BMAX-LM)

```
Input: Noisy observation $\boldsymbol{Y}$, $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{K}_\gamma \in \mathbb{R}^{d_\gamma \times d_\gamma}$; prior $\boldsymbol{\pi} \in \Lambda^{|\wp|}$;
    parameters $g, \nu, \omega^2$.
Output: Aggregate estimator $\boldsymbol{\psi}^{(k)}$.
  Let $\boldsymbol{\psi}^{(0)} = 0$, $\hat{\boldsymbol{f}}_\gamma = \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma$ where $\hat{\boldsymbol{\beta}}_\gamma$ is defined as (3.9) for $\gamma \in \wp$.
  for $k = 1, 2, \ldots$ do
    Choose step size $t_k = s \in (0, 2/A_3)$.
    Calculate
        $$\nabla \log J(\boldsymbol{\psi}^{(k-1)}) = \frac{1-\nu}{\omega^2} \sum_{\gamma \in \wp} \lambda_\gamma^{(k-1)} \boldsymbol{W}_\gamma (\boldsymbol{\psi}^{(k-1)} - \hat{\boldsymbol{f}}_\gamma)$$
    where $\boldsymbol{\lambda}^{(k-1)} \in \Lambda^{|\wp|}$ and
        $$\lambda_\gamma^{(k-1)} \propto \pi_\gamma |\frac{1}{g}\boldsymbol{K}_\gamma|^{1/2} |\nu \boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \frac{1}{g}\boldsymbol{K}_\gamma|^{-1/2}$$
        $$\cdot \exp\left(\frac{1-\nu}{2\omega^2} (\boldsymbol{\psi}^{(k-1)} - \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma)^\top \boldsymbol{W}_\gamma (\boldsymbol{\psi}^{(k-1)} - \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma)\right)$$
        $$\cdot \exp\left(-\frac{1}{2\omega^2} \|\boldsymbol{Y} - \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma\|_2^2 - \frac{(\hat{\boldsymbol{\beta}}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)^\top \boldsymbol{K}_\gamma (\hat{\boldsymbol{\beta}}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)}{2g\omega^2}\right)$$
    Calculate
        $$\boldsymbol{\psi}^{(k)} = \boldsymbol{\psi}^{(k-1)} - t_k \nabla \log J(\boldsymbol{\psi}^{(k-1)}).$$
  end for
```

Lemma 4. *For any $\boldsymbol{\psi} \in \mathbb{R}^n$ we have*

$$\nabla^2 \log J(\boldsymbol{\psi}) \geq A_1 \boldsymbol{I}_n \ , \tag{3.21}$$

and if condition (3.19) *satisfies, then for $k > 0$ and $\alpha \in [0, 1]$ it holds that*

$$\nabla^2 \log J((1-\alpha)\boldsymbol{\psi}^{(k-1)} + \alpha \boldsymbol{\psi}^{(k)}) \leq A_3 \boldsymbol{I}_n \ , \tag{3.22}$$

where A_1, A_3 are defined as (2.25) *and* (3.20), *$\boldsymbol{\psi}^{(k)}$ is the k-th iteration output from Algorithm 5.*

The strong convexity of $\log J(\boldsymbol{\psi})$ described by Lemma 4 implies the following lemma, which measures the quantity of $[\log J(\boldsymbol{\psi}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))]$.

Lemma 5. *For any $\boldsymbol{\psi} \in \mathbb{R}^n$ we have the following inequalities*

$$\log J(\boldsymbol{\psi}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu)) \leq \frac{1}{2A_1} \|\nabla \log J(\boldsymbol{\psi})\|_2^2 \ , \tag{3.23}$$

where A_1 is defined as (2.25).

Proposition 10. *Given condition* (3.19) *is satisfied, for $\boldsymbol{\psi}^{(k)}$ output from Algorithm 5 and choose fixed step size $t_k = s \in (0, 2/A_3)$ for $k > 0$, then*

$$\begin{aligned} & \log J(\boldsymbol{\psi}^{(k)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu)) \\ \leq \ & \left[1 - 2A_1(s - (A_3/2)s^2)\right]^k \left(\log J(\boldsymbol{\psi}^{(0)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))\right) \ . \end{aligned}$$

Remark 12. *We can simply take $s = 1/A_3$ to minimize the right hand side of above inequality, it results that*

$$\log J(\boldsymbol{\psi}^{(k)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu)) \leq (1 - A_1/A_3)^k \left(\log J(\boldsymbol{\psi}^{(0)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))\right) \ .$$

The proof of Proposition 10 is almost the same as that of Proposition 5. And it states that Algorithm 5 converges to the minimum of $\log J(\boldsymbol{\psi})$ with a geographic rate.

When $|\wp| = 2^d$ is large, it is not practical to directly calculate $\nabla \log J(\boldsymbol{\psi}^{(k-1)})$ out. Instead, we may use Monte Carlo methods such as Metropolis-Hastings algorithm (Algorithm 6) to approximate $\nabla \log J(\boldsymbol{\psi}^{(k-1)})$ of k-th step in Algorithm 5.

Algorithm 6 Metropolis-Hastings Sampler for estimating $\nabla \log J(\boldsymbol{\psi}^{(k-1)})$ at k-th step in Algorithm 5

```
Input: Noisy observation $\boldsymbol{Y}$; for $\gamma \in \wp$, $\boldsymbol{X}_\gamma \in \mathbb{R}^{n \times d_\gamma}$ and $\boldsymbol{K}_\gamma \in \mathbb{R}^{d_\gamma \times d_\gamma}$; prior
    $\boldsymbol{\pi} \in \Lambda^{|\wp|}$; parameters $g, \nu, \omega^2$, $(k-1)$-th step estimator $\boldsymbol{\psi}^{(k-1)}$.
Output: $\boldsymbol{v}_T^{(k-1)}$ as estimator of $\nabla \log J(\boldsymbol{\psi}^{(k-1)})$.
  Initialize $j(0) = 0$.
  for $t = 1, \cdots, T_0 + T$ do
    Generate $\tilde{\gamma} \sim q(\cdot|\gamma(t-1))$.
    Compute
        $\rho(\gamma(t-1), \tilde{\gamma}) = \min\left(\frac{q(\gamma(t-1)|\tilde{\gamma})\theta(\tilde{\gamma})}{q(\tilde{\gamma}|\gamma(t-1))\theta(\gamma(t-1))}, 1\right)$ ,
    where
        $\theta(\gamma) = \pi_\gamma|\frac{1}{g}\boldsymbol{K}_\gamma|^{1/2}|\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \frac{1}{g}\boldsymbol{K}_\gamma|^{-1/2}$
            $\cdot \exp\left(\frac{1-\nu}{2\omega^2}(\boldsymbol{\psi}^{(k-1)} - \boldsymbol{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma)^\top\boldsymbol{W}_\gamma(\boldsymbol{\psi}^{(k-1)} - \boldsymbol{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma)\right)$
            $\cdot \exp\left(-\frac{1}{2\omega^2}\|\boldsymbol{Y} - \boldsymbol{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma\|_2^2 - \frac{(\hat{\boldsymbol{\beta}}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)^\top\boldsymbol{K}_\gamma(\hat{\boldsymbol{\beta}}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)}{2g\omega^2}\right)$
    Generate a random variable
        $\gamma(t) = \begin{cases} \tilde{\gamma}\ , & \text{with probability} \quad \rho(\gamma(t-1), \tilde{\gamma})\ ; \\ \gamma(t-1)\ , & \text{with probability} \quad 1 - \rho(\gamma(t-1), \tilde{\gamma})\ . \end{cases}$
  end for
  Calculate
        $\boldsymbol{v}_T^{(k-1)} = \frac{1-\nu}{\omega^2}\frac{1}{T}\sum_{t=T_0+1}^{T_0+T} W_{\gamma(t)}(\boldsymbol{\psi}^{(k-1)} - \boldsymbol{X}_{\gamma(t)}\hat{\boldsymbol{\beta}}_{\gamma(t)})$
```

3.3 A Frequentist's Approach: Apply Q-Aggregation To Affine Estimators

In Section 2.5, we introduced Q-aggregation (Dai et al., 2012), and proved that Q-aggregation (with KL entropy) is essentially a dual representation of the aggregate by BMAX, under the assumption that candidate estimators $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M\}$ are static, being independent of noise $\boldsymbol{\xi}$.

In last section BMAX is applied to linear models with Gaussian priors, and the Bayes estimator is an aggregation of ridge regression estimators which are not independent of noise $\boldsymbol{\xi}$.

Now we define affine estimators $\hat{\boldsymbol{f}}_\gamma$ indexed by parameter $\gamma \in \wp$,

$$\hat{\boldsymbol{f}}_\gamma = A_\gamma \boldsymbol{Y} + \boldsymbol{b}_\gamma \ , \tag{3.24}$$

where symmetric $A_\gamma \in \mathbb{R}^{d_\gamma \times d_\gamma}$ satisfies

$$\begin{cases} A_\gamma \geq 0 \ , \\ \max_{\gamma \in \wp} \|A_\gamma\|_{\text{op}} \leq V \ , \end{cases} \tag{3.25}$$

for some constant V and $\|\cdot\|_{\text{op}}$ is the matrix spectral norm. Denote the set size of $\wp$ as $|\wp|$.

In this section, given observation $\boldsymbol{Y} \in \mathbb{R}^n$, we propose a model averaging approach by aggregation of affine estimators $\hat{\boldsymbol{f}}_\gamma$.

Affine estimators are frequently used in the statistical literature and the following are several examples (see, e.g., Dalalyan and Salmon, 2012).

- Ordinary least squares: we assume $\boldsymbol{b}_\gamma = 0$ and $A_\gamma \boldsymbol{Y}$ is the projection of $\boldsymbol{Y}$ to a linear subspace $\mathcal{L}_\gamma$ of $\mathbb{R}^n$. In our special example with γ being the sparsity pattern of a dictionary, we denote by $\boldsymbol{X}_\gamma$ the design matrix $\boldsymbol{X}$ restricted to the columns indicated by the sparsity pattern γ. Let $A_\gamma = \boldsymbol{X}_\gamma(\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma)^- \boldsymbol{X}_\gamma^\top$, where B^- represents the pseudo-inverse of a matrix B, then we have

$$\hat{\boldsymbol{f}}_\gamma = A_\gamma \boldsymbol{Y} = \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma ,$$

where

$$\hat{\boldsymbol{\beta}}_\gamma \in \operatorname*{argmin}_{\boldsymbol{\beta}_\gamma \in \mathbb{R}^{d_\gamma}} \|\boldsymbol{Y} - \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma\|_2^2 .$$

- Ridge regression: with γ being the sparsity pattern of a given design matrix $\boldsymbol{X}$, set $\boldsymbol{b}_\gamma = 0$ and $A_\gamma = \boldsymbol{X}_\gamma(\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \lambda \boldsymbol{I}_{d_\gamma})^{-1}\boldsymbol{X}_\gamma^\top$ for some $\lambda > 0$, which gives the estimator

$$\hat{\boldsymbol{f}}_\gamma = A_\gamma \boldsymbol{Y} = \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma ,$$

where $\hat{\boldsymbol{\beta}}_\gamma$ is the solution of the ridge regression problem

$$\hat{\boldsymbol{\beta}}_\gamma \in \operatorname*{argmin}_{\boldsymbol{\beta}_\gamma \in \mathbb{R}^{d_\gamma}} \left[\|\boldsymbol{Y} - \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma\|_2^2 + \lambda \|\boldsymbol{\beta}_\gamma\|_2^2\right] .$$

- Diagonal filters: the matrices A_γ's are diagonal; that is, $A_\gamma = \operatorname{diag}(a_1, \cdots, a_n)$. An example given in Dalalyan and Salmon (2012), called truncated SVD, corresponds to the choice of $a_k = \mathbf{1}_{k \le \gamma}$ for some integer $\gamma \in \wp = \{1, \ldots, n\}$.

The aggregation is taken in the form of

$$\hat{\mathfrak{f}}_{\boldsymbol{\lambda}} = \sum_{j=1}^{|\wp|} \lambda_j \hat{\boldsymbol{f}}_j . \tag{3.26}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{|\wp|})^\top \in \Lambda^{|\wp|}$ and $\hat{\boldsymbol{f}}_j$ defined as (3.24). Our goal is to find a proper $\boldsymbol{\lambda}$ such that $\hat{\mathfrak{f}}_{\boldsymbol{\lambda}}$ is a good approximation of $\boldsymbol{\eta} = \mathbb{E}\boldsymbol{Y}$.

Leung and Barron (2006) proposed an aggregation method over estimators that are least-square projections onto linear subspaces. For instance, under a given $n \times d_\gamma$ design matrix $\boldsymbol{X}_\gamma$ ($d_\gamma \le n$), the project estimator to the linear subspace spanned by the columns of $\boldsymbol{X}_\gamma$ is given by the least squares estimator $\hat{\boldsymbol{f}}_\gamma = \boldsymbol{X}_\gamma(\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma)^{-1}\boldsymbol{X}_\gamma^\top \boldsymbol{Y}$. In Leung and Barron (2006), the coefficients used in aggregation formula (3.26) is based on exponential weighting, given by

$$\lambda_\gamma \propto \pi_\gamma \exp\left(-(\|\hat{\boldsymbol{f}}_\gamma - \boldsymbol{Y}\|_2^2 + \sigma^2(2d_\gamma - n))/(2\omega^2)\right) ,$$

where $\{\pi_\gamma\}$ is a prior distribution over the dictionary of models. For appropriately chosen $\omega^2 \geq 2\sigma^2$, Leung and Barron (2006) proved optimal oracle inequalities in expectation for the exponential weighted aggregation method, with $\Delta(n, M, \sigma^2)$ in (1.17) being of the order $\sigma^2 \log M/n$. Extensions of Leung and Barron (2006) have been made in multiple directions. Rigollet and Tsybakov (2011) focused on high-dimension models and treated the sparsity in particular. They proposed the Exponential Screening (ES) estimator by choosing a specific discrete prior. The ES estimator benefits from three types of sparsity simultaneously, which includes the low rank of design matrix $\boldsymbol{X}$, ℓ_0 and ℓ_1 norm of the parameter vector. Dalalyan and Salmon (2012) extended projection estimators to general affine estimators which take the more general form of $\hat{\boldsymbol{f}}_\gamma = A_\gamma \boldsymbol{Y} + \boldsymbol{b}_\gamma$. However, previous work using the exponential weighting scheme only led to oracle inequalities in expectation. There are so far no deviation results that hold in high probability for affine model aggregation.

Now we propose an aggregation estimator that can achieve a proper deviation bound, by applying a modified version of Q-aggregations (Rigollet, 2012; Dai et al., 2012) to affine estimators (3.24) that satisfy conditions (3.25). Note that different from Dalalyan and Salmon (2012), we do not require the affine matrices A_γ's to be exchangeable and we make no assumption on $\boldsymbol{b}_\gamma$.

Specifically, define

$$\hat{\boldsymbol{f}}^{\mathrm{Q}} = \hat{\mathfrak{f}}_{\hat{\boldsymbol{\theta}}} = \sum_{\gamma \in \wp} \hat{\theta}_\gamma \hat{\boldsymbol{f}}_\gamma, \tag{3.27}$$

where

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \Lambda^{|\wp|}} \left\{ (1-\nu) \|\hat{\mathfrak{f}}_{\boldsymbol{\theta}} - \boldsymbol{Y}\|_2^2 + \nu \sum_{\gamma \in \wp} \theta_\gamma \|\hat{\boldsymbol{f}}_\gamma - \boldsymbol{Y}\|_2^2 + \Phi \sum_{\gamma \in \wp} \theta_\gamma C_\gamma + \Phi \mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\pi}) \right\} , \tag{3.28}$$

$\nu \in (0, 1)$ and C_γ is defined as

$$C_\gamma = \frac{8\sigma^4 \mathrm{tr}(A_\gamma^2)}{\Phi^2 - 8V\Phi\sigma^2} + \frac{2\sigma^2 \mathrm{tr}(A_\gamma)}{\Phi} \, . \tag{3.29}$$

The following theorem shows the modified Q-aggregation of affine estimators has an oracle inequality with a proper deviation bound.

Theorem 6. *Consider affine estimators* $\hat{\boldsymbol{f}}_\gamma$ *of* (3.24) *that satisfy conditions* (3.25)*. Given* $\nu \in (0,1)$*, let* $\hat{\boldsymbol{f}}^Q$ *be the aggregation estimator defined by* (3.27)*,* (3.28) *and* (3.29)*. If* $\Phi \geq 32\left[V \vee (\min\{\nu, 1-\nu\})^{-1}\right]\sigma^2$*, then for any fixed* $q \in \wp$*, we have*

$$\|\hat{\boldsymbol{f}}^Q - \boldsymbol{\eta}\|_2^2 \leq \|\hat{\boldsymbol{f}}_q - \boldsymbol{\eta}\|_2^2 + \Phi C_q + \Phi \log(\frac{1}{\pi_q \delta}) \ ,$$

with probability at least $1 - \delta$*.*

Theorem 6 states that Q-aggregation on affine estimators is competitive to any single affine estimators in the given dictionary, while Theorem 5 states that the aggregate by BMAX for linear models with Gaussian priors is competitive to any single linear estimator given the design matrix $\boldsymbol{X}$. The remainder term in oracle inequalities of both theorems are of order $O(d_\gamma)$, and such dimension related term seems extra compared to the results of Leung and Barron (2006) and Dalalyan and Salmon (2012), which actually can not be eliminated due the chi-square type noise term from projection of Gaussian noise.

Chapter 4

Fully Bayes Approach with Hyper Prior on g

Rather than using a plug-in estimate to eliminate g, a natural alternative is Fully Bayes (FB) with the integrated marginal likelihood under a proper hyper prior on g. Consequently, a prior on g leads to a mixture of priors on the coefficients $\boldsymbol{\beta}_\gamma$, and it typically provides more robust inference. Although Zellner and Siow (1980) did not explicitly use a g-prior formulation with a prior on g, their recommendation of a multivariate Cauchy form for $p(\boldsymbol{\beta}_\gamma|\sigma^2)$ implicitly corresponds to using a g-prior with an Inv-Gamma(1/2,n/2) prior on g, namely,

$$p(\boldsymbol{\beta}_\gamma|\sigma^2) \propto \int N\left(\boldsymbol{\beta}_\gamma \,|\, 0, g\sigma^2(\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma)^{-1}\right) \pi(g)\, dg \; ,$$

with

$$\pi(g) = \frac{(n/2)^{1/2}}{\Gamma(1/2)} g^{-3/2} e^{-n/(2g)} \; .$$

Besides the above Zellner-Siow prior on g, both Liang et al. (2008) and Cui and George (2008) investigated hyper-g prior on g in the form of

$$\pi(g) = \frac{a-2}{2}(1+g)^{-a/2} \; , \;\; g > 0 \; ,$$

which is a proper distribution for $a > 2$. This family of priors includes priors used by Strawderman (1971) to provide improved mean square risk over ordinary maximum likelihood estimates in the normal means problem. Liang et al. (2008) also modified the hyper-g prior to hyper-g/n prior

$$\pi(g) = \frac{a-2}{2n}(1+\frac{g}{n})^{-a/2} \; ,$$

for model selection consistency under the null model.

4.1 BMAX framework and settings for Linear Models with Gaussian priors and priors on g

We adopt the Bayesian framework for linear models in Chapter 3, the only difference is that in this chapter we put a prior on the parameter g. Given $\boldsymbol{Y} \in \mathbb{R}^n$ and $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, our Bayesian framework for this chapter is

$$\boldsymbol{Y}|\boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \omega^2 \boldsymbol{I}_n) \ , \tag{4.1}$$

$$\boldsymbol{\mu}|\mathcal{M}_\gamma, \boldsymbol{\beta}_\gamma = \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma \ , \tag{4.2}$$

$$\boldsymbol{\beta}_\gamma | g_\gamma, \mathcal{M}_\gamma \sim N(\tilde{\boldsymbol{\beta}}_\gamma, g_\gamma \omega^2 \boldsymbol{K}_\gamma^{-1}) \ , \tag{4.3}$$

$$g_\gamma|\mathcal{M}_\gamma \sim \text{Inv-Gamma}(\alpha_\gamma, (d_\gamma/2 + \alpha_\gamma) g_0) \ , \tag{4.4}$$

and model prior is

$$p(\mathcal{M}_\gamma) = \pi_\gamma \ , \tag{4.5}$$

where $\gamma \in \wp$ is sparsity pattern with respective to subset of $\boldsymbol{X}$ columns, $\boldsymbol{K}_\gamma \in \mathbb{R}^{d_\gamma \times d_\gamma}$ is positive definite, $\tilde{\boldsymbol{\beta}}_\gamma \in \mathbb{R}^{d_\gamma}$, and $\alpha_\gamma, \ g_0, \ \omega > 0$ are given.

Consider Bayes estimator $\hat{\boldsymbol{\psi}}$

$$\hat{\boldsymbol{\psi}} = \underset{\boldsymbol{\psi} \in \mathbb{R}^n}{\operatorname{argmin}} \, \mathbb{E} \left[L(\boldsymbol{\psi}, \boldsymbol{\mu}) | \boldsymbol{Y} \right] \tag{4.6}$$

where L is some loss function.

It is easy to see that

$$
\begin{aligned}
& \mathbb{E}\left[L(\boldsymbol{\psi},\boldsymbol{\mu})|\boldsymbol{Y}\right] \\
= & \sum_{\gamma\in\wp}\int L(\boldsymbol{\psi},\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)p(\boldsymbol{\beta}_\gamma|\boldsymbol{Y},\mathcal{M}_\gamma)p(\mathcal{M}_\gamma|\boldsymbol{Y})\;d\boldsymbol{\beta}_\gamma \\
= & \frac{1}{p(\boldsymbol{Y})}\sum_{\gamma\in\wp}\int L(\boldsymbol{\psi},\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)p(\boldsymbol{Y}|\mathcal{M}_\gamma,\boldsymbol{\beta}_\gamma)p(\boldsymbol{\beta}_\gamma|\mathcal{M}_\gamma)p(\mathcal{M}_\gamma)\;d\boldsymbol{\beta}_\gamma \\
= & \frac{(2\pi\omega^2)^{-n/2}}{p(\boldsymbol{Y})}\sum_{\gamma\in\wp}\pi_\gamma\int L(\boldsymbol{\psi},\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)\exp\left(-\frac{1}{2\omega^2}\|\boldsymbol{Y}-\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma\|_2^2\right)p(\boldsymbol{\beta}_\gamma|\mathcal{M}_\gamma)\;d\boldsymbol{\beta}_\gamma
\end{aligned}
$$

For exponentiated least square loss $L(\boldsymbol{\psi},\boldsymbol{\mu})=\exp\{\frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi}-\boldsymbol{\mu}\|_2^2\}$, the Bayes estimator is

$$
\boldsymbol{\psi}_X(\omega^2,\nu)=\underset{\boldsymbol{\psi}\in\mathbb{R}^n}{\operatorname{argmin}}\,J(\boldsymbol{\psi}) \tag{4.7}
$$

where

$$
J(\boldsymbol{\psi})=\sum_{\gamma\in\wp}\pi_\gamma\int\exp\left(\frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi}-\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma\|_2^2-\frac{1}{2\omega^2}\|\boldsymbol{Y}-\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma\|_2^2\right)p(\boldsymbol{\beta}_\gamma|\mathcal{M}_\gamma)\;\;d\boldsymbol{\beta}_\gamma\;. \tag{4.8}
$$

Under assumptions (4.3) and (4.4), since $p(\boldsymbol{\beta}_\gamma|\mathcal{M}_\gamma)=\int_0^\infty p(\boldsymbol{\beta}_\gamma|g_\gamma,\mathcal{M}_\gamma)p(g_\gamma|\mathcal{M}_\gamma)\;dg_\gamma$, with simple algebra we have

$$
\begin{aligned}
p(\boldsymbol{\beta}_\gamma|\mathcal{M}_\gamma) = & \left(2\pi\omega^2(d_\gamma/2+\alpha_\gamma)g_0\right)^{-d_\gamma/2}|\boldsymbol{K}_\gamma|^{1/2}\frac{\Gamma(d_\gamma/2+\alpha_\gamma)}{\Gamma(\alpha_\gamma)} \\
& \cdot\left(1+\frac{(\boldsymbol{\beta}_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)^\top\boldsymbol{K}_\gamma(\boldsymbol{\beta}_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)}{2(d_\gamma/2+\alpha_\gamma)g_0\omega^2}\right)^{-(d_\gamma/2+\alpha_\gamma)}
\end{aligned} \tag{4.9}
$$

The following theorem states that with a prior on g_γ for each model $\mathcal{M}_\gamma$, the prediction performance of the BMAX estimator $\boldsymbol{\psi}_X(\omega^2,\nu)$ is competitive to the BMAX estimator with any chosen g in Chapter 3.

Theorem 7. *Consider BMAX estimator $\boldsymbol{\psi}_X(\omega^2,\nu)$ as defined in (4.7) and (4.8), with Gaussian priors as (4.3) and (4.4). Assume $\nu\in(0,1)$ and if $\omega^2\geq\frac{\sigma^2}{\min(\nu,1-\nu)}$,* *we have oracle inequality*

$$
\begin{aligned}
&\|\boldsymbol{\psi}_X(\omega^2,\nu)-\boldsymbol{\eta}\|_2^2\\
&\le \min_{\substack{\gamma\in\wp\\ \boldsymbol{\beta}^*_\gamma\in\mathbb{R}^{d_\gamma}}} \left\{\|\boldsymbol{X}_\gamma\boldsymbol{\beta}^*_\gamma-\boldsymbol{\eta}\|_2^2+\frac{1}{g_0}(\boldsymbol{\beta}^*_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)^\top\boldsymbol{K}_\gamma(\boldsymbol{\beta}^*_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)+2\omega^2\log(\frac{1}{\pi_\gamma\delta})\right.\\
&\qquad\qquad \left.+\omega^2\log(|\nu g_0\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma\boldsymbol{K}_\gamma^{-1}+\boldsymbol{I}_{d_\gamma}|)+2\omega^2\log\left(\frac{(d_\gamma/2+\alpha_\gamma)^{d_\gamma/2}\Gamma(\alpha_\gamma)}{\Gamma(d_\gamma/2+\alpha_\gamma)}\right)\right\}\ ,
\end{aligned}
\tag{4.10}
$$

with probability at least $1-\delta$*. Moreover,*

$$
\begin{aligned}
&\mathbb{E}\|\boldsymbol{\psi}_X(\omega^2,\nu)-\boldsymbol{\eta}\|_2^2\\
&\le \min_{\substack{\gamma\in\wp\\ \boldsymbol{\beta}^*_\gamma\in\mathbb{R}^{d_\gamma}}} \left\{\|\boldsymbol{X}_\gamma\boldsymbol{\beta}^*_\gamma-\boldsymbol{\eta}\|_2^2+\frac{1}{g_0}(\boldsymbol{\beta}^*_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)^\top\boldsymbol{K}_\gamma(\boldsymbol{\beta}^*_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)+2\omega^2\log(\frac{1}{\pi_\gamma})\right.\\
&\qquad\qquad \left.+\omega^2\log(|\nu g_0\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma\boldsymbol{K}_\gamma^{-1}+\boldsymbol{I}_{d_\gamma}|)+2\omega^2\log\left(\frac{(d_\gamma/2+\alpha_\gamma)^{d_\gamma/2}\Gamma(\alpha_\gamma)}{\Gamma(d_\gamma/2+\alpha_\gamma)}\right)\right\}\ .
\end{aligned}
\tag{4.11}
$$

Remark 13. *For any given* $g>0$ *in Theorem 5, we can simply set the hyper prior (4.4) with* $g_0=g$*, then the upper bounds of oracle inequalities in Theorem 5 and Theorem 7 becomes almost same with only one additional term of* $2\omega^2\log\left(\frac{(d_\gamma/2+\alpha_\gamma)^{d_\gamma/2}\Gamma(\alpha_\gamma)}{\Gamma(d_\gamma/2+\alpha_\gamma)}\right)$ *in (4.10) and (4.11) of Theorem 7. And since* $\Gamma(z)\approx z^{z-1/2}e^{-z}\sqrt{2\pi}$*,*

$$
\begin{aligned}
\log\left(\frac{(d_\gamma/2+\alpha_\gamma)^{d_\gamma/2}\Gamma(\alpha_\gamma)}{\Gamma(d_\gamma/2+\alpha_\gamma)}\right) &\approx \log\left(\frac{(d_\gamma/2+\alpha_\gamma)^{d_\gamma/2}\Gamma(\alpha_\gamma)}{(d_\gamma/2+\alpha_\gamma)^{d_\gamma/2+\alpha_\gamma-1/2}e^{-(d_\gamma/2+\alpha_\gamma)}\sqrt{2\pi}}\right)\\
&= \log\left((d_\gamma/2+\alpha_\gamma)^{1/2}e^{d_\gamma/2+\alpha_\gamma}\frac{\Gamma(\alpha_\gamma)}{\sqrt{2\pi}}\right)=O(d_\gamma)\ ,
\end{aligned}
$$

thus when d is fixed, aggregate by BMAX in linear models with mixture of g-priors is competitive to that with any fixed g chosen in Chapter 3 on the prediction accuracy on model averaging problem.

4.2 Gradient Descent Algorithm for Solving BMAX in Linear Models with Gaussian Priors and Priors on g

Since we can change the order of ∇^2 and $\int(\cdot)dg$, below we listed a corollary with results paralleled to those of Lemma 4 and Lemma 5.

Corollary 5. *For any $\boldsymbol{\psi} \in \mathbb{R}^n$ we have*

$$\nabla^2 \log J(\boldsymbol{\psi}) \geq A_1 \boldsymbol{I}_n \ , \tag{4.12}$$

and

$$\log J(\boldsymbol{\psi}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu)) \leq \frac{1}{2A_1} \|\nabla \log J(\boldsymbol{\psi})\|_2^2 \ , \tag{4.13}$$

and if condition (3.19) *holds, then for $k > 0$ and $0 \leq \alpha \leq 1$*

$$\nabla^2 \log J((1-\alpha)\boldsymbol{\psi}^{(k-1)} + \alpha \boldsymbol{\psi}^{(k)}) \leq A_3 \boldsymbol{I}_n \ , \tag{4.14}$$

where A_1 and A_3 are defined as (2.25) *and* (3.20)*.*

Below we propose a gradient descent algorithm to solve $\boldsymbol{\psi}_X(\omega^2, \nu)$ defined by (4.7) and (4.8).

Algorithm 7 Gradient Descent Algorithm to solve BMAX in Linear Models with Gaussian Priors and priors on g (GD-BMAX-LM-g)

```
Input: Noisy observation $\boldsymbol{Y}$; for $\gamma \in \wp$, $\boldsymbol{X}_\gamma \in \mathbb{R}^{n \times d_\gamma}$ and $K_\gamma \in \mathbb{R}^{d_\gamma \times d_\gamma}$;prior
  $\boldsymbol{\pi} \in \Lambda^{|\wp|}$; parameters $g_0, \nu, \omega^2$.
Output: Aggregate estimator $\boldsymbol{\psi}^{(k)}$.
  Initialize $\boldsymbol{\psi}^{(0)} = 0$.
  for $k = 1, 2, \ldots$ do
    Choose step size $t_k = s \in (0, 2/A_3)$.
    Calculate
        $\boldsymbol{\psi}^{(k)} = \boldsymbol{\psi}^{(k-1)} - t_k \nabla \log J(\boldsymbol{\psi}^{(k-1)})$ .
  end for
```

With Corollary 5 we can show that Algorithm 7 converges to the minimum of $\log J(\boldsymbol{\psi})$ in a geographic rate.

Proposition 11. *Given condition* (3.19) *is satisfied, for* $\boldsymbol{\psi}^{(k)}$ *output from Algorithm 7 and choose fixed step size* $t_k = s \in (0, 2/A_3)$ *for* $k > 0$*, then*

$$\begin{aligned} & \log J(\boldsymbol{\psi}^{(k)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu)) \\ \leq \ & [1 - 2A_1(s - (A_3/2)s^2)]^k \left(\log J(\boldsymbol{\psi}^{(0)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))\right) \end{aligned}$$

Remark 14. *We can simply take* $s = 1/A_3$ *to minimize the right hand side of above inequality, it results that*

$$\log J(\boldsymbol{\psi}^{(k)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu)) \leq (1 - A_1/A_3)^k \left(\log J(\boldsymbol{\psi}^{(0)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))\right) \ .$$

The proof of Proposition 11 is almost the same as that of Proposition 10. And it states that Algorithm 7 converges to the minimum of $\log J(\boldsymbol{\psi})$ with a geographic rate. The calculation of $\nabla \log J(\boldsymbol{\psi}^{(k-1)})$ is similar to that of Algorithm 5, the only difference is that we need to integrate with priors on g, which can be done by Monte Carlo methods, namely, sample $\{g_1, \ldots, g_N\}$ from priors, and with each fixed g_ℓ ($\ell = 1, \ldots, N$), the calculation is exactly the same as Algorithm 5, then we just need to average the results over sample $\{g_1, \ldots, g_N\}$ to approximate $\nabla \log J(\boldsymbol{\psi}^{(k-1)})$.

Appendix A
Proofs

A.1 Proof of Proposition 1

Note first that by homogeneity, one may assume that $\sigma = 1$. Moreover, write for simplicity $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{\text{EXP}}$. If we assume $\lambda_1 \le 1/2$, we obtain

$$\begin{aligned}
\|\mathsf{f}_{\boldsymbol{\lambda}}\|_2^2 - \|\boldsymbol{f}_1\|_2^2 &\ge |\lambda_1 \boldsymbol{f}_1 + (1-\lambda_1)\boldsymbol{f}_2|_2^2 - |\boldsymbol{f}_1|_2^2 \\
&= (1-\lambda_1)^2 |\boldsymbol{f}_2|_2^2 - (1-\lambda_1^2)|\boldsymbol{f}_1|_2^2 \\
&\ge 2(1-\lambda_1)^2\sqrt{n} + [(1-\lambda_1)^2 - (1-\lambda_1^2)]n \\
&\ge \sqrt{n}/2 - 2\lambda_1 n\,.
\end{aligned} \tag{A.1}$$

We first treat the low temperature case where ω is chosen as in (2.7). Define the event

$$E = \{n\widehat{\text{MSE}}(\boldsymbol{f}_2) + 2\sqrt{n} \le n\widehat{\text{MSE}}(\boldsymbol{f}_1)\}\,,$$

and observe that $\boldsymbol{\eta} \equiv 0$ gives

$$E = \left\{2\langle \boldsymbol{f}_2 - \boldsymbol{f}_1, \boldsymbol{\xi}\rangle_2 \ge \|\boldsymbol{f}_2\|_2^2 - \|\boldsymbol{f}_1\|_2^2 + 2\sqrt{n}\right\}\ . \tag{A.2}$$

On the one hand, we have $\|\boldsymbol{f}_2\|_2^2 - \|\boldsymbol{f}_1\|_2^2 = 1 + 2\sqrt{n}$ and on the other hand

$$\|\boldsymbol{f}_2 - \boldsymbol{f}_1\|_2^2 = \|\boldsymbol{f}_2\|_2^2 + \|\boldsymbol{f}_1\|_2^2 = (2n + 2\sqrt{n} + 1) \ge \frac{1}{8}(1 + 4\sqrt{n})^2\,.$$

Thus, we have

$$\mathbb{P}(E) \ge \mathbb{P}(2\langle \boldsymbol{f}_2 - \boldsymbol{f}_1, \boldsymbol{\xi}\rangle_2 \ge 2\sqrt{2}\|\boldsymbol{f}_2 - \boldsymbol{f}_1\|_2) = \mathbb{P}(Z \ge \sqrt{2}) \ge 0.07\,. \tag{A.3}$$

where $Z \sim \mathcal{N}(0,1)$. In view of (2.4), on the event E, we have

$$\lambda_1 \le \lambda_2 e^{-\frac{1}{\omega^2}\sqrt{n}} \le \frac{1}{8\sqrt{n}} \le \frac{1}{2}\,,$$

for low temperature ω chosen as in (2.7). Together with (A.1), it yields

$$\|\mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2 - \|\boldsymbol{f}_1\|^2 \ge \frac{\sqrt{n}}{4}.$$

We now turn to the case of potentially high temperatures. Actually, the following proof holds for *any* temperature ω as long as the α_js are chosen small enough. In this case, we can expect the M exponential weights to take comparable values. To that end, define for each $j = 2, \ldots, M$, the event

$$F_j = \left\{\widehat{\mathrm{MSE}}(\boldsymbol{f}_j) \le \widehat{\mathrm{MSE}}(\boldsymbol{f}_1)\right\}\,,$$

Define $F = \bigcap_{j=2}^M F_j$ and denote by F_j^c the complement of F_j. Recall that $\|\boldsymbol{f}_j\|_2^2 = \|\boldsymbol{f}_2\|_2^2 + \alpha_j^2$ so that

$$\begin{aligned} F_j^c &= \left\{2\langle \boldsymbol{f}_j - \boldsymbol{f}_1, \boldsymbol{\xi}\rangle_2 \le \|\boldsymbol{f}_j\|_2^2 - \|\boldsymbol{f}_1\|_2^2\right\} \\ &= \left\{2\langle \boldsymbol{f}_2 - \boldsymbol{f}_1, \boldsymbol{\xi}\rangle_2 + 2\langle \boldsymbol{f}_j - \boldsymbol{f}_2, \boldsymbol{\xi}\rangle_2 \le \|\boldsymbol{f}_2\|_2^2 - \|\boldsymbol{f}_1\|_2^2 + \alpha_j^2\right\} \\ &\subset E^c \cup G_j\,, \end{aligned}$$

where the E is defined in (A.2) and G_j is defined as

$$G_j = \left\{2\langle \boldsymbol{f}_j - \boldsymbol{f}_2, \boldsymbol{\xi}\rangle_2 \le \alpha_j^2 - 2\sqrt{n}\right\}\,.$$

In view of (2.8), we have

$$\mathbb{P}(G_j) \le \mathbb{P}\left(2\langle \boldsymbol{f}_j - \boldsymbol{f}_2, \boldsymbol{\xi}\rangle_2 \le -\alpha_j^2\right) \le \mathbb{P}\left(Z \ge \sqrt{2\log(100M)}\right) \le \frac{0.01}{M}\,.$$

Therefore,

$$\mathbb{P}(F^c) \le \mathbb{P}(E^c) + \sum_{j=2}^M \mathbb{P}(G_j) \le 0.93 + 0.01 = 0.94\,.$$

Note now that on the event F, for any $j = 2, \ldots, M$, we have $\lambda_j \ge \lambda_1$ so that $\lambda_1 \le 1/M \le 1/2$. Together with (A.1), it yields

$$\|\mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2 - \|\boldsymbol{f}_1\|_2^2 \ge \frac{\sqrt{n}}{2} - \frac{2n}{M} \ge \frac{\sqrt{n}}{4}\,,$$

where, in the last inequality, we used the fact that $M \ge 8\sqrt{n}$. ∎

A.2 Proof of Proposition 2

Note first that by homogeneity, one may assume that $\sigma = 1$. Next, observe that $\mathfrak{f}_{\boldsymbol{\lambda}^{\text{PROJ}}} = (\mathsf{P}_m\boldsymbol{\xi}, 0, \ldots, 0)^\top \in \mathbb{R}^n$, where $\mathsf{P}_m\boldsymbol{\xi} \in \mathbb{R}^m$ is the projection of $\tilde{\boldsymbol{\xi}} = (\xi_1, \ldots, \xi_m)^\top$ onto $\mathcal{B}_1^m(\sqrt{n})$, the ℓ_1-ball of $\mathbb{R}^m$ with radius $\sqrt{n}$.

Let E denote the event on which $\|\tilde{\boldsymbol{\xi}}\|_1 \le \sqrt{n}$ and observe that, on this event, we have $\mathsf{P}_m\boldsymbol{\xi} = \tilde{\boldsymbol{\xi}}$. It yields

$$n\,\mathrm{MSE}(\mathfrak{f}_{\boldsymbol{\lambda}^{\text{PROJ}}}) = \sum_{j=1}^{m} \xi_j^2 = \|\tilde{\boldsymbol{\xi}}\|_2^2\,,$$

Let now F denote the event on which $\|\tilde{\boldsymbol{\xi}}\|_2^2 \ge m/2$ and note that on $E \cap F$, it holds

$$\mathrm{MSE}(\mathfrak{f}_{\boldsymbol{\lambda}^{\text{PROJ}}}) \ge \frac{m}{2n} \ge \sqrt{\frac{1}{13n}}$$

To conclude our proof, it remains to bound from below the probability of $E \cap F$. The bounds below follow from the fact that $\|\tilde{\boldsymbol{\xi}}\|_2^2$ follows a chi-squared distribution with m degrees of freedom. We begin by the event E. Using Hölder's inequality, we have

$$\mathbb{P}(E^c) \le \mathbb{P}\big(\|\tilde{\boldsymbol{\xi}}\|_2^2 \ge \frac{n}{m}\big) = \mathbb{P}\big(\|\tilde{\boldsymbol{\xi}}\|_2^2 - \mathbb{E}\|\tilde{\boldsymbol{\xi}}\|_2^2 \ge \frac{n}{m} - m\big)$$

Next, using the fact that $m^2 \le 8n/13$ together with Laurent and Massart (2000a, Lemma 1) we get

$$\mathbb{P}(E^c) \le \mathbb{P}\big(\|\tilde{\boldsymbol{\xi}}\|_2^2 - \mathbb{E}\|\tilde{\boldsymbol{\xi}}\|_2^2 \ge \frac{5m}{8}\big) \le e^{-m/16}\,.$$

Moreover, using Laurent and Massart (2000a, Lemma 1), we also get that

$$\mathbb{P}(F^c) = \mathbb{P}\big(\|\tilde{\boldsymbol{\xi}}\|_2^2 - \mathbb{E}\|\tilde{\boldsymbol{\xi}}\|_2^2 \le -\frac{m}{2}\big) \le e^{-m/16}\,.$$

Therefore, since $n \ge 416$ implies $m \ge 16$, we get

$$\mathbb{P}(E \cap F) \ge 1 - \mathbb{P}(E^c) - \mathbb{P}(F^c) \ge 1 - 2e^{-m/16} \ge 1 - 2/e \ge 1/4\,.$$

$\blacksquare$

A.3 Proof of Theorem 1

Proposition 12. *For any $\boldsymbol{\lambda} \in \Lambda^M$, real sequence $\{x_j\}_{j=1}^M$, and $a > 0$, we have*

$$\sum_{j=1}^{M} \lambda_j x_j - a\mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi}) \le a \log\left(\sum_{j=1}^{M} \pi_j e^{x_j/a}\right).$$

Proof. The result follows directly from Jesen's Inequality as

$$\exp\left(\sum_{j=1}^{M} \lambda_j((x_j/a) - \log(\lambda_j/\pi_j))\right) \le \sum_{j=1}^{M} \pi_j e^{x_j/a}$$

□

We also need the following lemma to prove the theorem.

Lemma 6. *For any $\boldsymbol{\psi} \in \mathbb{R}^n$, let $\boldsymbol{\lambda} \in \Lambda^M$ defined as*

$$\lambda_j \propto \pi_j \exp\left(-\frac{1}{2\omega^2}\|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2 + \frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi} - \boldsymbol{f}_j\|_2^2\right)$$

Then we have the following equation

$$\frac{\nabla J(\boldsymbol{\psi})}{J(\boldsymbol{\psi})} = \frac{1-\nu}{\omega^2}(\boldsymbol{\psi} - \mathfrak{f}_{\boldsymbol{\lambda}}),$$

and

$$\begin{aligned}
& \|\mathfrak{f}_{\boldsymbol{\lambda}} - \boldsymbol{\eta}\|_2^2 - \left(\nu \sum_{j=1}^{M} \theta_j \|\boldsymbol{f}_j - \boldsymbol{\eta}\|_2^2 + (1-\nu)\|\mathfrak{f}_{\boldsymbol{\theta}} - \boldsymbol{\eta}\|_2^2\right) \\
= \quad & -\nu \sum_{j=1}^{M} \lambda_j \|\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2 - (1-\nu)\|\mathfrak{f}_{\boldsymbol{\theta}} - \mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2 + 2\boldsymbol{\xi}^\top(\mathfrak{f}_{\boldsymbol{\lambda}} - \mathfrak{f}_{\boldsymbol{\theta}}) - 2\omega^2\mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi}) \\
& + 2\omega^2\mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\pi}) - 2\omega^2\mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\lambda}) + 2(1-\nu)(\mathfrak{f}_{\boldsymbol{\theta}} - \mathfrak{f}_{\boldsymbol{\lambda}})^\top(\mathfrak{f}_{\boldsymbol{\lambda}} - \boldsymbol{\psi}).
\end{aligned}$$

Proof. Since

$$J(\boldsymbol{\psi}) = \sum_{j=1}^{M} \pi_j \exp\left(-\frac{1}{2\omega^2}\|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2 + \frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi} - \boldsymbol{f}_j\|_2^2\right),$$

and

$$\nabla J(\boldsymbol{\psi}) = \sum_{j=1}^{M} \pi_j \exp\left(-\frac{1}{2\omega^2}\|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2 + \frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi} - \boldsymbol{f}_j\|_2^2\right) \frac{1-\nu}{\omega^2}(\boldsymbol{\psi} - \boldsymbol{f}_j),$$

Then we have

$$\frac{\nabla J(\boldsymbol{\psi})}{J(\boldsymbol{\psi})} = \frac{1-\nu}{\omega^2}(\boldsymbol{\psi} - \mathfrak{f}_{\boldsymbol{\lambda}}).$$

From the definition of $\boldsymbol{\lambda}$ we have,

$$\frac{\lambda_i}{\lambda_j} = \frac{\pi_i \exp\left(-\frac{1}{2\omega^2}\|\boldsymbol{f}_i - \boldsymbol{Y}\|_2^2 + \frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi} - \boldsymbol{f}_i\|_2^2\right)}{\pi_j \exp\left(-\frac{1}{2\omega^2}\|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2 + \frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi} - \boldsymbol{f}_j\|_2^2\right)}$$

then it follows that

$$\begin{aligned} & 2\omega^2 \log(\lambda_i/\pi_i) + \|\boldsymbol{f}_i - \boldsymbol{Y}\|_2^2 - (1-\nu)\|\boldsymbol{\psi} - \boldsymbol{f}_i\|_2^2 \\ = \; & 2\omega^2 \log(\lambda_j/\pi_j) + \|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2 - (1-\nu)\|\boldsymbol{\psi} - \boldsymbol{f}_j\|_2^2 \end{aligned}$$

Sum up each hand side of the above equation with weight $\boldsymbol{\lambda}$ and any chosen $\boldsymbol{\theta} \in \Lambda^M$,

$$\begin{aligned} & \sum_{j=1}^{M} \lambda_j \|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2 - (1-\nu)\sum_{j=1}^{M} \lambda_j \|\boldsymbol{f}_j - \boldsymbol{\psi}\|_2^2 + 2\omega^2 \sum_{j=1}^{M} \lambda_j \log(\lambda_j/\pi_j) \\ = \; & \sum_{j=1}^{M} \theta_j \|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2 - (1-\nu)\sum_{j=1}^{M} \theta_j \|\boldsymbol{f}_j - \boldsymbol{\psi}\|_2^2 + 2\omega^2 \sum_{j=1}^{M} \theta_j \log(\lambda_j/\pi_j) \end{aligned}$$

Combine the above equation and the following facts that

$$\sum_{j=1}^{M} \lambda_j \|\boldsymbol{f}_j - \boldsymbol{\psi}\|_2^2 = \|\mathfrak{f}_{\boldsymbol{\lambda}} - \boldsymbol{\psi}\|_2^2 + \sum_{j=1}^{M} \lambda_j \|\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2$$

and

$$\begin{aligned} \sum_{j=1}^{M} \theta_j \|\boldsymbol{f}_j - \boldsymbol{\psi}\|_2^2 &= \|\mathfrak{f}_{\boldsymbol{\theta}} - \boldsymbol{\psi}\|_2^2 + \sum_{j=1}^{M} \theta_j \|\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\theta}}\|_2^2 \\ &= \|\mathfrak{f}_{\boldsymbol{\theta}} - \mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2 + \|\mathfrak{f}_{\boldsymbol{\lambda}} - \boldsymbol{\psi}\|_2^2 - 2(\mathfrak{f}_{\boldsymbol{\theta}} - \mathfrak{f}_{\boldsymbol{\lambda}})^\top (\mathfrak{f}_{\boldsymbol{\lambda}} - \boldsymbol{\psi}) \\ &\quad + \sum_{j=1}^{M} \theta_j \|\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\theta}}\|_2^2 \end{aligned}$$

we have

$$\begin{aligned}
& \sum_{j=1}^{M}\lambda_j\|\boldsymbol{f}_j-\boldsymbol{Y}\|_2^2-(1-\nu)\sum_{j=1}^{M}\lambda_j\|\boldsymbol{f}_j-\mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2+2\omega^2\mathcal{K}(\boldsymbol{\lambda},\boldsymbol{\pi}) \\
= & \sum_{j=1}^{M}\theta_j\|\boldsymbol{f}_j-\boldsymbol{Y}\|_2^2-(1-\nu)\sum_{j=1}^{M}\theta_j\|\boldsymbol{f}_j-\mathfrak{f}_{\boldsymbol{\theta}}\|_2^2+2\omega^2\mathcal{K}(\boldsymbol{\theta},\boldsymbol{\pi})-2\omega^2\mathcal{K}(\boldsymbol{\theta},\boldsymbol{\lambda}) \\
& -(1-\nu)\|\mathfrak{f}_{\boldsymbol{\theta}}-\mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2+2(1-\nu)(\mathfrak{f}_{\boldsymbol{\theta}}-\mathfrak{f}_{\boldsymbol{\lambda}})^\top(\mathfrak{f}_{\boldsymbol{\lambda}}-\boldsymbol{\psi})
\end{aligned}$$

Plug the following two equations to each hand side of (A.4)

$$\sum_{j=1}^{M}\lambda_j\|\boldsymbol{f}_j-\boldsymbol{Y}\|_2^2-(1-\nu)\sum_{j=1}^{M}\lambda_j\|\boldsymbol{f}_j-\mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2=\|\mathfrak{f}_{\boldsymbol{\lambda}}-\boldsymbol{Y}\|_2^2+\nu\sum_{j=1}^{M}\lambda_j\|\boldsymbol{f}_j-\mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2$$

$$\sum_{j=1}^{M}\theta_j\|\boldsymbol{f}_j-\boldsymbol{Y}\|_2^2-(1-\nu)\sum_{j=1}^{M}\theta_j\|\boldsymbol{f}_j-\mathfrak{f}_{\boldsymbol{\theta}}\|_2^2=\nu\sum_{j=1}^{M}\theta_j\|\boldsymbol{f}_j-\boldsymbol{Y}\|_2^2+(1-\nu)\|\mathfrak{f}_{\boldsymbol{\theta}}-\boldsymbol{Y}\|_2^2$$

and rearrange the terms we have

$$\begin{aligned}
& \|\mathfrak{f}_{\boldsymbol{\lambda}}-\boldsymbol{Y}\|_2^2-\left(\nu\sum_{j=1}^{M}\theta_j\|\boldsymbol{f}_j-\boldsymbol{Y}\|_2^2+(1-\nu)\|\mathfrak{f}_{\boldsymbol{\theta}}-\boldsymbol{Y}\|_2^2\right) \\
= & -\nu\sum_{j=1}^{M}\lambda_j\|\boldsymbol{f}_j-\mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2-(1-\nu)\|\mathfrak{f}_{\boldsymbol{\theta}}-\mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2-2\omega^2\mathcal{K}(\boldsymbol{\lambda},\boldsymbol{\pi})+2\omega^2\mathcal{K}(\boldsymbol{\theta},\boldsymbol{\pi}) \\
& -2\omega^2\mathcal{K}(\boldsymbol{\theta},\boldsymbol{\lambda})+2(1-\nu)(\mathfrak{f}_{\boldsymbol{\theta}}-\mathfrak{f}_{\boldsymbol{\lambda}})^\top(\mathfrak{f}_{\boldsymbol{\lambda}}-\boldsymbol{\psi})
\end{aligned}$$

then by combining the above equation with $\boldsymbol{Y}=\boldsymbol{\eta}+\boldsymbol{\xi}$ it follows that

$$\begin{aligned}
& \|\mathfrak{f}_{\boldsymbol{\lambda}}-\boldsymbol{\eta}\|_2^2-\left(\nu\sum_{j=1}^{M}\theta_j\|\boldsymbol{f}_j-\boldsymbol{\eta}\|_2^2+(1-\nu)\|\mathfrak{f}_{\boldsymbol{\theta}}-\boldsymbol{\eta}\|_2^2\right) \\
= & -\nu\sum_{j=1}^{M}\lambda_j\|\boldsymbol{f}_j-\mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2-(1-\nu)\|\mathfrak{f}_{\boldsymbol{\theta}}-\mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2+2\boldsymbol{\xi}^\top(\mathfrak{f}_{\boldsymbol{\lambda}}-\mathfrak{f}_{\boldsymbol{\theta}})-2\omega^2\mathcal{K}(\boldsymbol{\lambda},\boldsymbol{\pi}) \\
& +2\omega^2\mathcal{K}(\boldsymbol{\theta},\boldsymbol{\pi})-2\omega^2\mathcal{K}(\boldsymbol{\theta},\boldsymbol{\lambda})+2(1-\nu)(\mathfrak{f}_{\boldsymbol{\theta}}-\mathfrak{f}_{\boldsymbol{\lambda}})^\top(\mathfrak{f}_{\boldsymbol{\lambda}}-\boldsymbol{\psi})
\end{aligned}$$

□

Now we are ready to prove Theorem 1.

From the definition of $\boldsymbol{\psi}_X(\omega^2,\nu)$ (2.15), $\boldsymbol{\psi}_X(\omega^2,\nu)$ is the minimizer of $J(\boldsymbol{\psi})$, thus $\nabla J(\boldsymbol{\psi}_X(\omega^2,\nu))=0$. By using Lemma 6, $\boldsymbol{\psi}_X(\omega^2,\nu)=\mathfrak{f}_{\boldsymbol{\lambda}}$ with $\boldsymbol{\lambda}\in\Lambda^M$

defined as

$$\lambda_j \propto \pi_j \exp\left(-\frac{1}{2\omega^2}\|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2 + \frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi}_X(\omega^2,\nu) - \boldsymbol{f}_j\|_2^2\right).$$

Also from Lemma 6 we have

$$\begin{aligned} & \|\mathfrak{f}_{\boldsymbol{\lambda}} - \boldsymbol{\eta}\|_2^2 - \left(\nu\sum_{j=1}^{M}\theta_j\|\boldsymbol{f}_j - \boldsymbol{\eta}\|_2^2 + (1-\nu)\|\mathfrak{f}_{\boldsymbol{\theta}} - \boldsymbol{\eta}\|_2^2\right) \\ = \;& -\nu\sum_{j=1}^{M}\lambda_j\|\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2 - (1-\nu)\|\mathfrak{f}_{\boldsymbol{\theta}} - \mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2 + 2\boldsymbol{\xi}^\top(\mathfrak{f}_{\boldsymbol{\lambda}} - \mathfrak{f}_{\boldsymbol{\theta}}) - 2\omega^2\mathcal{K}(\boldsymbol{\lambda},\boldsymbol{\pi}) \\ & +2\omega^2\mathcal{K}(\boldsymbol{\theta},\boldsymbol{\pi}) - 2\omega^2\mathcal{K}(\boldsymbol{\theta},\boldsymbol{\lambda}) \end{aligned}$$

It is also easy to verify the following inequality

$$-\nu\sum_{j=1}^{M}\lambda_j\|\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2 - (1-\nu)\|\mathfrak{f}_{\boldsymbol{\theta}} - \mathfrak{f}_{\boldsymbol{\lambda}}\|_2^2 \le -\nu_1\sum_{j=1}^{M}\lambda_j\|\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\theta}}\|_2^2$$

where $\nu_1 = \min(\nu, 1-\nu)$.

Combining the above inequality and $2\omega^2\mathcal{K}(\boldsymbol{\theta},\boldsymbol{\lambda}) \ge 0$ we have

$$\begin{aligned} & \|\mathfrak{f}_{\boldsymbol{\lambda}} - \boldsymbol{\eta}\|_2^2 - \left(\nu\sum_{j=1}^{M}\theta_j\|\boldsymbol{f}_j - \boldsymbol{\eta}\|_2^2 + (1-\nu)\|\mathfrak{f}_{\boldsymbol{\theta}} - \boldsymbol{\eta}\|_2^2\right) \\ \le \;& -\nu_1\sum_{j=1}^{M}\lambda_j\|\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\theta}}\|_2^2 + 2\boldsymbol{\xi}^\top(\mathfrak{f}_{\boldsymbol{\lambda}} - \mathfrak{f}_{\boldsymbol{\theta}}) - 2\omega^2\mathcal{K}(\boldsymbol{\lambda},\boldsymbol{\pi}) + 2\omega^2\mathcal{K}(\boldsymbol{\theta},\boldsymbol{\pi}) \\ = \;& \sum_{j=1}^{M}\lambda_j\left(-\nu_1\|\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\theta}}\|_2^2 + 2\boldsymbol{\xi}^\top(\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\theta}})\right) - 2\omega^2\mathcal{K}(\boldsymbol{\lambda},\boldsymbol{\pi}) + 2\omega^2\mathcal{K}(\boldsymbol{\theta},\boldsymbol{\pi}) \\ \le \;& 2\omega^2\log\left(\sum_{j=1}^{M}\pi_j\exp\left\{\frac{-\nu_1\|\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\theta}}\|_2^2 + 2\boldsymbol{\xi}^\top(\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\theta}})}{2\omega^2}\right\}\right) + 2\omega^2\mathcal{K}(\boldsymbol{\theta},\boldsymbol{\pi}) \end{aligned}$$

where the last inequality is using Proposition 12 with $x_j = -\nu_1\|\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\theta}}\|_2^2 + 2\boldsymbol{\xi}^\top(\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\theta}})$ and $a = 2\omega^2$.

Put expectation for $\boldsymbol{\xi}$ at each side of the above inequality,

$$\begin{aligned}
& \mathbb{E}\|\mathfrak{f}_{\boldsymbol{\lambda}} - \boldsymbol{\eta}\|_2^2 - \left(\nu \sum_{j=1}^{M} \theta_j \|\boldsymbol{f}_j - \boldsymbol{\eta}\|_2^2 + (1-\nu)\|\mathfrak{f}_{\boldsymbol{\theta}} - \boldsymbol{\eta}\|_2^2\right) \\
\leq\ & 2\omega^2 \mathbb{E} \log \left(\sum_{j=1}^{M} \pi_j \exp\left\{ \frac{-\nu_1 \|\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\theta}}\|_2^2 + 2\boldsymbol{\xi}^\top (\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\theta}})}{2\omega^2} \right\} \right) + 2\omega^2 \mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\pi}) \\
\leq\ & 2\omega^2 \log \left(\sum_{j=1}^{M} \pi_j \mathbb{E} \exp\left\{ \frac{-\nu_1 \|\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\theta}}\|_2^2 + 2\boldsymbol{\xi}^\top (\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\theta}})}{2\omega^2} \right\} \right) + 2\omega^2 \mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\pi}) \\
\leq\ & 2\omega^2 \log \left(\sum_{j=1}^{M} \pi_j \exp\left\{ (-\nu_1 + \sigma^2/\omega^2) \frac{\|\boldsymbol{f}_j - \mathfrak{f}_{\boldsymbol{\theta}}\|_2^2}{2\omega^2} \right\} \right) + 2\omega^2 \mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\pi}) \\
\leq\ & 2\omega^2 \mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\pi})
\end{aligned}$$

where the second inequality is from concavity of $\log(t)$, the third comes from Gaussian assumption, and the last one is because of assumption $\omega^2 \geq \sigma^2/\nu_1 = \frac{\sigma^2}{\min(\nu, 1-\nu)}$.

Also by Chernoff bound with probability at least $1-\delta$,

$$\|\mathfrak{f}_{\boldsymbol{\lambda}} - \boldsymbol{\eta}\|_2^2 - \left(\nu \sum_{j=1}^{M} \theta_j \|\boldsymbol{f}_j - \boldsymbol{\eta}\|_2^2 + (1-\nu)\|\mathfrak{f}_{\boldsymbol{\theta}} - \boldsymbol{\eta}\|_2^2\right) \leq 2\omega^2 \mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\pi}) + 2\omega^2 \log(1/\delta).$$

■

A.4 Proof of Lemma 1

Define $\boldsymbol{\lambda} \in \Lambda^M$ as

$$\lambda_j \propto \pi_j \exp\left(-\frac{1}{2\omega^2} \|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2 + \frac{1-\nu}{2\omega^2} \|\boldsymbol{\psi} - \boldsymbol{f}_j\|_2^2 \right)$$

It follows that

$$\frac{\nabla J(\boldsymbol{\psi})}{J(\boldsymbol{\psi})} = \frac{1-\nu}{\omega^2} (\boldsymbol{\psi} - \mathfrak{f}_{\boldsymbol{\lambda}})$$

and

$$\frac{\nabla^2 J(\boldsymbol{\psi})}{J(\boldsymbol{\psi})} = \sum_{j=1}^{M} \lambda_j \left(\left(\frac{1-\nu}{\omega^2}\right)^2 (\boldsymbol{\psi} - \boldsymbol{f}_j)(\boldsymbol{\psi} - \boldsymbol{f}_j)^\top + \left(\frac{1-\nu}{\omega^2}\right) \boldsymbol{I}_n \right)$$

Then we have

$$\begin{aligned}\nabla^2 \log J(\boldsymbol{\psi}) &= \frac{(\nabla^2 J(\boldsymbol{\psi}))J(\boldsymbol{\psi}) - (\nabla J(\boldsymbol{\psi}))(\nabla J(\boldsymbol{\psi}))^\top}{J^2(\boldsymbol{\psi})} \\ &= \sum_{j=1}^{M} \lambda_j \left(\left(\frac{1-\nu}{\omega^2}\right)^2 (\boldsymbol{\psi} - \boldsymbol{f}_j)(\boldsymbol{\psi} - \boldsymbol{f}_j)^\top + \left(\frac{1-\nu}{\omega^2}\right) \boldsymbol{I}_n \right) \\ &\quad - \left(\frac{1-\nu}{\omega^2}\right)^2 (\boldsymbol{\psi} - \mathfrak{f}_{\boldsymbol{\lambda}})(\boldsymbol{\psi} - \mathfrak{f}_{\boldsymbol{\lambda}})^\top \\ &= \left(\frac{1-\nu}{\omega^2}\right) \boldsymbol{I}_n + \left(\frac{1-\nu}{\omega^2}\right)^2 FAF^\top\end{aligned}$$

where $F = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M) \in \mathbb{R}^{n\times M}$ and $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_M) - \boldsymbol{\lambda}\boldsymbol{\lambda}^\top \geq 0$.

Therefore $\nabla^2 \log J(\boldsymbol{\psi}) \geq \left(\frac{1-\nu}{\omega^2}\right) \boldsymbol{I}_n$.

With assumption that $\|\boldsymbol{f}_j\|_2 \leq L$ for all j, for any $u \in \mathbb{R}^n$,

$$\begin{aligned}u^\top F\operatorname{diag}(\lambda_1, \ldots, \lambda_M)Fu &= \sum_{j=1}^{M} \lambda_j u^\top \boldsymbol{f}_j \boldsymbol{f}_j^\top u = \sum_{j=1}^{M} \lambda_j (\boldsymbol{f}_j^\top u)^2 \\ &\leq \sum_{j=1}^{M} \lambda_j (\|\boldsymbol{f}_j\|_2 \|u\|_2)^2 \leq L^2 \|u\|_2^2\end{aligned}$$

which results that

$$FAF^\top \leq F\operatorname{diag}(\lambda_1, \ldots, \lambda_M)F^\top \leq L^2 \boldsymbol{I}_n,$$

and it follows that $\nabla^2 \log J(\boldsymbol{\psi}) \leq \left(\left(\frac{1-\nu}{\omega^2}\right) + \left(\frac{1-\nu}{\omega^2}\right)^2 L^2\right) \boldsymbol{I}_n$.

∎

A.5 Proof of Lemma 2

From inequality (2.28), for any $\boldsymbol{\psi}_1 \in \mathbb{R}^n$ we have

$$\log J(\boldsymbol{\psi}_1) \geq \log J(\boldsymbol{\psi}_2) + (\boldsymbol{\psi}_1 - \boldsymbol{\psi}_2)^\top \frac{\nabla J(\boldsymbol{\psi}_2)}{J(\boldsymbol{\psi}_2)} + (A_1/2)\|\boldsymbol{\psi}_1 - \boldsymbol{\psi}_2\|_2^2$$

The righthand side of the above inequality is a convex quadratic function of $\boldsymbol{\psi}_1$ (for fixed $\boldsymbol{\psi}_2$). Setting the gradient with respect to $\boldsymbol{\psi}_1$ equal to zero, we find that

$\tilde{\boldsymbol{\psi}}_1 = \boldsymbol{\psi}_2 - (1/A_1)\frac{\nabla J(\boldsymbol{\psi}_2)}{J(\boldsymbol{\psi}_2)}$ minimizes the righthand side. Therefore we have

$$\begin{aligned}
\log J(\boldsymbol{\psi}_1) &\geq \log J(\boldsymbol{\psi}_2) + (\boldsymbol{\psi}_1 - \boldsymbol{\psi}_2)^\top \frac{\nabla J(\boldsymbol{\psi}_2)}{J(\boldsymbol{\psi}_2)} + (A_1/2)\|\boldsymbol{\psi}_1 - \boldsymbol{\psi}_2\|_2^2 \\
&\geq \log J(\boldsymbol{\psi}_2) + (\tilde{\boldsymbol{\psi}}_1 - \boldsymbol{\psi}_2)^\top \frac{\nabla J(\boldsymbol{\psi}_2)}{J(\boldsymbol{\psi}_2)} + (A_1/2)\|\tilde{\boldsymbol{\psi}}_1 - \boldsymbol{\psi}_2\|_2^2 \\
&= \log J(\boldsymbol{\psi}_2) - \frac{1}{2A_1}\left\|\frac{\nabla J(\boldsymbol{\psi}_2)}{J(\boldsymbol{\psi}_2)}\right\|_2^2
\end{aligned}$$

Since this holds for any $\boldsymbol{\psi}_1 \in \mathbb{R}^n$, we have

$$\log J(\boldsymbol{\psi}_X(\omega^2, \nu)) \geq \log J(\boldsymbol{\psi}_2) - \frac{1}{2A_1}\left\|\frac{\nabla J(\boldsymbol{\psi}_2)}{J(\boldsymbol{\psi}_2)}\right\|_2^2$$

Similarly, from inequality (2.29), for any $\boldsymbol{\psi}_1 \in \mathbb{R}^n$ we have

$$\log J(\boldsymbol{\psi}_1) \leq \log J(\boldsymbol{\psi}_2) + (\boldsymbol{\psi}_1 - \boldsymbol{\psi}_2)^\top \frac{\nabla J(\boldsymbol{\psi}_2)}{J(\boldsymbol{\psi}_2)} + (A_2/2)\|\boldsymbol{\psi}_1 - \boldsymbol{\psi}_2\|_2^2,$$

minimizing each side over $\boldsymbol{\psi}_1$ will give us

$$\log J(\boldsymbol{\psi}_X(\omega^2, \nu)) \leq \log J(\boldsymbol{\psi}_2) - \frac{1}{2A_2}\left\|\frac{\nabla J(\boldsymbol{\psi}_2)}{J(\boldsymbol{\psi}_2)}\right\|_2^2.$$

■

A.6 Proof of Proposition 3

As in the proof of Theorem 1, $\boldsymbol{\psi}_X(\omega^2, \nu) = \mathfrak{f}_{\boldsymbol{\lambda}}$ with $\boldsymbol{\lambda} \in \Lambda^M$ defined as

$$\lambda_j \propto \pi_j \exp\left(-\frac{1}{2\omega^2}\|\boldsymbol{f}_j - \boldsymbol{Y}\|_2^2 + \frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{f}_j\|_2^2\right).$$

For any $j = 1, \ldots, M$,

$$\begin{aligned}
&\log J(\boldsymbol{\psi}^{(k)}) = \log J\left(\boldsymbol{\psi}^{(k-1)} + \alpha_k(\boldsymbol{f}_{J^{(k)}} - \boldsymbol{\psi}^{(k-1)})\right) \\
\leq\; &\log J\left(\boldsymbol{\psi}^{(k-1)} + \alpha_k(\boldsymbol{f}_j - \boldsymbol{\psi}^{(k-1)})\right) \\
\leq\; &\log J(\boldsymbol{\psi}^{(k-1)}) + \alpha_k(\boldsymbol{f}_j - \boldsymbol{\psi}^{(k-1)})^\top \frac{\nabla J(\boldsymbol{\psi}^{(k-1)})}{J(\boldsymbol{\psi}^{(k-1)})} + 2\alpha_k^2 D
\end{aligned}$$

where the first inequality comes from definition, the second inequality is from Taylor expansion at $\boldsymbol{\psi}^{(k-1)}$ and (2.29) in Lemma 1 with the fact that $\|\boldsymbol{f}_j - \boldsymbol{\psi}^{(k-1)}\|_2^2 \leq 4L^2$.

Then we sum the above inequality over $\boldsymbol{\lambda}$ which results that

$$\begin{aligned}
\log J(\boldsymbol{\psi}^{(k)}) &\leq \log J(\boldsymbol{\psi}^{(k-1)}) + \alpha_k \sum_{j=1}^{M} \lambda_j (\boldsymbol{f}_j - \boldsymbol{\psi}^{(k-1)})^\top \frac{\nabla J(\boldsymbol{\psi}^{(k-1)})}{J(\boldsymbol{\psi}^{(k-1)})} + 2\alpha_k^2 D \\
&= \log J(\boldsymbol{\psi}^{(k-1)}) + \alpha_k (\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\psi}^{(k-1)})^\top \frac{\nabla J(\boldsymbol{\psi}^{(k-1)})}{J(\boldsymbol{\psi}^{(k-1)})} + 2\alpha_k^2 D \\
&\leq \log J(\boldsymbol{\psi}^{(k-1)}) + \alpha_k (\log J(\boldsymbol{\psi}_X(\omega^2, \nu)) - \log J(\boldsymbol{\psi}^{(k-1)})) + 2\alpha_k^2 D
\end{aligned}$$

where the last inequality comes from the convexity of $\log J(\boldsymbol{\psi})$.

Denote $\delta_k = \log J(\boldsymbol{\psi}^{(k)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))$, it follows that

$$\delta_k \leq (1 - \alpha_k)\delta_{k-1} + 2\alpha_k^2 D$$

Since

$$\begin{aligned}
\delta_0 &= \log J(\boldsymbol{\psi}^{(0)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu)) \leq \log J(\boldsymbol{\psi}^{(0)}) \\
&\leq \log \left(\sum_{j=1}^{M} \pi_j \exp \left(\frac{1-\nu}{2\omega^2} \|\boldsymbol{\psi}^{(0)} - \boldsymbol{f}_j\|_2^2 \right) \right) \\
&\leq \frac{1-\nu}{\omega^2} 2L^2 \leq 2D
\end{aligned}$$

By mathematical induction if $\delta_{k-1} \leq \frac{8D}{k+2}$ then

$$\begin{aligned}
\delta_k &\leq (1 - \alpha_k)\delta_{k-1} + 2\alpha_k^2 D \\
&\leq (1 - 2/(k+1)) \frac{8D}{k+2} + 2(2/(k+1))^2 D \leq \frac{8D}{k+3}
\end{aligned}$$

Therefore

$$\log J(\boldsymbol{\psi}^{(k)}) \leq \log J(\boldsymbol{\psi}_X(\omega^2, \nu)) + \frac{8D}{k+3}$$

$\blacksquare$

A.7 Proof of Proposition 4

$$\begin{aligned}
\|\boldsymbol{\psi}^{(k)} - \boldsymbol{\psi}_X(\omega^2, \nu)\|_2^2 &\leq \frac{2}{A_1}\Big(\log J(\boldsymbol{\psi}^{(k)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))\Big) \\
&\leq \frac{2}{A_1}\frac{8D}{k+3} = \frac{16D}{A_1(k+3)}
\end{aligned}$$

where the first inequality comes from Taylor expansion at point $\boldsymbol{\psi}_X(\omega^2, \nu)$, with using (2.28) in Lemma 1 and $\nabla J(\boldsymbol{\psi}_2)$; and the second inequality is from Proposition 3.

It follows that

$$\begin{aligned}
\|\boldsymbol{\psi}^{(k)} - \boldsymbol{\eta}\|_2^2 &= \|(\boldsymbol{\psi}^{(k)} - \boldsymbol{\psi}_X(\omega^2, \nu)) + (\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta})\|_2^2 \\
&\leq \|\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta}\|_2^2 \\
&\quad +2\|\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta}\|_2\|\boldsymbol{\psi}^{(k)} - \boldsymbol{\psi}_X(\omega^2, \nu)\|_2 + \|\boldsymbol{\psi}^{(k)} - \boldsymbol{\psi}_X(\omega^2, \nu)\|_2^2 \\
&\leq \|\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta}\|_2^2 + 2\sqrt{\frac{16D}{A_1(k+3)}}\|\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta}\|_2 + \frac{16D}{A_1(k+3)}
\end{aligned}$$

Then the proposition follows using Theorem 1. ∎

A.8 Proof of Proposition 5

$$\boldsymbol{\psi}^{(k)} = \boldsymbol{\psi}^{(k-1)} - t_k \nabla \log J(\boldsymbol{\psi}^{(k-1)})$$

$$\begin{aligned}
\log J(\boldsymbol{\psi}^{(k)}) &= \log J(\boldsymbol{\psi}^{(k-1)} - t_k \nabla \log J(\boldsymbol{\psi}^{(k-1)})) \\
&\leq \log J(\boldsymbol{\psi}^{(k-1)}) - t_k \|\nabla \log J(\boldsymbol{\psi}^{(k-1)}))\|_2^2 + (A_2/2)t_k^2 \|\nabla \log J(\boldsymbol{\psi}^{(k-1)}))\|_2^2 \\
&= \log J(\boldsymbol{\psi}^{(k-1)}) - (t_k - (A_2/2)t_k^2)\|\nabla \log J(\boldsymbol{\psi}^{(k-1)}))\|_2^2
\end{aligned}$$

where the inequality is from (2.29).

Then by subtracting $\log J(\boldsymbol{\psi}_X(\omega^2,\nu))$ by each side, we have

$$\begin{aligned}&\log J(\boldsymbol{\psi}^{(k)}) - \log J(\boldsymbol{\psi}_X(\omega^2,\nu))\\ &\leq \log J(\boldsymbol{\psi}^{(k-1)}) - \log J(\boldsymbol{\psi}_X(\omega^2,\nu)) - (t_k - (A_2/2)t_k^2)\|\nabla \log J(\boldsymbol{\psi}^{(k-1)}))\|_2^2 \end{aligned}\tag{A.4}$$

Also from (2.30) we have

$$\|\nabla \log J(\boldsymbol{\psi}^{(k-1)}))\|_2^2 \geq 2A_1 \left(\log J(\boldsymbol{\psi}^{(k-1)}) - \log J(\boldsymbol{\psi}_X(\omega^2,\nu))\right) \tag{A.5}$$

Choose fixed step size $t_k = s \in (0, 2/A_2)$ for any $k > 0$, combining (A.4) and (A.5) results

$$\begin{aligned}&\log J(\boldsymbol{\psi}^{(k)}) - \log J(\boldsymbol{\psi}_X(\omega^2,\nu))\\ \leq\ & [1-2A_1(s-(A_2/2)s^2)]\left(\log J(\boldsymbol{\psi}^{(k-1)}) - \log J(\boldsymbol{\psi}_X(\omega^2,\nu))\right)\end{aligned}$$

It follows that

$$\begin{aligned}&\log J(\boldsymbol{\psi}^{(k)}) - \log J(\boldsymbol{\psi}_X(\omega^2,\nu))\\ \leq\ & [1-2A_1(s-(A_2/2)s^2)]^k\left(\log J(\boldsymbol{\psi}^{(0)}) - \log J(\boldsymbol{\psi}_X(\omega^2,\nu))\right)\end{aligned}$$

■

A.9 Proof of Proposition 6

$$\begin{aligned}\|\boldsymbol{\psi}^{(k)} - \boldsymbol{\psi}_X(\omega^2,\nu)\|_2^2 &\leq \frac{2}{A_1}\left(\log J(\boldsymbol{\psi}^{(k)}) - \log J(\boldsymbol{\psi}_X(\omega^2,\nu))\right)\\ &\leq \frac{2}{A_1}(1-A_1/A_2)^k \log J(\boldsymbol{\psi}^{(0)})\\ &\leq \frac{2}{A_1}(1-A_1/A_2)^k \frac{1-\nu}{2\omega^2}L^2 = L^2(1-A_1/A_2)^k\end{aligned}$$

where the first inequality comes from Taylor expansion at point $\boldsymbol{\psi}_X(\omega^2,\nu)$, with using (2.28) in Lemma 1 and $\nabla \log J(\boldsymbol{\psi}_X(\omega^2,\nu)) = 0$; the second inequality is from Proposition 5; and the third inequality is from assumption (2.24) resulting $\log J(\boldsymbol{\psi}^{(0)}) \leq \frac{1-\nu}{2\omega^2}L^2$.

It follows that

$$\begin{aligned}
\|\boldsymbol{\psi}^{(k)} - \boldsymbol{\eta}\|_2^2 &= \|(\boldsymbol{\psi}^{(k)} - \boldsymbol{\psi}_X(\omega^2,\nu)) + (\boldsymbol{\psi}_X(\omega^2,\nu) - \boldsymbol{\eta})\|_2^2 \\
&\leq \|\boldsymbol{\psi}_X(\omega^2,\nu) - \boldsymbol{\eta}\|_2^2 \\
&\quad +2\|\boldsymbol{\psi}_X(\omega^2,\nu) - \boldsymbol{\eta}\|_2\|\boldsymbol{\psi}^{(k)} - \boldsymbol{\psi}_X(\omega^2,\nu)\|_2 + \|\boldsymbol{\psi}^{(k)} - \boldsymbol{\psi}_X(\omega^2,\nu)\|_2^2 \\
&\leq \|\boldsymbol{\psi}_X(\omega^2,\nu) - \boldsymbol{\eta}\|_2^2 \\
&\quad +2\sqrt{L^2(1-A_1/A_2)^k}\|\boldsymbol{\psi}_X(\omega^2,\nu) - \boldsymbol{\eta}\|_2 + L^2(1-A_1/A_2)^k.
\end{aligned}$$

Then the proposition follows using Theorem 1.

∎

A.10 Proof of Proposition 7

$\boldsymbol{Y}$ is given, the following expectation is respect to the randomness from the MH algorithm. For $k > 0$, $\boldsymbol{u}_T^{(k-1)}$ from Algorithm 3 is estimator of $\mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}} = \sum_{j=1}^M \lambda_j^{(k-1)} \boldsymbol{f}_j$. Then in Algorithm 2 we update $\boldsymbol{\psi}^{(k)}$ by

$$\boldsymbol{\psi}^{(k)} = \boldsymbol{\psi}^{(k-1)} - t_k \frac{1-\nu}{\omega^2}(\boldsymbol{\psi}^{(k-1)} - \boldsymbol{u}_T^{(k-1)})$$

Denote $v^{(k-1)} = \frac{1-\nu}{\omega^2}(\boldsymbol{\psi}^{(k-1)} - \boldsymbol{u}_T^{(k-1)})$, then we have

$$\mathbb{E}[v^{(k-1)}|\boldsymbol{\psi}^{(k-1)}] = \frac{1-\nu}{\omega^2}(\boldsymbol{\psi}^{(k-1)} - \mathfrak{f}_{\boldsymbol{\lambda}^{(k-1)}}) = \nabla \log J(\boldsymbol{\psi}^{(k-1)})$$

and

$$\|\mathrm{COV}[v^{(k-1)}|\boldsymbol{\psi}^{(k-1)}]\|_{op} = \left(\frac{1-\nu}{\omega^2}\right)^2 \|\mathrm{COV}[\boldsymbol{u}_T^{(k-1)}|\boldsymbol{\psi}^{(k-1)}]\|_{op} \leq \left(\frac{1-\nu}{\omega^2}\right)^2 s^2$$

It follows that

$$\begin{aligned}
\log J(\boldsymbol{\psi}^{(k)}) &= \log J(\boldsymbol{\psi}^{(k-1)} - t_k v^{(k-1)}) \\
&\leq \log J(\boldsymbol{\psi}^{(k-1)}) - t_k \nabla \log J(\boldsymbol{\psi}^{(k-1)})^\top v^{(k-1)} + (A_2/2)t_k^2\|v^{(k-1)}\|_2^2
\end{aligned}$$

where the inequality is from (2.29).

Then by subtracting $\log J(\boldsymbol{\psi}_X(\omega^2,\nu))$ by each side and take expectation condition on $\boldsymbol{\psi}^{(k-1)}$, also denote $\delta_k = \log J(\boldsymbol{\psi}^{(k)}) - \log J(\boldsymbol{\psi}_X(\omega^2,\nu))$, we have

$$\begin{aligned}
\mathbb{E}[\delta_k|\boldsymbol{\psi}^{(k-1)}] &\le \delta_{k-1} - t_k \nabla \log J(\boldsymbol{\psi}^{(k-1)})^\top \mathbb{E}[v^{(k-1)}|\boldsymbol{\psi}^{(k-1)}] \\
&\quad +(A_2/2)t_k^2 \mathbb{E}[\|v^{(k-1)}\|_2^2|\boldsymbol{\psi}^{(k-1)}] \\
&\le \delta_{k-1} - t_k\|\nabla \log J(\boldsymbol{\psi}^{(k-1)})\|_2^2 \\
&\quad +(A_2/2)t_k^2\left(\|\nabla \log J(\boldsymbol{\psi}^{(k-1)})\|_2^2 + n\left(\frac{1-\nu}{\omega^2}\right)^2 s^2\right) \\
&= \delta_{k-1} - \frac{1}{2A_2}\|\nabla \log J(\boldsymbol{\psi}^{(k-1)})\|_2^2 + \frac{1}{2A_2}\left(\frac{1-\nu}{\omega^2}\right)^2 ns^2
\end{aligned}$$

Combine the above inequality with

$$\|\nabla \log J(\boldsymbol{\psi}^{(k-1)}))\|_2^2 \ge 2A_1\left(\log J(\boldsymbol{\psi}^{(k-1)}) - \log J(\boldsymbol{\psi}_X(\omega^2,\nu))\right) \tag{A.6}$$

which is from (2.30).

It results that

$$\mathbb{E}[\delta_k|\boldsymbol{\psi}^{(k-1)}] \le \delta_{k-1}(1 - A_1/A_2) + \frac{A_1^2}{2A_2}ns^2$$

And it directly follows that

$$\mathbb{E}[\delta_k] \le \mathbb{E}[\delta_{k-1}](1 - A_1/A_2) + \frac{A_1^2}{2A_2}ns^2$$

Therefore

$$\mathbb{E}[\delta_k] \le \mathbb{E}[\delta_0](1 - A_1/A_2)^k + \frac{A_1}{2}ns^2$$

∎

A.11 Proof of Lemma 3

Assume $(\boldsymbol{\lambda}^0, \boldsymbol{h}^0) \in A \cap B$. We have

$$Q(\boldsymbol{\lambda}^0) \ge \min_{\boldsymbol{\lambda}\in\Lambda^M} Q(\boldsymbol{\lambda}) = \min_{\boldsymbol{\lambda}\in\Lambda^M}\max_{\boldsymbol{h}\in\mathbb{R}^n} S(\boldsymbol{\lambda},\boldsymbol{h}) \ge \max_{\boldsymbol{h}\in\mathbb{R}^n}\min_{\boldsymbol{\lambda}\in\Lambda^M} S(\boldsymbol{\lambda},\boldsymbol{h})$$

The second inequality is from simple algebra and the third inequality is from Lemma 36.1 in Rockafellar (1997).

Also we have

$$\begin{aligned}
&\max_{\boldsymbol{h}\in\mathbb{R}^n}\min_{\boldsymbol{\lambda}\in\Lambda^M} S(\boldsymbol{\lambda},\boldsymbol{h})\\
&= \max_{\boldsymbol{h}\in\mathbb{R}^n}\left[\min_{\boldsymbol{\lambda}\in\Lambda^M}\left(\nu\sum_{j=1}^M \lambda_j\|\boldsymbol{f}_j-\boldsymbol{h}\|_2^2+2\omega^2\mathcal{K}(\boldsymbol{\lambda},\boldsymbol{\pi})\right)-\frac{\nu}{1-\nu}\|\boldsymbol{h}-\boldsymbol{Y}\|_2^2\right]\\
&= \max_{\boldsymbol{h}\in\mathbb{R}^n}\left[-\frac{\nu}{1-\nu}\|\boldsymbol{h}-\boldsymbol{Y}\|_2^2-2\omega^2\log\left(\sum_{j=1}^M \pi_j e^{-\nu\|\boldsymbol{f}_j-\boldsymbol{h}\|_2^2/2\omega^2}\right)\right]\\
&= \max_{\boldsymbol{h}\in\mathbb{R}^n} T(\boldsymbol{h}) = T(\hat{\boldsymbol{h}}) \geq T(\boldsymbol{h}^0)
\end{aligned}$$

The second equality comes from Jessen's inequality

$$\exp\left(\sum_{j=1}^M \lambda_j\left(-\frac{\nu}{2\omega^2}\|\boldsymbol{f}_j-\boldsymbol{h}\|_2^2-\log\frac{\lambda_j}{\pi_j}\right)\right)\leq\sum_{j=1}^M \lambda_j\exp\left(-\frac{\nu}{2\omega^2}\|\boldsymbol{f}_j-\boldsymbol{h}\|_2^2-\log\frac{\lambda_j}{\pi_j}\right).$$

Now we have

$$Q(\boldsymbol{\lambda}^0)\geq\min_{\boldsymbol{\lambda}\in\Lambda^M}Q(\boldsymbol{\lambda})=\min_{\boldsymbol{\lambda}\in\Lambda^M}\max_{\boldsymbol{h}\in\mathbb{R}^n}S(\boldsymbol{\lambda},\boldsymbol{h})\geq\max_{\boldsymbol{h}\in\mathbb{R}^n}\min_{\boldsymbol{\lambda}\in\Lambda^M}S(\boldsymbol{\lambda},\boldsymbol{h})=\max_{\boldsymbol{h}\in\mathbb{R}^n}T(\boldsymbol{h})\geq T(\boldsymbol{h}^0)$$

Our target is now to prove $Q(\boldsymbol{\lambda}^0)=T(\boldsymbol{h}^0)$. Since $(\boldsymbol{\lambda}^0,\boldsymbol{h}^0)\in A\cap B$ we have

$$\begin{cases}\boldsymbol{h}^0=\dfrac{1}{\nu}\boldsymbol{Y}-\dfrac{1-\nu}{\nu}\boldsymbol{f}_{\boldsymbol{\lambda}^0},\\[2ex] \lambda_j^0=\dfrac{\exp\left(-\frac{\nu}{2\omega^2}\|\boldsymbol{f}_j-\boldsymbol{h}^0\|_2^2\right)\pi_j}{\sum_{i=1}^M\exp\left(-\frac{\nu}{2\omega^2}\|\boldsymbol{f}_i-\boldsymbol{h}^0\|_2^2\right)\pi_i}\end{cases}$$

Then

$$\sum_{i=1}^M\exp\left(-\frac{\nu}{2\omega^2}\|\boldsymbol{f}_i-\boldsymbol{h}^0\|_2^2\right)\pi_i=\frac{\exp\left(-\frac{\nu}{2\omega^2}\|\boldsymbol{f}_j-\boldsymbol{h}^0\|_2^2\right)\pi_j}{\lambda_j^0}$$

which results that

$$\begin{aligned}
\log\left(\sum_{i=1}^M\exp\left(-\frac{\nu}{2\omega^2}\|\boldsymbol{f}_i-\boldsymbol{h}^0\|_2^2\right)\pi_i\right) &= -\frac{\nu}{2\omega^2}\|\boldsymbol{f}_j-\boldsymbol{h}^0\|_2^2-\log(\lambda_j^0/\pi_j)\\
&= \sum_{i=1}^M\lambda_i^0\left(-\frac{\nu}{2\omega^2}\|\boldsymbol{f}_i-\boldsymbol{h}^0\|_2^2-\log(\lambda_i^0/\pi_i)\right)
\end{aligned}$$

Plug back into $T(\boldsymbol{h}^0)$,

$$\begin{aligned}
T(\boldsymbol{h}^0) &= -\frac{\nu}{1-\nu}\|\boldsymbol{h}^0-\boldsymbol{Y}\|_2^2 - 2\omega^2\left[\sum_{i=1}^M \lambda_i^0\left(-\frac{\nu}{2\omega^2}\|\boldsymbol{f}_i-\boldsymbol{h}^0\|_2^2-\log(\lambda_i^0/\pi_i)\right)\right] \\
&= -\frac{\nu}{1-\nu}\|\boldsymbol{h}^0-\boldsymbol{Y}\|_2^2 + \nu\sum_{i=1}^M \lambda_i^0\|\boldsymbol{f}_i-\boldsymbol{h}^0\|_2^2 + 2\omega^2\mathcal{K}(\boldsymbol{\lambda}^0,\boldsymbol{\pi}) \\
&= \|\boldsymbol{f}_{\boldsymbol{\lambda}^0}-\boldsymbol{Y}\|_2^2 + \nu\sum_{i=1}^M \lambda_i^0\|\boldsymbol{f}_i-\boldsymbol{f}_{\boldsymbol{\lambda}^0}\|_2^2 + 2\omega^2\mathcal{K}(\boldsymbol{\lambda}^0,\boldsymbol{\pi}) \\
&= Q(\boldsymbol{\lambda}^0)
\end{aligned}$$

The third equality is obtained by plugging in $\boldsymbol{h}^0 = \frac{1}{\nu}\boldsymbol{Y} - \frac{1-\nu}{\nu}\boldsymbol{f}_{\boldsymbol{\lambda}^0}$.

Therefore

$$Q(\boldsymbol{\lambda}^0) = \min_{\boldsymbol{\lambda}\in\Lambda^M} Q(\boldsymbol{\lambda}) = \min_{\boldsymbol{\lambda}\in\Lambda^M}\max_{\boldsymbol{h}\in\mathbb{R}^n} S(\boldsymbol{\lambda},\boldsymbol{h}) = \max_{\boldsymbol{h}\in\mathbb{R}^n}\min_{\boldsymbol{\lambda}\in\Lambda^M} S(\boldsymbol{\lambda},\boldsymbol{h}) = \max_{\boldsymbol{h}\in\mathbb{R}^n} T(\boldsymbol{h}) = T(\boldsymbol{h}^0)$$

So $\hat{\boldsymbol{h}} = \boldsymbol{h}^0$ and $\boldsymbol{\lambda}^Q = \boldsymbol{\lambda}^0$, combining with $\boldsymbol{h}^0 = \frac{1}{\nu}\boldsymbol{Y} - \frac{1-\nu}{\nu}\boldsymbol{f}_{\boldsymbol{\lambda}^0}$, we have

$$\hat{\boldsymbol{h}} = \frac{1}{\nu}\boldsymbol{Y} - \frac{1-\nu}{\nu}\mathfrak{f}_{\boldsymbol{\lambda}^Q}.$$

Then $A\cap B$ has unique point $(\boldsymbol{\lambda}^Q,\hat{\boldsymbol{h}})$. ∎

A.12 Proof of (3.7)

$$\begin{aligned}
\mathbb{E}(\boldsymbol{\mu}|\boldsymbol{Y}) &= \sum_{\gamma\in\wp}\mathbb{E}(\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma|\boldsymbol{Y},\mathcal{M}_\gamma)p(\mathcal{M}_\gamma|\boldsymbol{Y}) \\
&= \sum_{\gamma\in\wp}\boldsymbol{X}_\gamma\mathbb{E}(\boldsymbol{\beta}_\gamma|\boldsymbol{Y},\mathcal{M}_\gamma)p(\boldsymbol{Y}|\mathcal{M}_\gamma)p(\mathcal{M}_\gamma)/p(\boldsymbol{Y}) \\
&= \frac{1}{p(\boldsymbol{Y})}\sum_{\gamma\in\wp}\pi_\gamma\boldsymbol{X}_\gamma\mathbb{E}(\boldsymbol{\beta}_\gamma|\boldsymbol{Y},\mathcal{M}_\gamma)p(\boldsymbol{Y}|\mathcal{M}_\gamma) \\
&= \frac{1}{p(\boldsymbol{Y})}\sum_{\gamma\in\wp}\pi_\gamma\boldsymbol{X}_\gamma\int\boldsymbol{\beta}_\gamma p(\boldsymbol{\beta}_\gamma|\boldsymbol{Y},\mathcal{M}_\gamma)p(\boldsymbol{Y}|\mathcal{M}_\gamma)\,d\boldsymbol{\beta}_\gamma \\
&= \frac{1}{p(\boldsymbol{Y})}\sum_{\gamma\in\wp}\pi_\gamma\boldsymbol{X}_\gamma\int\boldsymbol{\beta}_\gamma p(\boldsymbol{Y}|\boldsymbol{\beta}_\gamma,\mathcal{M}_\gamma)p(\boldsymbol{\beta}_\gamma|\mathcal{M}_\gamma)\,d\boldsymbol{\beta}_\gamma
\end{aligned}$$

$$\begin{aligned}
& \int \boldsymbol{\beta}_\gamma p(\boldsymbol{Y}|\boldsymbol{\beta}_\gamma, \mathcal{M}_\gamma) p(\boldsymbol{\beta}_\gamma|\mathcal{M}_\gamma)\, d\boldsymbol{\beta}_\gamma \\
= & \int \boldsymbol{\beta}_\gamma (2\pi\omega^2)^{-n/2} \exp\left(-\frac{\|\boldsymbol{Y} - \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma\|_2^2}{2\omega^2}\right) \\
& (2\pi\omega^2)^{-d_\gamma/2} |\boldsymbol{K}_\gamma/g|^{1/2} \exp\left(-\frac{(\boldsymbol{\beta}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)^\top \boldsymbol{K}_\gamma (\boldsymbol{\beta}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)}{2g\omega^2}\right)\, d\boldsymbol{\beta}_\gamma \\
= & (2\pi\omega^2)^{-n/2} (2\pi\omega^2)^{-d_\gamma/2} |\boldsymbol{K}_\gamma/g|^{1/2} \int \boldsymbol{\beta}_\gamma \exp(-G_\gamma/2\omega^2)\, d\boldsymbol{\beta}_\gamma
\end{aligned}$$

where

$$G_\gamma = \|\boldsymbol{Y} - \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma\|_2^2 + (\boldsymbol{\beta}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)^\top \boldsymbol{K}_\gamma (\boldsymbol{\beta}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)/g$$

Since $\hat{\boldsymbol{\beta}}_\gamma = (\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g)^{-1}(\boldsymbol{X}_\gamma^\top \boldsymbol{Y} + \frac{1}{g}\boldsymbol{K}_\gamma \tilde{\boldsymbol{\beta}}_\gamma)$, then

$$\begin{aligned}
G_\gamma &= (\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma)^\top (\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g)(\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma) \\
& \quad + \|\boldsymbol{Y}\|_2^2 + \tilde{\boldsymbol{\beta}}_\gamma^\top \boldsymbol{K}_\gamma \tilde{\boldsymbol{\beta}}_\gamma/g - \hat{\boldsymbol{\beta}}_\gamma^\top (\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g)\hat{\boldsymbol{\beta}}_\gamma \\
&= (\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma)^\top (\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g)(\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma) \\
& \quad + \|\boldsymbol{Y} - \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma\|_2^2 + (\hat{\boldsymbol{\beta}}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)^\top \boldsymbol{K}_\gamma (\hat{\boldsymbol{\beta}}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)/g
\end{aligned}$$

It follows that

$$\begin{aligned}
& \int \boldsymbol{\beta}_\gamma p(\boldsymbol{Y}|\boldsymbol{\beta}_\gamma, \mathcal{M}_\gamma) p(\boldsymbol{\beta}_\gamma|\mathcal{M}_\gamma)\, d\boldsymbol{\beta}_\gamma \\
= & (2\pi\omega^2)^{-n/2} |\boldsymbol{K}_\gamma/g|^{1/2} |\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g|^{-1/2} \\
& \exp\left(-\frac{\|\boldsymbol{Y} - \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma\|_2^2 + (\hat{\boldsymbol{\beta}}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)^\top \boldsymbol{K}_\gamma (\hat{\boldsymbol{\beta}}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)/g}{2\omega^2}\right) \hat{\boldsymbol{\beta}}_\gamma
\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}(\boldsymbol{\mu}|\boldsymbol{Y}) &= \frac{(2\pi\omega^2)^{-n/2}}{p(\boldsymbol{Y})}\sum_{\gamma\in\wp}\pi_\gamma(\boldsymbol{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma)|\boldsymbol{K}_\gamma/g|^{1/2}|\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma+\boldsymbol{K}_\gamma/g|^{-1/2} \\ &\quad \exp\left(-\frac{\|\boldsymbol{Y}-\boldsymbol{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma\|_2^2+(\hat{\boldsymbol{\beta}}_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)^\top\boldsymbol{K}_\gamma(\hat{\boldsymbol{\beta}}_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)/g}{2\omega^2}\right)\end{aligned}$$

Also notice that

$$\begin{aligned}p(\boldsymbol{Y}) &= \sum_{\gamma\in\wp}p(\boldsymbol{Y}|\mathcal{M}_\gamma)p(\mathcal{M}_\gamma) \\ &= \sum_{\gamma\in\wp}p(\mathcal{M}_\gamma)\int_{\boldsymbol{\beta}_\gamma}p(\boldsymbol{Y}|\boldsymbol{\beta}_\gamma,\mathcal{M}_\gamma)p(\boldsymbol{\beta}_\gamma|\mathcal{M}_\gamma)\,d\boldsymbol{\beta}_\gamma \\ &= \sum_{\gamma\in\wp}\pi_\gamma(2\pi\omega^2)^{-n/2}|\boldsymbol{K}_\gamma/g|^{1/2}|\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma+\boldsymbol{K}_\gamma/g|^{-1/2} \\ &\quad \exp\left(-\frac{\|\boldsymbol{Y}-\boldsymbol{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma\|_2^2+(\hat{\boldsymbol{\beta}}_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)^\top\boldsymbol{K}_\gamma(\hat{\boldsymbol{\beta}}_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)/g}{2\omega^2}\right)\end{aligned}$$

Therefore,

$$\mathbb{E}(\boldsymbol{\mu}|\boldsymbol{Y}) = \sum_{\gamma\in\wp}\lambda_\gamma(\boldsymbol{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma)$$

where

$$\begin{aligned}\lambda_\gamma &\propto \pi_\gamma|\boldsymbol{K}_\gamma/g|^{1/2}|\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma+\boldsymbol{K}_\gamma/g|^{-1/2} \\ &\quad \exp\left(-\frac{\|\boldsymbol{Y}-\boldsymbol{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma\|_2^2+(\hat{\boldsymbol{\beta}}_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)^\top\boldsymbol{K}_\gamma(\hat{\boldsymbol{\beta}}_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)/g}{2\omega^2}\right)\end{aligned}$$

and $\sum_{\gamma\in\wp}\lambda_\gamma = 1$ and $\lambda_\gamma > 0$.

∎

A.13 Proof of Proposition 9

$$\begin{aligned}
J(\boldsymbol{\psi}) &= \sum_{\gamma\in\wp} \pi_\gamma \int \exp\left(\frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi}-\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma\|_2^2 - \frac{1}{2\omega^2}\|\boldsymbol{Y}-\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma\|_2^2\right) \\
&\quad (2\pi g\omega^2)^{-d_\gamma/2}|\boldsymbol{K}_\gamma|^{1/2}\exp\left(-\frac{(\boldsymbol{\beta}_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)^\top\boldsymbol{K}_\gamma(\boldsymbol{\beta}_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)}{2g\omega^2}\right)\, d\boldsymbol{\beta}_\gamma \\
&= \sum_{\gamma\in\wp} \pi_\gamma (2\pi g\omega^2)^{-d_\gamma/2}|\boldsymbol{K}_\gamma|^{1/2} \int \exp\left(-\frac{H_\gamma}{2\omega^2}\right)\, d\boldsymbol{\beta}_\gamma
\end{aligned}$$

where

$$\begin{aligned}
H_\gamma &= -(1-\nu)\|\boldsymbol{\psi}-\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma\|_2^2 + \|\boldsymbol{Y}-\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma\|_2^2 + (\boldsymbol{\beta}_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)^\top\boldsymbol{K}_\gamma(\boldsymbol{\beta}_\gamma-\tilde{\boldsymbol{\beta}}_\gamma)/g \\
&= \left(\boldsymbol{\beta}_\gamma + \left(\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1}\left((1-\nu)\boldsymbol{X}_\gamma^\top\boldsymbol{\psi} - \boldsymbol{X}_\gamma^\top\boldsymbol{Y} - \boldsymbol{K}_\gamma\tilde{\boldsymbol{\beta}}_\gamma/g\right)\right)^\top \\
&\quad \cdot \left(\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right) \\
&\quad \cdot \left(\boldsymbol{\beta}_\gamma + \left(\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1}\left((1-\nu)\boldsymbol{X}_\gamma^\top\boldsymbol{\psi} - \boldsymbol{X}_\gamma^\top\boldsymbol{Y} - \boldsymbol{K}_\gamma\tilde{\boldsymbol{\beta}}_\gamma/g\right)\right) \\
&\quad - \left((1-\nu)\boldsymbol{X}_\gamma^\top\boldsymbol{\psi} - \boldsymbol{X}_\gamma^\top\boldsymbol{Y} - \boldsymbol{K}_\gamma\tilde{\boldsymbol{\beta}}_\gamma/g\right)^\top \\
&\quad \cdot \left(\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1}\left((1-\nu)\boldsymbol{X}_\gamma^\top\boldsymbol{\psi} - \boldsymbol{X}_\gamma^\top\boldsymbol{Y} - \boldsymbol{K}_\gamma\tilde{\boldsymbol{\beta}}_\gamma/g\right)^\top \\
&\quad -(1-\nu)\|\boldsymbol{\psi}\|_2^2 + \|\boldsymbol{Y}\|_2^2 + \tilde{\boldsymbol{\beta}}_\gamma^\top\boldsymbol{K}_\gamma\tilde{\boldsymbol{\beta}}_\gamma/g
\end{aligned}$$

$$\begin{aligned}
J(\boldsymbol{\psi}) &= \sum_{\gamma\in\wp} \pi_\gamma (2\pi g\omega^2)^{-d_\gamma/2}|\boldsymbol{K}_\gamma|^{1/2}(2\pi\omega^2)^{d_\gamma/2}|\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g|^{-1/2}e^{\frac{G_\gamma}{2\omega^2}} \\
&= \sum_{\gamma\in\wp} \pi_\gamma |\boldsymbol{K}_\gamma/g|^{1/2}|\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g|^{-1/2}\exp\left(\frac{G_\gamma}{2\omega^2}\right)
\end{aligned}$$

where

$$\begin{aligned}
G_\gamma &= -\left((1-\nu)\boldsymbol{X}_\gamma^\top \boldsymbol{\psi} - \boldsymbol{X}_\gamma^\top \boldsymbol{Y} - \boldsymbol{K}_\gamma \tilde{\boldsymbol{\beta}}_\gamma/g\right)^\top \\
&\quad \cdot \left(\nu \boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1} \left((1-\nu)\boldsymbol{X}_\gamma^\top \boldsymbol{\psi} - \boldsymbol{X}_\gamma^\top \boldsymbol{Y} - \boldsymbol{K}_\gamma \tilde{\boldsymbol{\beta}}_\gamma/g\right)^\top \\
&\quad -(1-\nu)\|\boldsymbol{\psi}\|_2^2 + \|\boldsymbol{Y}\|_2^2 + \tilde{\boldsymbol{\beta}}_\gamma^\top \boldsymbol{K}_\gamma \tilde{\boldsymbol{\beta}}_\gamma/g \\
&= (1-\nu)\boldsymbol{\psi}^\top \left[\boldsymbol{I}_n + (1-\nu)\boldsymbol{X}_\gamma \left(\nu \boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1} \boldsymbol{X}_\gamma^\top\right] \boldsymbol{\psi} \\
&\quad -2(1-\nu)\boldsymbol{\psi}^\top \boldsymbol{X}_\gamma \left(\nu \boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1} (\boldsymbol{X}_\gamma^\top \boldsymbol{Y} + \boldsymbol{K}_\gamma \tilde{\boldsymbol{\beta}}_\gamma/g) \\
&\quad +(\boldsymbol{X}_\gamma^\top \boldsymbol{Y} + \boldsymbol{K}_\gamma \tilde{\boldsymbol{\beta}}_\gamma/g)^\top \left(\nu \boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1} (\boldsymbol{X}_\gamma^\top \boldsymbol{Y} + \boldsymbol{K}_\gamma \tilde{\boldsymbol{\beta}}_\gamma/g) \\
&\quad -\|\boldsymbol{Y}\|_2^2 - \tilde{\boldsymbol{\beta}}_\gamma^\top \boldsymbol{K}_\gamma \tilde{\boldsymbol{\beta}}_\gamma \\
&= (1-\nu)(\boldsymbol{\psi} - \mu_\gamma)^\top \left[\boldsymbol{I}_n + (1-\nu)\boldsymbol{X}_\gamma \left(\nu \boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1} \boldsymbol{X}_\gamma^\top\right] (\boldsymbol{\psi} - \mu_\gamma) \\
&\quad -(1-\nu)\mu_\gamma^\top \left[\boldsymbol{I}_n + (1-\nu)\boldsymbol{X}_\gamma \left(\nu \boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1} \boldsymbol{X}_\gamma^\top\right] \mu_\gamma \\
&\quad +(\boldsymbol{X}_\gamma^\top \boldsymbol{Y} + \boldsymbol{K}_\gamma \tilde{\boldsymbol{\beta}}_\gamma/g)^\top \left(\nu \boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1} (\boldsymbol{X}_\gamma^\top \boldsymbol{Y} + \boldsymbol{K}_\gamma \tilde{\boldsymbol{\beta}}_\gamma/g) \\
&\quad -\|\boldsymbol{Y}\|_2^2 - \tilde{\boldsymbol{\beta}}_\gamma^\top \boldsymbol{K}_\gamma \tilde{\boldsymbol{\beta}}_\gamma/g
\end{aligned}$$

where

$$\begin{aligned}
\mu_\gamma &= \left[\boldsymbol{I}_n + (1-\nu)\boldsymbol{X}_\gamma \left(\nu \boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1} \boldsymbol{X}_\gamma^\top\right]^{-1} \\
&\quad \cdot \boldsymbol{X}_\gamma \left(\nu \boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1} (\boldsymbol{X}_\gamma^\top \boldsymbol{Y} + \boldsymbol{K}_\gamma \tilde{\boldsymbol{\beta}}_\gamma/g)
\end{aligned}$$

From definition (3.9) we have

$$(\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g)\hat{\boldsymbol{\beta}}_\gamma = (\boldsymbol{X}_\gamma^\top \boldsymbol{Y} + \boldsymbol{K}_\gamma \tilde{\boldsymbol{\beta}}_\gamma/g)$$

Then

$$\begin{aligned}
\mu_\gamma &= \left[\boldsymbol{I}_n + (1-\nu)\boldsymbol{X}_\gamma\left(\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1}\boldsymbol{X}_\gamma^\top\right]^{-1}\boldsymbol{X}_\gamma\left(\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1} \\
&\quad (\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g)\hat{\boldsymbol{\beta}}_\gamma \\
&= \left[\boldsymbol{I}_n + (1-\nu)\boldsymbol{X}_\gamma\left(\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1}\boldsymbol{X}_\gamma^\top\right]^{-1}\boldsymbol{X}_\gamma\left(\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1} \\
&\quad (\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g + (1-\nu)\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma)\hat{\boldsymbol{\beta}}_\gamma \\
&= \left[\boldsymbol{I}_n + (1-\nu)\boldsymbol{X}_\gamma\left(\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1}\boldsymbol{X}_\gamma^\top\right]^{-1} \\
&\quad \boldsymbol{X}_\gamma\left(\boldsymbol{I}_{d_\gamma} + (1-\nu)(\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g)^{-1}\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma\right)\hat{\boldsymbol{\beta}}_\gamma \\
&= \left[\boldsymbol{I}_n + (1-\nu)\boldsymbol{X}_\gamma\left(\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1}\boldsymbol{X}_\gamma^\top\right]^{-1} \\
&\quad \left(\boldsymbol{I}_n + (1-\nu)\boldsymbol{X}_\gamma(\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g)^{-1}\boldsymbol{X}_\gamma^\top\right)\boldsymbol{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma \\
&= \boldsymbol{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma
\end{aligned}$$

Also notice that

$$\begin{aligned}
& -(1-\nu)\mu_\gamma^\top\left[\boldsymbol{I}_n + (1-\nu)\boldsymbol{X}_\gamma\left(\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1}\boldsymbol{X}_\gamma^\top\right]\mu_\gamma \\
& +(\boldsymbol{X}_\gamma^\top\boldsymbol{Y} + \boldsymbol{K}_\gamma\tilde{\boldsymbol{\beta}}_\gamma/g)^\top\left(\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1}(\boldsymbol{X}_\gamma^\top\boldsymbol{Y} + \boldsymbol{K}_\gamma\tilde{\boldsymbol{\beta}}_\gamma/g) \\
& -\|\boldsymbol{Y}\|_2^2 - \tilde{\boldsymbol{\beta}}_\gamma^\top\boldsymbol{K}_\gamma\tilde{\boldsymbol{\beta}}_\gamma/g \\
= \; & -(1-\nu)\hat{\boldsymbol{\beta}}_\gamma^\top\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma\left(\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1}(\boldsymbol{X}_\gamma^\top\boldsymbol{Y} + \boldsymbol{K}_\gamma\tilde{\boldsymbol{\beta}}_\gamma/g) \\
& +(\boldsymbol{X}_\gamma^\top\boldsymbol{Y} + \boldsymbol{K}_\gamma\tilde{\boldsymbol{\beta}}_\gamma/g)^\top\left(\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1}(\boldsymbol{X}_\gamma^\top\boldsymbol{Y} + \boldsymbol{K}_\gamma\tilde{\boldsymbol{\beta}}_\gamma/g) \\
& -\|\boldsymbol{Y}\|_2^2 - \tilde{\boldsymbol{\beta}}_\gamma^\top\boldsymbol{K}_\gamma\tilde{\boldsymbol{\beta}}_\gamma/g \\
= \; & -(1-\nu)\hat{\boldsymbol{\beta}}_\gamma^\top\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma\left(\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1}(\boldsymbol{X}_\gamma^\top\boldsymbol{Y} + \boldsymbol{K}_\gamma\tilde{\boldsymbol{\beta}}_\gamma/g) \\
& +\hat{\boldsymbol{\beta}}_\gamma^\top(\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g)\left(\nu\boldsymbol{X}_\gamma^\top\boldsymbol{X}_\gamma + \boldsymbol{K}_\gamma/g\right)^{-1}(\boldsymbol{X}_\gamma^\top\boldsymbol{Y} + \boldsymbol{K}_\gamma\tilde{\boldsymbol{\beta}}_\gamma/g) \\
& -\|\boldsymbol{Y}\|_2^2 - \tilde{\boldsymbol{\beta}}_\gamma^\top\boldsymbol{K}_\gamma\tilde{\boldsymbol{\beta}}_\gamma/g \\
= \; & \hat{\boldsymbol{\beta}}_\gamma^\top(\boldsymbol{X}_\gamma^\top\boldsymbol{Y} + \boldsymbol{K}_\gamma\tilde{\boldsymbol{\beta}}_\gamma/g) - \|\boldsymbol{Y}\|_2^2 - \tilde{\boldsymbol{\beta}}_\gamma^\top\boldsymbol{K}_\gamma\tilde{\boldsymbol{\beta}}_\gamma/g \\
= \; & -\|\boldsymbol{Y} - \boldsymbol{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma\|_2^2 - (\hat{\boldsymbol{\beta}}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)^\top\boldsymbol{K}_\gamma(\hat{\boldsymbol{\beta}}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)/g
\end{aligned}$$

Therefore,

$$
\begin{aligned}
J(\boldsymbol{\psi}) = \sum_{\gamma \in \wp} & \pi_\gamma |\frac{1}{g}\boldsymbol{K}_\gamma|^{1/2} |\nu \boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \frac{1}{g}\boldsymbol{K}_\gamma|^{-1/2} \\
& \cdot \exp\left(\frac{1-\nu}{2\omega^2} (\boldsymbol{\psi} - \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma)^\top \boldsymbol{W}_\gamma (\boldsymbol{\psi} - \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma) \right) \\
& \cdot \exp\left(-\frac{1}{2\omega^2} \|\boldsymbol{Y} - \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma\|_2^2 - \frac{(\hat{\boldsymbol{\beta}}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)^\top \boldsymbol{K}_\gamma \, , (\hat{\boldsymbol{\beta}}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)}{2g\omega^2} \right)
\end{aligned}
$$

where

$$
\boldsymbol{W}_\gamma = \boldsymbol{I}_n + (1-\nu)\boldsymbol{X}_\gamma \left(\nu \boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \frac{1}{g}\boldsymbol{K}_\gamma \right)^{-1} \boldsymbol{X}_\gamma^\top \, .
$$

$\blacksquare$

A.14 Proof of Theorem 5

Proposition 13. *For any given positive definite matrix $A \in \mathbb{R}^{d \times d}$, we have the following inequality*

$$
tr(AB^{-1}) + \log(|B|) \geq d + \log(|A|),
$$

for any positive definite matrix $B \in \mathbb{R}^{d \times d}$, and the equation is held when $B = A$.

Proof.

$$
\mathrm{tr}(AB^{-1}) + \log(|B|) = \mathrm{tr}(B^{-1/2}AB^{-1/2}) + \log(|B|) = \mathrm{tr}(C) - \log(|C|) + \log(|A|)
$$

where $C = B^{-1/2}AB^{-1/2}$ is also positive definite. And by singular value decomposition $C = O^\top U O$ where $O \in \mathbb{R}^{d \times d}$ is orthogonal matrix and $U = \mathrm{diag}(u_1, \ldots, u_d) > 0$. Then

$$
\mathrm{tr}(C) - \log(|C|) = \mathrm{tr}(U) - \log(|U|) = \sum_{i=1}^{d} (u_i - \log(u_i)) \geq d
$$

and last equation holds when $u_i = 1 \; \forall i$, which is equivalent to $A = B$. $\square$

For some fixed $l \in \wp$, specify $\boldsymbol{f}_\gamma$ with distribution $\Theta(\gamma)$ in Corollary 2 as following:

$$\boldsymbol{f}_\gamma | \boldsymbol{\beta}_l, \mathcal{M}_l = \boldsymbol{X}_l \boldsymbol{\beta}_l,$$

and $\boldsymbol{\beta}_l | \mathcal{M}_l \sim N(\boldsymbol{\beta}_l^*, \omega^2 \boldsymbol{D}_l^{-1})$ and $p(\mathcal{M}_l) = 1$.

While $\pi(\gamma)$ is defined as,

$$\boldsymbol{f}_\gamma | \boldsymbol{\beta}_\gamma, \mathcal{M}_\gamma = \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma,$$

and $p(\boldsymbol{\beta}_\gamma | \mathcal{M}_\gamma)$ is as defined in (3.4) and $p(\mathcal{M}_\gamma) = \pi_\gamma$.

Then we apply Corollary 2, write the right hand side of oracle deviation inequality (2.22) as

$$RHS = \nu \int_\Omega \|\boldsymbol{f}_\gamma - \boldsymbol{\eta}\|_2^2 \Theta(\gamma)\, d\gamma + (1-\nu) \left\| \int_\Omega \boldsymbol{f}_\gamma \Theta(\gamma)\, d\gamma - \boldsymbol{\eta} \right\|_2^2 + 2\omega^2 \mathcal{K}(\Theta, \pi)$$

The notation of $\mathbb{E}(\cdot)$ below is for expectation with respect to $\Theta(\gamma)$, specifically $\boldsymbol{f}_\gamma = \boldsymbol{X}_l \boldsymbol{\beta}_l$ with

$$\boldsymbol{\beta}_l \sim N(\boldsymbol{\beta}_l^*, \omega^2 \boldsymbol{D}_l^{-1}),$$

then we have

$$\begin{aligned}
\int_\Omega \|\boldsymbol{f}_\gamma - \boldsymbol{\eta}\|_2^2 \Theta(\gamma)\, d\gamma &= \mathbb{E}\|\boldsymbol{X}_l \boldsymbol{\beta}_l - \boldsymbol{\eta}\|_2^2 \\
&= \mathrm{tr}(\mathrm{COV}(\boldsymbol{X}_l \boldsymbol{\beta}_l)) + \|\boldsymbol{X}_l \boldsymbol{\beta}_l^* - \boldsymbol{\eta}\|_2^2 \\
&= \omega^2 \mathrm{tr}(\boldsymbol{X}_l^\top \boldsymbol{X}_l \boldsymbol{D}_l^{-1}) + \|\boldsymbol{X}_l \boldsymbol{\beta}_l^* - \boldsymbol{\eta}\|_2^2
\end{aligned}$$

and

$$\left\| \int_\Omega \boldsymbol{f}_\gamma \Theta(\gamma)\, d\gamma - \boldsymbol{\eta} \right\|_2^2 = \|\boldsymbol{X}_l \boldsymbol{\beta}_l^* - \boldsymbol{\eta}\|_2^2$$

Also

$$\begin{aligned}
\mathcal{K}(\Theta,\pi) &= \mathbb{E}\left[\log\frac{(2\pi\omega^2)^{-d_l/2}|\boldsymbol{D}_l|^{1/2}\exp\left(-\frac{(\boldsymbol{\beta}_l-\boldsymbol{\beta}_l^*)^\top\boldsymbol{D}_l(\boldsymbol{\beta}_l-\boldsymbol{\beta}_l^*)}{2\omega^2}\right)}{\pi_l(2\pi\omega^2)^{-d_l/2}|\boldsymbol{K}_l/g|^{1/2}\exp\left(-\frac{(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)^\top\boldsymbol{K}_l(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)}{2g\omega^2}\right)}\right] \\
&= \log(1/\pi_l)+\frac{1}{2}\log(|\boldsymbol{D}_l|/|\boldsymbol{K}_l/g|) \\
&\quad +\mathbb{E}\left(-\frac{(\boldsymbol{\beta}_l-\boldsymbol{\beta}_l^*)^\top\boldsymbol{D}_l(\boldsymbol{\beta}_l-\boldsymbol{\beta}_l^*)}{2\omega^2}+\frac{(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)^\top\boldsymbol{K}_l(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)}{2g\omega^2}\right)
\end{aligned}$$

Notice that

$$\begin{aligned}
\mathbb{E}(\boldsymbol{\beta}_l-\boldsymbol{\beta}_l^*)^\top\boldsymbol{D}_l(\boldsymbol{\beta}_l-\boldsymbol{\beta}_l^*) &= \mathbb{E}\mathrm{tr}\left[\left(\boldsymbol{D}_l^{1/2}(\boldsymbol{\beta}_l-\boldsymbol{\beta}_l^*)\right)\left(\boldsymbol{D}_l^{1/2}(\boldsymbol{\beta}_l-\boldsymbol{\beta}_l^*)\right)^\top\right] \\
&= \mathrm{tr}\left[\mathrm{COV}\left(\boldsymbol{D}_l^{1/2}(\boldsymbol{\beta}_l-\boldsymbol{\beta}_l^*)\right)\right]=d_l\omega^2
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)^\top\boldsymbol{K}_l(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l) &= \mathbb{E}\mathrm{tr}\left[\left(\boldsymbol{K}_l^{1/2}(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)\right)\left(\boldsymbol{K}_l^{1/2}(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)\right)^\top\right] \\
&= \omega^2\mathrm{tr}(\boldsymbol{K}_l\boldsymbol{D}_l^{-1})+(\boldsymbol{\beta}_l^*-\tilde{\boldsymbol{\beta}}_l)^\top\boldsymbol{K}_l(\boldsymbol{\beta}_l^*-\tilde{\boldsymbol{\beta}}_l)
\end{aligned}$$

Therefore we have

$$\begin{aligned}
RHS &= \|\boldsymbol{X}_l\boldsymbol{\beta}_l^*-\boldsymbol{\eta}\|_2^2+2\omega^2\left(\log(1/\pi_l)+\frac{1}{2}\log(|\boldsymbol{D}_l|/|\boldsymbol{K}_l/g|)\right) \\
&\quad +\omega^2\nu\mathrm{tr}(\boldsymbol{X}_l^\top\boldsymbol{X}_l\boldsymbol{D}_l^{-1})-d_l\omega^2+\omega^2\mathrm{tr}(\boldsymbol{K}_l\boldsymbol{D}_l^{-1}/g) \\
&\quad +(\boldsymbol{\beta}_l^*-\tilde{\boldsymbol{\beta}}_l)^\top\boldsymbol{K}_l(\boldsymbol{\beta}_l^*-\tilde{\boldsymbol{\beta}}_l)/g \\
&= 2\omega^2\log(1/\pi_l)-\omega^2\log(|\boldsymbol{K}_l/g|)-d_l\omega^2 \\
&\quad +\|\boldsymbol{X}_l\boldsymbol{\beta}_l^*-\boldsymbol{\eta}\|_2^2+(\boldsymbol{\beta}_l^*-\tilde{\boldsymbol{\beta}}_l)^\top\boldsymbol{K}_l(\boldsymbol{\beta}_l^*-\tilde{\boldsymbol{\beta}}_l)/g \\
&\quad +\omega^2\mathrm{tr}\left((\nu\boldsymbol{X}_l^\top\boldsymbol{X}_l+\boldsymbol{K}_l/g)\boldsymbol{D}_l^{-1}\right)+\omega^2\log(|\boldsymbol{D}_l|)
\end{aligned}$$

where $\boldsymbol{\beta}_l^*$ and $\boldsymbol{D}_l$ are to be decided to minimize RHS.

Now we minimize RHS over $\boldsymbol{D}_l\in\mathbb{R}^{d_l\times d_l}$ being positive definite. By using Proposition 13, when

$$\boldsymbol{D}_l=\nu\boldsymbol{X}_l^\top\boldsymbol{X}_l+\boldsymbol{K}_l/g$$

the minimum is obtained, and RHS becomes

$$\begin{aligned} RHS &= 2\omega^2 \log(1/\pi_l) + \omega^2 \log(|\nu g \boldsymbol{X}_l^\top \boldsymbol{X}_l \boldsymbol{K}_l^{-1} + \boldsymbol{I}_{d_l}|) \\ &\quad + \|\boldsymbol{X}_l \boldsymbol{\beta}_l^* - \boldsymbol{\eta}\|_2^2 + (\boldsymbol{\beta}_l^* - \tilde{\boldsymbol{\beta}}_l)^\top \boldsymbol{K}_l (\boldsymbol{\beta}_l^* - \tilde{\boldsymbol{\beta}}_l)/g \end{aligned}$$

∎

A.15 Proof of Corollary 4

The oracle inequalities can be obtained directly by restricting $\pi_k = 1$ in Theorem 5. Here we just show the explicit expression for $\boldsymbol{\psi}_X(\omega^2, \nu)$. With Proposition 9, $\boldsymbol{\psi}_X(\omega^2, \nu)$ is the minimizer for $J(\boldsymbol{\psi})$ as in (3.12) with $\pi_k = 1$. So it is obvious that

$$\boldsymbol{\psi}_X(\omega^2, \nu) = \boldsymbol{X}_k \hat{\boldsymbol{\beta}}_k$$

where $\hat{\boldsymbol{\beta}}_k = (\boldsymbol{X}_k^\top \boldsymbol{X}_k + \boldsymbol{K}_k/g)^{-1}(\boldsymbol{X}_k^\top \boldsymbol{Y} + \boldsymbol{K}_k \tilde{\boldsymbol{\beta}}_k/g)$ as defined in (3.9), and $\hat{\boldsymbol{\beta}}_k$ is the MAP estimator for minimizing

$$\|\boldsymbol{Y} - \boldsymbol{X}_k \hat{\boldsymbol{\beta}}_k\|_2^2 + (\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k)^\top \boldsymbol{K}_k (\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k)/g.$$

over $\boldsymbol{\beta}_k \in \mathbb{R}^{d_k}$ for fixed sparsity pattern $k \in \wp$. ∎

A.16 Proof of Lemma 4

Proposition 14. $\boldsymbol{W}_\gamma$ *as defined in* (3.13) *satisfies the following inequality*

$$\boldsymbol{I}_n \leq \boldsymbol{W}_\gamma \leq (1/\nu)\boldsymbol{I}_n\,,$$

for any $\gamma \in \wp$.

Proof. Firstly it is easy to verify the following inequality by using SVD on matrix $\boldsymbol{K}_\gamma^{-1/2}(\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma)\boldsymbol{K}_\gamma^{-1/2}$,

$$\left(\boldsymbol{K}_\gamma^{-1/2}(\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma)\boldsymbol{K}_\gamma^{-1/2}\right)^{-1} \geq \left(\boldsymbol{K}_\gamma^{-1/2}(\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma)\boldsymbol{K}_\gamma^{-1/2} + \frac{1}{\nu g}\boldsymbol{I}_n\right)^{-1}$$

Then it follows that

$$\begin{aligned}
& (\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma)^{-1} - (\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + (1/\nu g)\boldsymbol{K}_\gamma)^{-1} \\
= \ & \boldsymbol{K}_\gamma^{-1/2}\left[\left(\boldsymbol{K}_\gamma^{-1/2}(\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma)\boldsymbol{K}_\gamma^{-1/2}\right)^{-1} - \left(\boldsymbol{K}_\gamma^{-1/2}(\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma)\boldsymbol{K}_\gamma^{-1/2} + \frac{1}{\nu g}\boldsymbol{I}_n\right)^{-1}\right] \\
& \cdot \boldsymbol{K}_\gamma^{-1/2} \\
\geq \ & 0
\end{aligned}$$

which results

$$\boldsymbol{X}_\gamma(\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + (1/\nu g)\boldsymbol{K}_\gamma)^{-1}\boldsymbol{X}_\gamma^\top \leq \boldsymbol{X}_\gamma(\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma)^{-1}\boldsymbol{X}_\gamma^\top$$

Then we have

$$\begin{aligned}
\boldsymbol{W}_\gamma &= \boldsymbol{I}_n + \frac{1-\nu}{\nu}\boldsymbol{X}_\gamma\left(\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \frac{1}{\nu g}\boldsymbol{K}_\gamma\right)^{-1}\boldsymbol{X}_\gamma^\top \\
&\leq \boldsymbol{I}_n + \frac{1-\nu}{\nu}\boldsymbol{X}_\gamma(\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma)^{-1}\boldsymbol{X}_\gamma^\top \leq (1/\nu)\boldsymbol{I}_n
\end{aligned}$$

where the last inequality is because $\boldsymbol{X}_\gamma(\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma)^{-1}\boldsymbol{X}_\gamma^\top \leq \boldsymbol{I}_n$. And $\boldsymbol{W}_\gamma \geq \boldsymbol{I}_n$ is obvious since $\boldsymbol{X}_\gamma\left(\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \frac{1}{\nu g}\boldsymbol{K}_\gamma\right)^{-1}\boldsymbol{X}_\gamma^\top \geq 0$. □

Proposition 15. *For any $k > 0$ and $0 \leq \alpha \leq 1$,*

$$\|\alpha\boldsymbol{\psi}^{(k-1)} + (1-\alpha)\boldsymbol{\psi}^{(k)}\|_2^2 \leq (1/\nu)L_2^2 .$$

Proof. Denote $\hat{\boldsymbol{f}}_\gamma = \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma$ and $a = t_k \frac{1-\nu}{\omega^2}$, then we have

$$\begin{aligned}
\boldsymbol{\psi}^{(k)} &= \boldsymbol{\psi}^{(k-1)} - t_k \nabla \log J(\boldsymbol{\psi}^{(k-1)}) \\
&= \boldsymbol{\psi}^{(k-1)} - a \sum_{j=1}^{M} \lambda_\gamma^{(k-1)} \boldsymbol{W}_\gamma (\boldsymbol{\psi}^{(k-1)} - \hat{\boldsymbol{f}}_\gamma) \\
&= \sum_{j=1}^{M} \lambda_\gamma^{(k-1)} \left[(\boldsymbol{I}_n - a\boldsymbol{W}_\gamma)\, \boldsymbol{\psi}^{(k-1)} + a\boldsymbol{W}_\gamma \hat{\boldsymbol{f}}_\gamma \right]
\end{aligned}$$

where $\boldsymbol{\lambda}^{(k-1)} \in \Lambda^{|}\wp|$ and

$$\begin{aligned}
&\lambda_\gamma^{(k-1)} \propto \pi_\gamma |\frac{1}{g}\boldsymbol{K}_\gamma|^{1/2} |\nu \boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma + \frac{1}{g}\boldsymbol{K}_\gamma|^{-1/2} \\
&\cdot \exp\left(\frac{1-\nu}{2\omega^2} (\boldsymbol{\psi}^{(k-1)} - \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma)^\top \boldsymbol{W}_\gamma (\boldsymbol{\psi}^{(k-1)} - \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma) \right) \\
&\cdot \exp\left(-\frac{1}{2\omega^2} \|\boldsymbol{Y} - \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma\|_2^2 - \frac{(\hat{\boldsymbol{\beta}}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)^\top \boldsymbol{K}_\gamma (\hat{\boldsymbol{\beta}}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)}{2g\omega^2} \right)
\end{aligned}$$

It follows that

$$\|\boldsymbol{\psi}^{(k)}\|_2^2 \;\leq\; \sum_{\gamma \in \wp} \lambda_\gamma^{(k-1)} \left\| (\boldsymbol{I}_n - a\boldsymbol{W}_\gamma)\boldsymbol{\psi}^{(k-1)} + a\boldsymbol{W}_\gamma \hat{\boldsymbol{f}}_\gamma \right\|_2^2$$

For each $\gamma \in \wp$, define $\mu_1^{(\gamma)}$ and $\mu_n^{(\gamma)}$ as the smallest and biggest eigenvalue of $\boldsymbol{W}_\gamma$. From Proposition 14, $t_k \leq \frac{\nu}{1-\nu}\omega^2$ implies $a \leq \nu \leq 1/\mu_n^{(\gamma)}$, and it follows that

$$(\boldsymbol{I}_n - a\boldsymbol{W}_\gamma) a\boldsymbol{W}_\gamma \geq 0 \ ,$$

and then we have

$$\begin{aligned}
& \left\| (\boldsymbol{I}_n - a\boldsymbol{W}_\gamma)\boldsymbol{\psi}^{(k-1)} + a\boldsymbol{W}_\gamma \hat{\boldsymbol{f}}_\gamma \right\|_2^2 \\
= \;& \left\| (\boldsymbol{I}_n - a\boldsymbol{W}_\gamma)\boldsymbol{\psi}^{(k-1)} \right\|_2^2 + \left\| a\boldsymbol{W}_\gamma \hat{\boldsymbol{f}}_\gamma \right\|_2^2 + 2\langle (\boldsymbol{I}_n - a\boldsymbol{W}_\gamma)\boldsymbol{\psi}^{(k-1)}, a\boldsymbol{W}_\gamma \hat{\boldsymbol{f}}_\gamma \rangle_2 \\
\leq \;& \left\| (\boldsymbol{I}_n - a\boldsymbol{W}_\gamma)\boldsymbol{\psi}^{(k-1)} \right\|_2^2 + \left\| a\boldsymbol{W}_\gamma \hat{\boldsymbol{f}}_\gamma \right\|_2^2 \\
& + \boldsymbol{\psi}^{(k-1)^\top} (\boldsymbol{I}_n - a\boldsymbol{W}_\gamma) a\boldsymbol{W}_\gamma \boldsymbol{\psi}^{(k-1)} + \hat{\boldsymbol{f}}_\gamma^\top (\boldsymbol{I}_n - a\boldsymbol{W}_\gamma) a\boldsymbol{W}_\gamma \hat{\boldsymbol{f}}_\gamma \\
= \;& \boldsymbol{\psi}^{(k-1)^\top} (\boldsymbol{I}_n - a\boldsymbol{W}_\gamma) \boldsymbol{\psi}^{(k-1)} + \hat{\boldsymbol{f}}_\gamma^\top a\boldsymbol{W}_\gamma \hat{\boldsymbol{f}}_\gamma
\end{aligned}$$

where the inequality is because: write $(\boldsymbol{I}_n - a\boldsymbol{W}_\gamma)a\boldsymbol{W}_\gamma = A^2 \geq 0$ for some positive-definite matrix $A \in \mathbb{R}^{n\times n}$ and $2\boldsymbol{h}_1^\top A^2\boldsymbol{h}_2 \leq \boldsymbol{h}_1^\top A^2\boldsymbol{h}_1 + \boldsymbol{h}_2^\top A^2\boldsymbol{h}_2$ for any $\boldsymbol{h}_1, \boldsymbol{h}_2 \in \mathbb{R}^n$.

Then we have

$$\begin{aligned}
\|\boldsymbol{\psi}^{(k)}\|_2^2 &\leq \sum_{j=1}^{M}\left[\boldsymbol{\psi}^{(k-1)^\top}(\boldsymbol{I}_n - a\boldsymbol{W}_\gamma)\boldsymbol{\psi}^{(k-1)} + \hat{\boldsymbol{f}}_\gamma^\top a\boldsymbol{W}_\gamma\hat{\boldsymbol{f}}_\gamma\right] \\
&\leq \sum_{j=1}^{M}\left[(1 - a\mu_1^{(\gamma)})\|\boldsymbol{\psi}^{(k-1)}\|_2^2 + a\mu_n^{(\gamma)}\|\hat{\boldsymbol{f}}_\gamma\|_2^2\right] \\
&\leq (1-a)\|\boldsymbol{\psi}^{(k-1)}\|_2^2 + (a/\nu)\|\hat{\boldsymbol{f}}_\gamma\|_2^2 \leq (1-a)\|\boldsymbol{\psi}^{(k-1)}\|_2^2 + (a/\nu)L_2^2
\end{aligned}$$

where the third inequality is using Proposition 14.

Then by mathematical induction we can conclude that for any $k \geq 0$, $\|\boldsymbol{\psi}^{(k)}\|_2^2 \leq (1/\nu)L_2^2$ and the reasons are as following: firstly, $\|\boldsymbol{\psi}^{(0)}\|_2^2 = 0 \leq (1/\nu)L_2^2$; secondly if $\|\boldsymbol{\psi}^{(k-1)}\|_2^2 \leq (1/\nu)L_2^2$ then

$$\begin{aligned}
\|\boldsymbol{\psi}^{(k)}\|_2^2 &\leq (1-a)\|\boldsymbol{\psi}^{(k-1)}\|_2^2 + (a/\nu)L_2^2 \\
&\leq (1-a)(1/\nu)L_2^2 + (a/\nu)L_2^2 = (1/\nu)L_2^2
\end{aligned}$$

Thus for any $k > 0$ and $0 \leq \alpha \leq 1$,

$$\|\alpha\boldsymbol{\psi}^{(k-1)} + (1-\alpha)\boldsymbol{\psi}^{(k)}\|_2^2 \leq \alpha\|\boldsymbol{\psi}^{(k-1)}\|_2^2 + (1-\alpha)\|\boldsymbol{\psi}^{(k)}\|_2^2 \leq (1/\nu)L_2^2.$$

□

Now we are ready to prove Lemma 4. It is easy to see that for any $\boldsymbol{\psi} \in \mathbb{R}^n$

that can be expressed as $\boldsymbol{\psi} = \alpha\boldsymbol{\psi}^{(k-1)} + (1-\alpha)\boldsymbol{\psi}^{(k)}$ for some $k > 0$ and $0 \le \alpha \le 1$,

$$\begin{aligned}
\nabla^2 \log J(\boldsymbol{\psi}) &= \frac{\nabla^2 J(\boldsymbol{\psi})}{J(\boldsymbol{\psi})} - (\nabla J(\boldsymbol{\psi})/J(\boldsymbol{\psi}))(\nabla J(\boldsymbol{\psi})/J(\boldsymbol{\psi}))^\top \\
&= \sum_{\gamma\in\wp} \lambda_\gamma \left[\left(\frac{1-\nu}{\omega^2}\right)^2 \boldsymbol{W}_\gamma(\boldsymbol{\psi} - \hat{\boldsymbol{f}}_\gamma)(\boldsymbol{\psi} - \hat{\boldsymbol{f}}_\gamma)^\top \boldsymbol{W}_\gamma + \frac{1-\nu}{\omega^2}\boldsymbol{W}_\gamma \right] \\
&\quad - \left(\frac{1-\nu}{\omega^2}\right)^2 \left(\sum_{\gamma\in\wp} \lambda_\gamma \boldsymbol{W}_\gamma(\boldsymbol{\psi} - \hat{\boldsymbol{f}}_\gamma) \right) \left(\sum_{\gamma\in\wp} \lambda_\gamma \boldsymbol{W}_\gamma(\boldsymbol{\psi} - \hat{\boldsymbol{f}}_\gamma) \right)^\top \\
&\le \sum_{\gamma\in\wp} \lambda_\gamma \left[\left(\frac{1-\nu}{\omega^2}\right)^2 \boldsymbol{W}_\gamma(\boldsymbol{\psi} - \hat{\boldsymbol{f}}_\gamma)(\boldsymbol{\psi} - \hat{\boldsymbol{f}}_\gamma)^\top \boldsymbol{W}_\gamma + \frac{1-\nu}{\omega^2}\boldsymbol{W}_\gamma \right]
\end{aligned}$$

Also notice that for any $\gamma \in \wp$,

$$\begin{aligned}
\operatorname{tr}\left((\boldsymbol{\psi} - \hat{\boldsymbol{f}}_\gamma)(\boldsymbol{\psi} - \hat{\boldsymbol{f}}_\gamma)^\top \right) &= \|\boldsymbol{\psi} - \hat{\boldsymbol{f}}_\gamma\|_2^2 \le 2(\|\boldsymbol{\psi}\|_2^2 + \|\hat{\boldsymbol{f}}_\gamma\|_2^2) \\
&\le 2(1/\nu + 1)L_2^2
\end{aligned}$$

where the last inequality is from Proposition 15.

With $\operatorname{rank}\left((\boldsymbol{\psi} - \hat{\boldsymbol{f}}_\gamma)(\boldsymbol{\psi} - \hat{\boldsymbol{f}}_\gamma)^\top \right) = 1$, there exists $0 < b_\gamma \le 2(1/\nu + 1)L_2^2$ and orthogonal matrix $\boldsymbol{O} \in \mathbb{R}^{n\times n}$ such that

$$(\boldsymbol{\psi} - \hat{\boldsymbol{f}}_\gamma)(\boldsymbol{\psi} - \hat{\boldsymbol{f}}_\gamma)^\top = \boldsymbol{O}^\top \operatorname{diag}(b_\gamma, 0, \dots, 0)\boldsymbol{O}$$

Then for any $u \in \mathbb{R}^n$,

$$\begin{aligned}
u^\top \boldsymbol{W}_\gamma(\boldsymbol{\psi} - \hat{\boldsymbol{f}}_\gamma)(\boldsymbol{\psi} - \hat{\boldsymbol{f}}_\gamma)^\top \boldsymbol{W}_\gamma u &= u^\top \boldsymbol{W}_\gamma \boldsymbol{O}^\top \operatorname{diag}(b_\gamma, 0, \dots, 0)\boldsymbol{O}\boldsymbol{W}_\gamma u \\
&\le b_\gamma \|\boldsymbol{O}\boldsymbol{W}_\gamma u\|_2^2 \le 2(1/\nu + 1)L_2^2 u^\top \boldsymbol{W}_\gamma^2 u \\
&\le 2(1/\nu + 1)L_2^2 (1/\nu)^2 \|u\|_2^2
\end{aligned}$$

where the last inequality is from $\boldsymbol{W}_\gamma \le (1/\nu)\boldsymbol{I}_n$. Then it follows that

$$\boldsymbol{W}_\gamma(\boldsymbol{\psi} - \hat{\boldsymbol{f}}_\gamma)(\boldsymbol{\psi} - \hat{\boldsymbol{f}}_\gamma)^\top \boldsymbol{W}_\gamma \le \frac{2(1+\nu)}{\nu^3} L_2^2 \boldsymbol{I}_n \,.$$

Therefore,

$$\nabla^2 \log J(\boldsymbol{\psi}) \le \left[\left(\tfrac{1-\nu}{\omega^2}\right)^2 \tfrac{2(1+\nu)}{\nu^3} L_2^2 + \tfrac{1-\nu}{\omega^2}(1/\nu) \right] \boldsymbol{I}_n = A_3 \boldsymbol{I}_n$$

Order all the sparsity pattern by $(\gamma^1, \ldots, \gamma^{|\wp|})$, it is easy to verify that for any $\boldsymbol{\psi} \in \mathbb{R}^n$,

$$\begin{aligned} \nabla^2 \log J(\boldsymbol{\psi}) &= \frac{1-\nu}{\omega^2} \sum_{\gamma \in \wp} \lambda_\gamma \boldsymbol{W}_\gamma + \left(\frac{1-\nu}{\omega^2}\right)^2 GFG^\top \\ &\geq \frac{1-\nu}{\omega^2} \sum_{\gamma \in \wp} \lambda_\gamma \boldsymbol{W}_\gamma \geq \frac{1-\nu}{\omega^2} \boldsymbol{I}_n \end{aligned}$$

where $G = (\boldsymbol{W}_{\gamma^1}(\boldsymbol{\psi} - \hat{\boldsymbol{f}}_{\gamma^1}), \ldots, \boldsymbol{W}_{\gamma^{|\wp|}}(\boldsymbol{\psi} - \hat{\boldsymbol{f}}_{\gamma^{|\wp|}})) \in \mathbb{R}^{n \times |\wp|}$, and

$$F = \operatorname{diag}(\lambda_1, \ldots, \lambda_M) - \boldsymbol{\lambda}\boldsymbol{\lambda}^\top \geq 0$$

with $\boldsymbol{\lambda} = (\lambda_{\gamma^1}, \ldots, \lambda_{\gamma^{|\wp|}})$.

∎

A.17 Proof of Theorem 6

Lemma 7. *Suppose $(Y_1, \cdots, Y_k)$ are i.i.d. standard Gaussian random variables. Let $a_1, \cdots, a_k$ be nonnegative numbers, and*

$$|a|_\infty = \sup_{i=1,\cdots,k} |a_i|, \quad |a|^2 = \left(\sum_{i=1}^k a_i^2\right)^{1/2},$$

and let

$$Z = \sum_{i=1}^k a_i(Y_i^2 - 1).$$

Then for any $u \in (0, \frac{1}{2|a|_\infty})$,

$$\mathbb{E}\left(e^{uZ}\right) \leq \exp\left(\frac{|a|_2^2 u^2}{1 - 2|a|_\infty u}\right).$$

Proof. This lemma follows directly from the proof of Lemma 1 in Laurent and Massart (2000b). □

Lemma 8. *Given any $\boldsymbol{\lambda} \in \Lambda^{|\wp|}$, when $\Phi \geq 32\sigma^2 V$, for any fixed $q \in \wp$ we have with probability at least $1 - \delta$:*

$$2\langle \boldsymbol{\xi}, \hat{\mathfrak{f}}_{\boldsymbol{\lambda}} - \hat{\boldsymbol{f}}_q \rangle_2 - \Phi \mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi}) - \Phi \sum_{\gamma \in \wp} \lambda_\gamma C_\gamma \leq \frac{32\sigma^2}{\Phi} \sum_{\gamma \in \wp} \lambda_\gamma \|\hat{\boldsymbol{f}}_\gamma - \hat{\boldsymbol{f}}_q\|_2^2 + \Phi \log\left(\frac{1}{\delta}\right).$$

Proof. Let $\Delta = 2\langle \boldsymbol{\xi}, \hat{\mathfrak{f}}_{\boldsymbol{\lambda}} - \hat{\boldsymbol{f}}_q \rangle_2 - \Phi \mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi}) - \Phi \sum_{\gamma\in\wp} \lambda_\gamma C_\gamma$, then we have

$$\begin{aligned}
& \mathbb{E}\left[\exp\left(\frac{\Delta}{\Phi} - \frac{32\sigma^2}{\Phi^2}\sum_{\gamma\in\wp}\lambda_\gamma \|\hat{\boldsymbol{f}}_\gamma - \hat{\boldsymbol{f}}_q\|_2^2\right)\right] \\
= \; & \mathbb{E}\left[\exp\left(\sum_{\gamma\in\wp}\lambda_\gamma\left(\frac{2}{\Phi}\langle \boldsymbol{\xi}, \hat{\boldsymbol{f}}_\gamma - \hat{\boldsymbol{f}}_q\rangle_2 - \log(\frac{\lambda_\gamma}{\pi_p}) - C_\gamma - \frac{32\sigma^2}{\Phi^2}\|\hat{\boldsymbol{f}}_\gamma - \hat{\boldsymbol{f}}_q\|_2^2\right)\right)\right] \\
\leq \; & \mathbb{E}\left[\sum_{\gamma\in\wp}\lambda_\gamma \exp\left(\frac{2}{\Phi}\langle \boldsymbol{\xi}, \hat{\boldsymbol{f}}_\gamma - \hat{\boldsymbol{f}}_q\rangle_2 - \log(\frac{\lambda_\gamma}{\pi_p}) - C_\gamma - \frac{32\sigma^2}{\Phi^2}\|\hat{\boldsymbol{f}}_\gamma - \hat{\boldsymbol{f}}_q\|_2^2\right)\right] \\
= \; & \mathbb{E}\left[\sum_{\gamma\in\wp}\pi_p \exp\left(\frac{2}{\Phi}\langle \boldsymbol{\xi}, \hat{\boldsymbol{f}}_\gamma - \hat{\boldsymbol{f}}_q\rangle_2 - C_\gamma - \frac{32\sigma^2}{\Phi^2}\|\hat{\boldsymbol{f}}_\gamma - \hat{\boldsymbol{f}}_q\|_2^2\right)\right] \\
= \; & \sum_{\gamma\in\wp}\pi_p \exp(-C_\gamma)\mathbb{E}\left[\exp\left(\frac{2}{\Phi}\boldsymbol{\xi}^\top(B_\gamma \boldsymbol{Y} + \boldsymbol{v}_\gamma) - \frac{32\sigma^2}{\Phi^2}(B_\gamma \boldsymbol{Y} + \boldsymbol{v}_\gamma)^\top(B_\gamma \boldsymbol{Y} + \boldsymbol{v}_\gamma)\right)\right]
\end{aligned}$$

where $B_\gamma = A_\gamma - A_q$ and $\boldsymbol{v}_\gamma = \boldsymbol{b}_\gamma - \boldsymbol{b}_q$. The inequality in the above derivation follows from Jensen's inequality.

Let's consider the term in expectation for each $\gamma \in \wp$,

$$\begin{aligned}
& \mathbb{E}\left[\exp\left(\frac{2}{\Phi}\boldsymbol{\xi}^\top(B_\gamma \boldsymbol{Y} + \boldsymbol{v}_\gamma) - \frac{32\sigma^2}{\Phi^2}(B_\gamma \boldsymbol{Y} + \boldsymbol{v}_\gamma)^\top(B_\gamma \boldsymbol{Y} + \boldsymbol{v}_\gamma)\right)\right] \\
&= \exp\left(-\frac{32\sigma^2}{\Phi^2}\boldsymbol{\eta}^\top B_\gamma^2 \boldsymbol{\eta} - \frac{64\sigma^2}{\Phi^2}\boldsymbol{\eta}^\top B_\gamma \boldsymbol{v}_\gamma - \frac{32\sigma^2}{\Phi^2}\boldsymbol{v}_\gamma^\top \boldsymbol{v}_\gamma\right) \\
&\cdot \mathbb{E}\left[\exp\left(\boldsymbol{\xi}^\top(\frac{2}{\Phi}B_\gamma - \frac{32\sigma^2}{\Phi^2}B_\gamma^2)\boldsymbol{\xi} + \boldsymbol{\xi}^\top(\frac{2}{\Phi}(B_\gamma \boldsymbol{\eta} + \boldsymbol{v}_\gamma) - \frac{64\sigma^2}{\Phi^2}B_\gamma^2 \boldsymbol{\eta} - \frac{64\sigma^2}{\Phi^2}B_\gamma \boldsymbol{v}_\gamma)\right)\right]
\end{aligned}$$

We obtain from Cauchy-Schwarz inequality that

$$\mathbb{E}\left[\exp\left(\frac{2}{\Phi}\boldsymbol{\xi}^\top(B_\gamma \boldsymbol{Y} + \boldsymbol{v}_\gamma) - \frac{32\sigma^2}{\Phi^2}(B_\gamma \boldsymbol{Y} + \boldsymbol{v}_\gamma)^\top(B_\gamma \boldsymbol{Y} + \boldsymbol{v}_\gamma)\right)\right] \leq P_1 P_2,$$

with

$$P_1 = \mathbb{E}\left[\exp\left(\frac{4}{\Phi}\boldsymbol{\xi}^\top(B_\gamma - \frac{16\sigma^2}{\Phi}B_\gamma^2)\boldsymbol{\xi}\right)\right]^{1/2},$$

$$\begin{aligned}
P_2 = \mathbb{E}&\left[\exp\left(\boldsymbol{\xi}^\top(\frac{4}{\Phi}(B_\gamma \boldsymbol{\eta} + \boldsymbol{v}_\gamma) - \frac{128\sigma^2}{\Phi^2}B_\gamma^2 \boldsymbol{\eta} - \frac{128\sigma^2}{\Phi^2}B_\gamma \boldsymbol{v}_\gamma)\right)\right]^{1/2} \\
&\cdot \exp\left(-\frac{32\sigma^2}{\Phi^2}\boldsymbol{\eta}^\top B_\gamma^2 \boldsymbol{\eta} - \frac{64\sigma^2}{\Phi^2}\boldsymbol{\eta}^\top B_\gamma \boldsymbol{v}_\gamma - \frac{32\sigma^2}{\Phi^2}\boldsymbol{v}_\gamma^\top \boldsymbol{v}_\gamma\right)
\end{aligned}$$

Now we will bound P_1 and P_2 separately and we consider P_1 first. By assumption, each matrix A_γ is symmetric and positive semi-definite and could be decomposed by SVD: $A_\gamma = Q_\gamma^\top D_\gamma Q_\gamma$, where $D_\gamma = \mathrm{diag}(\zeta_1^\gamma, \ldots, \zeta_n^\gamma)$ with $\zeta_1^\gamma \geq \ldots \geq \zeta_n^\gamma \geq 0$, and $Q_\gamma \in \mathbb{R}^{n\times n}$ is orthogonal. Therefore,

$$\frac{4}{\Phi}\boldsymbol{\xi}^\top(B_\gamma - \frac{16\sigma^2}{\Phi}B_\gamma^2)\boldsymbol{\xi} \leq \frac{4}{\Phi}\boldsymbol{\xi}^\top B_\gamma\boldsymbol{\xi} \leq \frac{4}{\Phi}\boldsymbol{\xi}^\top A_\gamma\boldsymbol{\xi} = \frac{4}{\Phi}[Q_\gamma\boldsymbol{\xi}]^\top D_\gamma[Q_\gamma\boldsymbol{\xi}] = \frac{4\sigma^2}{\Phi}\sum_{i=1}^{\mathrm{rk}(A_\gamma)}\zeta_i^\gamma Z_i^2,$$

with $Z = (Z_1, \cdots, Z_n)^\top = Q_\gamma\boldsymbol{\xi}/\sigma \sim \mathcal{N}(0, \boldsymbol{I}_n)$, $\mathrm{rk}(A_\gamma)$ denotes rank of A_γ.

Then we have

$$\begin{aligned} P_1 &\leq \mathbb{E}\left[\exp\left(\frac{4\sigma^2}{\Phi}\sum_{i=1}^{\mathrm{rk}(A_\gamma)}\zeta_i^\gamma Z_i^2\right)\right]^{1/2} \\ &\leq \exp\left(\frac{8\sigma^4\mathrm{tr}(A_p^2)}{\Phi^2 - 8V\Phi\sigma^2} + \frac{2\sigma^2\mathrm{tr}(A_\gamma)}{\Phi}\right) = \exp(C_\gamma)\ , \end{aligned}$$

where the second inequality is by applying Lemma 7 with $u = \frac{4\sigma^2}{\Phi}$, $a_i = \zeta_i^\gamma$, $k = \mathrm{rk}(A_\gamma)$.

To bound P_2, denote $\frac{32\sigma^2}{\Phi^2} = c$, we observe that

$$\begin{aligned} P_2^2 &= \exp\left(-2c\boldsymbol{\eta}^\top B_\gamma^2\boldsymbol{\eta} - 4c\boldsymbol{\eta}^\top B_\gamma\boldsymbol{v}_\gamma - 2c\|\boldsymbol{v}_\gamma\|_2^2\right) \\ &\cdot\ \mathbb{E}\left[\exp\left(\boldsymbol{\xi}^\top\left(\frac{4}{\Phi}(B_\gamma\boldsymbol{\eta} + \boldsymbol{v}_\gamma) - 4cB_\gamma^2\boldsymbol{\eta} - 4cB_\gamma\boldsymbol{v}_\gamma\right)\right)\right] \\ &\leq \exp\left(-2c\boldsymbol{\eta}^\top B_\gamma^2\boldsymbol{\eta} - 4c\boldsymbol{\eta}^\top B_\gamma\boldsymbol{v}_\gamma - 2c\|\boldsymbol{v}_\gamma\|_2^2\right) \\ &\quad \exp\left(\sigma^2(\boldsymbol{\eta}^\top(\frac{8}{\Phi^2}B_\gamma^2 - \frac{16c}{\Phi}B_\gamma^3 + 8c^2B_\gamma^4)\boldsymbol{\eta} + \boldsymbol{\eta}^\top(\frac{16}{\Phi^2}B_\gamma - \frac{32c}{\Phi}B_\gamma^2 + 16c^2B_\gamma^3)\boldsymbol{v}_\gamma)\right) \\ &\quad \exp\left(\sigma^2(\boldsymbol{v}_\gamma^\top(\frac{8}{\Phi^2} - \frac{16c}{\Phi}B_\gamma + 8c^2B_\gamma^2)\boldsymbol{v}_\gamma)\right) \\ &\leq \exp\left(c\left(-\frac{7}{4}\|B_\gamma\boldsymbol{\eta} + \boldsymbol{v}_\gamma\|_2^2 + \frac{16\sigma^2}{\Phi}\|B_\gamma\|_{\mathrm{op}}\|B_\gamma\boldsymbol{\eta} + \boldsymbol{v}_\gamma\|_2^2\right.\right. \\ &\quad \left.\left. + \frac{256\sigma^4}{\Phi^2}\|B_\gamma\|_{\mathrm{op}}^2\|B_\gamma\boldsymbol{\eta} + \boldsymbol{v}_\gamma\|_2^2\right)\right) \\ &\leq \exp\left(c\|B_\gamma\boldsymbol{\eta} + \boldsymbol{v}_\gamma\|_2^2(-\frac{7}{4} + \frac{16V\sigma^2}{\Phi} + \frac{256V^2\sigma^4}{\Phi^2})\right) \end{aligned}$$

In the above derivation, the first inequality is from $\mathbb{E}e^{\boldsymbol{\xi}^\top\boldsymbol{f}} \leq e^{\sigma^2\|\boldsymbol{f}\|_2^2/2}$, the second inequality is by simple algebra, and the third inequality follows because

$\|B_\gamma\|_{\mathrm{op}} \le V$. By our assumption of Φ, we know that

$$\frac{256V^2\sigma^4}{\Phi^2} + \frac{16V\sigma^2}{\Phi} \le \frac{7}{4}\,,$$

and hence $P_2 < 1$.

With the bounds on P_1 and P_2 and Markov inequality, we see

$$\begin{aligned}
&\mathbb{P}\left[\frac{\Delta}{\Phi} - \frac{32\sigma^2}{\Phi^2}\sum_{\gamma\in\wp}\lambda_\gamma\|\hat{\boldsymbol{f}}_\gamma - \hat{\boldsymbol{f}}_q\|_2^2 \ge \log\left(\frac{1}{\delta}\right)\right] \\
= \;&\mathbb{P}\left[\exp\left(\frac{\Delta}{\Phi} - \frac{32\sigma^2}{\Phi^2}\sum_{\gamma\in\wp}\lambda_\gamma\|\hat{\boldsymbol{f}}_\gamma - \hat{\boldsymbol{f}}_q\|_2^2\right) \ge \frac{1}{\delta}\right] \\
\le \;&\delta\mathbb{E}\exp\left(\frac{\Delta}{\Phi} - \frac{32\sigma^2}{\Phi^2}\sum_{\gamma\in\wp}\lambda_\gamma\|\hat{\boldsymbol{f}}_\gamma - \hat{\boldsymbol{f}}_q\|_2^2\right) \le \delta.
\end{aligned}$$

Hence with probability of at least $1-\delta$,

$$\Delta \le \frac{32\sigma^2}{\Phi}\sum_{\gamma\in\wp}\lambda_\gamma\|\hat{\boldsymbol{f}}_\gamma - \hat{\boldsymbol{f}}_q\|_2^2 + \Phi\log\left(\frac{1}{\delta}\right).$$

□

Define

$$\hat{S}(\boldsymbol{\theta}) = (1-\nu)\|\mathfrak{f}_{\boldsymbol{\theta}} - \boldsymbol{Y}\|_2^2 + \nu\sum_{\gamma\in\wp}\theta_\gamma\|\hat{\boldsymbol{f}}_\gamma - \boldsymbol{Y}\|_2^2\,,$$

and

$$S(\boldsymbol{\theta}) = (1-\nu)\|\mathfrak{f}_{\boldsymbol{\theta}} - \boldsymbol{\eta}\|_2^2 + \nu\sum_{\gamma\in\wp}\theta_\gamma\|\hat{\boldsymbol{f}}_\gamma - \boldsymbol{\eta}\|_2^2\,.$$

Lemma 9. *For any $q \in \wp$, and $\alpha \in (0,1)$, we have*

$$\begin{aligned}
\|\mathfrak{f}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\eta}\|_2^2 - \|\hat{\boldsymbol{f}}_q - \boldsymbol{\eta}\|_2^2 \le -(1-\nu)\|\mathfrak{f}_{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{f}}_q\|_2^2 + 2\langle\boldsymbol{\xi}, \mathfrak{f}_{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{f}}_q\rangle_2 - \nu\sum_{\gamma\in\wp}\hat{\theta}_\gamma\|\hat{\boldsymbol{f}}_\gamma - \mathfrak{f}_{\hat{\boldsymbol{\theta}}}\|_2^2 \\
- \Phi\sum_{\gamma\in\wp}\hat{\theta}_\gamma C_\gamma + \Phi C_q - \Phi\mathcal{K}(\hat{\boldsymbol{\theta}}, \boldsymbol{\pi}) + \Phi\log(\frac{1}{\pi_q})
\end{aligned}$$

Proof. Easily seen that, for $\forall\ \boldsymbol{\theta} \in \Lambda^{|\wp|}$:

$$\hat{S}(\boldsymbol{\theta}) - S(\boldsymbol{\theta}) = \|\boldsymbol{Y}\|_2^2 - \|\boldsymbol{\eta}\|_2^2 - 2\langle\boldsymbol{\xi}, \mathfrak{f}_{\boldsymbol{\theta}}\rangle_2.$$

And by definition,

$$\hat{S}(\hat{\boldsymbol{\theta}}) + \Phi \sum_{\gamma \in \wp} \hat{\theta}_\gamma C_\gamma + \Phi \mathcal{K}(\hat{\boldsymbol{\theta}}, \boldsymbol{\pi}) \leq \hat{S}(\boldsymbol{\theta}) + \Phi \sum_{\gamma \in \wp} \theta_\gamma C_\gamma + \Phi \mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\pi}).$$

Using the above equation and inequality, we have

$$S(\hat{\boldsymbol{\theta}}) \leq S(\boldsymbol{\theta}) + \Phi \sum_{\gamma \in \wp} (\theta_\gamma - \hat{\theta}_\gamma) C_\gamma + 2\langle \boldsymbol{\xi}, \mathfrak{f}_{\hat{\boldsymbol{\theta}}} - \mathfrak{f}_{\boldsymbol{\theta}} \rangle_2 + \Phi \mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\pi}) - \Phi \mathcal{K}(\hat{\boldsymbol{\theta}}, \boldsymbol{\pi})$$

which is equivalent to:

$$\begin{aligned}
&(1-\nu)[\|\mathfrak{f}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\eta}\|_2^2 - \|\mathfrak{f}_{\boldsymbol{\theta}} - \boldsymbol{\eta}\|_2^2] \\
&\leq 2\langle \boldsymbol{\xi}, \mathfrak{f}_{\hat{\boldsymbol{\theta}}} - \mathfrak{f}_{\boldsymbol{\theta}} \rangle_2 + \nu \sum_{\gamma \in \wp} (\theta_\gamma - \hat{\theta}_\gamma) \|\hat{\boldsymbol{f}}_\gamma - \boldsymbol{\eta}\|_2^2 + \Phi \sum_{\gamma \in \wp} (\theta_\gamma - \hat{\theta}_\gamma) C_\gamma + \Phi \mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\pi}) \\
&- \Phi \mathcal{K}(\hat{\boldsymbol{\theta}}, \boldsymbol{\pi}).
\end{aligned}$$

Now we pick $\mathsf{e}_q \in \Lambda^{|\wp|}$, where $\mathsf{e}_q \in \mathbb{R}^{|\wp|}$ has value 0 in any coordinate except at position q where it takes value 1. Therefore, $\mathfrak{f}_{\mathsf{e}_q} = \hat{\boldsymbol{f}}_q$. Furthermore, we pick $\boldsymbol{\theta} = (1-\alpha)\hat{\boldsymbol{\theta}} + \alpha \mathsf{e}_q$. By simple algebra, we get

$$\|\mathfrak{f}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\eta}\|_2^2 - \|\mathfrak{f}_{\boldsymbol{\theta}} - \boldsymbol{\eta}\|_2^2 = \alpha \|\mathfrak{f}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\eta}\|_2^2 - \alpha \|\hat{\boldsymbol{f}}_q - \boldsymbol{\eta}\|_2^2 + \alpha(1-\alpha) \|\mathfrak{f}_{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{f}}_q\|_2^2 .$$

Hence,

$$\begin{aligned}
&(1-\nu)\alpha[\|\mathfrak{f}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\eta}\|_2^2 - \|\hat{\boldsymbol{f}}_q - \boldsymbol{\eta}\|_2^2] \\
&\leq -(1-\nu)\alpha(1-\alpha)\|\mathfrak{f}_{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{f}}_q\|_2^2 + 2\alpha\langle \boldsymbol{\xi}, \mathfrak{f}_{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{f}}_q \rangle_2 - \nu\alpha \sum_{\gamma \in \wp} \hat{\theta}_\gamma \|\hat{\boldsymbol{f}}_\gamma - \boldsymbol{\eta}\|_2^2 \\
&+ \nu\alpha \|\hat{\boldsymbol{f}}_q - \boldsymbol{\eta}\|_2^2 - \Phi\alpha \sum_{\gamma \in \wp} \hat{\theta}_\gamma C_\gamma + \Phi\alpha C_q + \Phi\mathcal{K}((1-\alpha)\hat{\boldsymbol{\theta}} + \alpha\mathsf{e}_q, \boldsymbol{\pi}) - \Phi\mathcal{K}(\hat{\boldsymbol{\theta}}, \boldsymbol{\pi}) \\
&\leq -(1-\nu)\alpha(1-\alpha)\|\mathfrak{f}_{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{f}}_q\|_2^2 + 2\alpha\langle \boldsymbol{\xi}, \mathfrak{f}_{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{f}}_q \rangle_2 - \nu\alpha \sum_{\gamma \in \wp} \hat{\theta}_\gamma \|\hat{\boldsymbol{f}}_\gamma - \boldsymbol{\eta}\|_2^2 \\
&+ \nu\alpha \|\hat{\boldsymbol{f}}_q - \boldsymbol{\eta}\|_2^2 - \Phi\alpha \sum_{\gamma \in \wp} \hat{\theta}_\gamma C_\gamma + \Phi\alpha C_q - \Phi\alpha\mathcal{K}(\hat{\boldsymbol{\theta}}, \boldsymbol{\pi}) + \Phi\alpha\mathcal{K}(\mathsf{e}_q, \boldsymbol{\pi}).
\end{aligned}$$

The last inequality is due to the convexity of the Kullback-Leibler distance. Using the equation

$$\sum_{\gamma \in \wp} \hat{\theta}_\gamma \|\hat{\boldsymbol{f}}_\gamma - \boldsymbol{\eta}\|_2^2 = \|\mathfrak{f}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\eta}\|_2^2 + \sum_{\gamma \in \wp} \hat{\theta}_\gamma \|\hat{\boldsymbol{f}}_\gamma - \mathfrak{f}_{\hat{\boldsymbol{\theta}}}\|_2^2 .$$

Divide both hand sides of the above inequality by α and let $\alpha \to 0+$, we obtain the desired inequality by rearranging the terms.

□

We are now ready to prove Theorem 6. It is easy to verify that

$$-(1-\nu)\|\mathfrak{f}_{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{f}}_q\|_2^2 - \nu \sum_{\gamma \in \wp} \hat{\theta}_\gamma \|\hat{\boldsymbol{f}}_\gamma - \mathfrak{f}_{\hat{\boldsymbol{\theta}}}\|_2^2 \le -\nu_1 \sum_{\gamma \in \wp} \hat{\theta}_\gamma \|\hat{\boldsymbol{f}}_\gamma - \hat{\boldsymbol{f}}_q\|_2^2 \qquad \text{(A.7)}$$

where $\nu_1 = \min(\nu, 1-\nu)$.

Combining Lemma 9 and equation (A.7), we see that

$$\|\mathfrak{f}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\eta}\|_2^2 - \|\hat{\boldsymbol{f}}_q - \boldsymbol{\eta}\|_2^2 \le \Phi \log(\frac{1}{\pi_q}) + \Phi C_q + \Delta - \nu_1 \sum_{\gamma \in \wp} \hat{\theta}_\gamma \|\hat{\boldsymbol{f}}_\gamma - \hat{\boldsymbol{f}}_q\|_2^2 \qquad \text{(A.8)}$$

where $\Delta = 2\langle \boldsymbol{\xi}, \mathfrak{f}_{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{f}}_q \rangle_2 - \Phi \mathcal{K}(\hat{\boldsymbol{\theta}}, \boldsymbol{\pi}) - \Phi \sum_{\gamma \in \wp} \hat{\theta}_\gamma C_\gamma$.

Using Lemma 8, we know that with probability of at least $1 - \delta$,

$$\Delta \le \frac{32\sigma^2}{\Phi} \sum_{\gamma \in \wp} \hat{\theta}_\gamma \|\hat{\boldsymbol{f}}_\gamma - \hat{\boldsymbol{f}}_q\|_2^2 + \Phi \log\left(\frac{1}{\delta}\right) ,$$

where $\Phi \ge 32\sigma^2(V \vee (\min(\nu, 1-\nu))^{-1}) \ge 32\sigma^2 V$.

Combine the above inequality with (A.8), we obtain

$$\|\mathfrak{f}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\eta}\|_2^2 - \|\hat{\boldsymbol{f}}_q - \boldsymbol{\eta}\|_2^2 \le \Phi \log(\frac{1}{\pi_q \delta}) + \Phi C_q + (32\sigma^2/\Phi - \nu_1) \sum_{\gamma \in \wp} \hat{\theta}_\gamma \|\hat{\boldsymbol{f}}_\gamma - \hat{\boldsymbol{f}}_q\|_2^2$$

Then theorem is concluded with $32\sigma^2/\Phi - \nu_1 \le 0$ since

$$\Phi \ge 32\sigma^2(V \vee (\min(\nu, 1-\nu))^{-1}) \, .$$

■

A.18 Proof of Theorem 7

For any fixed $l \in \{1, \ldots, M\}$, we just specify $\boldsymbol{f}_\gamma$ with distribution $\Theta(\gamma)$ in Corollary 2 as following:

$$\boldsymbol{f}_\gamma | \boldsymbol{\beta}_l, \mathcal{M}_l = \boldsymbol{X}_l \boldsymbol{\beta}_l,$$

and $\boldsymbol{\beta}_l|\mathcal{M}_l \sim N(\boldsymbol{\beta}_l^*, \omega^2 \boldsymbol{D}_l^{-1})$ and $p(\mathcal{M}_l) = 1$.

While $\pi(\gamma)$ is defined by $\forall j \in \{1, \ldots, M\}$,

$$\boldsymbol{f}_\gamma|\boldsymbol{\beta}_j, \mathcal{M}_j = \boldsymbol{X}_j\boldsymbol{\beta}_j,$$

and $p(\boldsymbol{\beta}_j|g_j, \mathcal{M}_j)$ is as defined in (4.3), $p(g_j|\mathcal{M}_j) \sim$ Inv-Gamma(α, θ) where $\theta = (d_l/2 + \alpha)g_0$, and $p(\mathcal{M}_j) = \pi_j$.

Then we apply Corollary 2, write the right hand side of oracle deviation inequality (2.22) as

$$RHS = \nu \int_\Omega \|\boldsymbol{f}_\gamma - \boldsymbol{\eta}\|_2^2 \Theta(\gamma)\ d\gamma + (1-\nu) \left\| \int_\Omega \boldsymbol{f}_\gamma \Theta(\gamma)\ d\gamma - \boldsymbol{\eta} \right\|_2^2 + 2\omega^2 \mathcal{K}(\Theta, \pi)$$

The notation of $\mathbb{E}(\cdot)$ below is for expectation with respect to $\Theta(\gamma)$, specifically $\boldsymbol{f}_\gamma = \boldsymbol{X}_l\boldsymbol{\beta}_l$ with

$$\boldsymbol{\beta}_l \sim N(\boldsymbol{\beta}_l^*, \omega^2 \boldsymbol{D}_l^{-1}),$$

then we have

$$\begin{aligned} \int_\Omega \|\boldsymbol{f}_\gamma - \boldsymbol{\eta}\|_2^2 \Theta(\gamma)\ d\gamma &= \mathbb{E}\|\boldsymbol{X}_l\boldsymbol{\beta}_l - \boldsymbol{\eta}\|_2^2 \\ &= \mathrm{tr}(\mathrm{COV}(\boldsymbol{X}_l\boldsymbol{\beta}_l)) + \|\boldsymbol{X}_l\boldsymbol{\beta}_l^* - \boldsymbol{\eta}\|_2^2 \\ &= \omega^2 \mathrm{tr}(\boldsymbol{X}_l^\top \boldsymbol{X}_l \boldsymbol{D}_l^{-1}) + \|\boldsymbol{X}_l\boldsymbol{\beta}_l^* - \boldsymbol{\eta}\|_2^2 \end{aligned}$$

and

$$\left\| \int_\Omega \boldsymbol{f}_\gamma \Theta(\gamma)\ d\gamma - \boldsymbol{\eta} \right\|_2^2 = \|\boldsymbol{X}_l\boldsymbol{\beta}_l^* - \boldsymbol{\eta}\|_2^2$$

Also

$$\mathcal{K}(\Theta, \pi) = \mathbb{E}\left[\log \frac{(2\pi\omega^2)^{-d_l/2} |\boldsymbol{D}_l|^{1/2} \exp\left(-\frac{(\boldsymbol{\beta}_l - \boldsymbol{\beta}_l^*)^\top \boldsymbol{D}_l (\boldsymbol{\beta}_l - \boldsymbol{\beta}_l^*)}{2\omega^2} \right)}{\pi_l h(\boldsymbol{\beta}_l)} \right]$$

where

$$\begin{aligned} h(\boldsymbol{\beta}_l) &= (2\pi\omega^2)^{-d_l/2}|\boldsymbol{K}_l|^{1/2}\theta^{-d_l/2}\frac{\Gamma(d_l/2+\alpha)}{\Gamma(\alpha)} \\ &\quad \cdot \left(1+\frac{(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)^\top \boldsymbol{K}_l(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)}{2\theta\omega^2}\right)^{-(d_l/2+\alpha)} \end{aligned}$$

Notice that

$$\begin{aligned} \mathbb{E}(\boldsymbol{\beta}_l-\boldsymbol{\beta}_l^*)^\top \boldsymbol{D}_l(\boldsymbol{\beta}_l-\boldsymbol{\beta}_l^*) &= \mathbb{E}\mathrm{tr}\left[\left(\boldsymbol{D}_l^{1/2}(\boldsymbol{\beta}_l-\boldsymbol{\beta}_l^*)\right)\left(\boldsymbol{D}_l^{1/2}(\boldsymbol{\beta}_l-\boldsymbol{\beta}_l^*)\right)^\top\right] \\ &= \mathrm{tr}\left[\mathrm{COV}\left(\boldsymbol{D}_l^{1/2}(\boldsymbol{\beta}_l-\boldsymbol{\beta}_l^*)\right)\right] = d_l\omega^2 \end{aligned}$$

Also by combining

$$\mathbb{E}\log\left(1+\frac{(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)^\top \boldsymbol{K}_l(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)}{2\theta\omega^2}\right) \leq \mathbb{E}\frac{(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)^\top \boldsymbol{K}_l(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)}{2\theta\omega^2}$$

and

$$\begin{aligned} \mathbb{E}(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)^\top \boldsymbol{K}_l(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l) &= \mathbb{E}\mathrm{tr}\left[\left(\boldsymbol{K}_l^{1/2}(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)\right)\left(\boldsymbol{K}_l^{1/2}(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)\right)^\top\right] \\ &= \omega^2\mathrm{tr}(\boldsymbol{K}_l\boldsymbol{D}_l^{-1}) + (\boldsymbol{\beta}_l^*-\tilde{\boldsymbol{\beta}}_l)^\top \boldsymbol{K}_l(\boldsymbol{\beta}_l^*-\tilde{\boldsymbol{\beta}}_l) \end{aligned}$$

will results

$$\begin{aligned} &\mathbb{E}\log\left(1+\frac{(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)^\top \boldsymbol{K}_l(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)}{2n\theta\omega^2}\right) \\ \leq\ &\mathrm{tr}(\boldsymbol{K}_l\boldsymbol{D}_l^{-1})/(2\theta) + \frac{(\boldsymbol{\beta}_l^*-\tilde{\boldsymbol{\beta}}_l)^\top \boldsymbol{K}_l(\boldsymbol{\beta}_l^*-\tilde{\boldsymbol{\beta}}_l)}{2\theta\omega^2} \end{aligned}$$

Therefore we have

$$
\begin{aligned}
RHS &= \|\boldsymbol{X}_l\boldsymbol{\beta}_l^* - \boldsymbol{\eta}\|_2^2 + \nu\omega^2\mathrm{tr}(\boldsymbol{X}_l^\top\boldsymbol{X}_l\boldsymbol{D}_l^{-1}) \\
&\quad +2\omega^2\mathbb{E}\left[\log\frac{(2\pi\omega^2)^{-d_l/2}|\boldsymbol{D}_l|^{1/2}\exp\left(-\frac{(\boldsymbol{\beta}_l-\boldsymbol{\beta}_l^*)^\top\boldsymbol{D}_l(\boldsymbol{\beta}_l-\boldsymbol{\beta}_l^*)}{2\omega^2}\right)}{\pi_l h(\boldsymbol{\beta}_l)}\right] \\
&= \|\boldsymbol{X}_l\boldsymbol{\beta}_l^* - \boldsymbol{\eta}\|_2^2 + \nu\omega^2\mathrm{tr}(\boldsymbol{X}_l^\top\boldsymbol{X}_l\boldsymbol{D}_l^{-1}) \\
&\quad +2\omega^2\left(\log\left((2\pi\omega^2)^{-d_l/2}|\boldsymbol{D}_l|^{1/2}\right) - (d_l/2) + \log(1/\pi_l) - \mathbb{E}\log h(\boldsymbol{\beta}_l)\right) \\
&= \|\boldsymbol{X}_l\boldsymbol{\beta}_l^* - \boldsymbol{\eta}\|_2^2 + \nu\omega^2\mathrm{tr}(\boldsymbol{X}_l^\top\boldsymbol{X}_l\boldsymbol{D}_l^{-1}) \\
&\quad -d_l\omega^2\log(2\pi\omega^2) + \omega^2\log|\boldsymbol{D}_l| - d_l\omega^2 + 2\omega^2\log(1/\pi_l) \\
&\quad -2\omega^2\log\left((2\pi\omega^2)^{-d_l/2}|\boldsymbol{K}_l|^{1/2}\theta^{-d_l/2}\frac{\Gamma(d_l/2+\alpha)}{\Gamma(\alpha)}\right) \\
&\quad +2\omega^2(d_l/2+\alpha)\mathbb{E}ln\left(1+\frac{(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)^\top\boldsymbol{K}_l(\boldsymbol{\beta}_l-\tilde{\boldsymbol{\beta}}_l)}{2\theta\omega^2}\right) \\
&\le \|\boldsymbol{X}_l\boldsymbol{\beta}_l^* - \boldsymbol{\eta}\|_2^2 + \nu\omega^2\mathrm{tr}(\boldsymbol{X}_l^\top\boldsymbol{X}_l\boldsymbol{D}_l^{-1}) \\
&\quad +\omega^2\log|\boldsymbol{D}_l| - d_l\omega^2 + 2\omega^2\log(1/\pi_l) \\
&\quad -\omega^2\log|\boldsymbol{K}_l| + d_l\omega^2\log\theta - 2\omega^2\log\left(\frac{\Gamma(d_l/2+\alpha)}{\Gamma(\alpha)}\right) \\
&\quad +2\omega^2(d_l/2+\alpha)\left(\mathrm{tr}(\boldsymbol{K}_l\boldsymbol{D}_l^{-1})/(2\theta) + \frac{(\boldsymbol{\beta}_l^*-\tilde{\boldsymbol{\beta}}_l)^\top\boldsymbol{K}_l(\boldsymbol{\beta}_l^*-\tilde{\boldsymbol{\beta}}_l)}{2\theta\omega^2}\right) \\
&= \|\boldsymbol{X}_l\boldsymbol{\beta}_l^* - \boldsymbol{\eta}\|_2^2 + 2\omega^2\log(1/\pi_l) \\
&\quad +\omega^2\left(\log|\boldsymbol{D}_l| + \mathrm{tr}((\frac{d_l/2+\alpha}{\theta}\boldsymbol{K}_l + \nu\boldsymbol{X}_l^\top\boldsymbol{X}_l)\boldsymbol{D}_l^{-1}) - d_l\right) \\
&\quad -\omega^2\log|\boldsymbol{K}_l| + d_l\omega^2\log\theta - 2\omega^2\log\left(\frac{\Gamma(d_l/2+\alpha)}{\Gamma(\alpha)}\right) \\
&\quad +\frac{d_l/2+\alpha}{\theta}(\boldsymbol{\beta}_l^*-\tilde{\boldsymbol{\beta}}_l)^\top\boldsymbol{K}_l(\boldsymbol{\beta}_l^*-\tilde{\boldsymbol{\beta}}_l)
\end{aligned}
$$

where $\boldsymbol{\beta}_l^*$ and $\boldsymbol{D}_l$ are to be decided to minimize RHS.

Now we minimize RHS over $\boldsymbol{D}_l \in \mathbb{R}^{d_l\times d_l}$ being positive definite. By using Proposition 13, when

$$\boldsymbol{D}_l = \frac{d_l/2+\alpha}{\theta}\boldsymbol{K}_l + \nu\boldsymbol{X}_l^\top\boldsymbol{X}_l$$

the minimum is obtained, and RHS becomes

$$\begin{aligned} RHS &\leq \|\boldsymbol{X}_l\boldsymbol{\beta}_l^* - \boldsymbol{\eta}\|_2^2 + 2\omega^2\log(1/\pi_l) \\ &+\omega^2\log|\frac{d_l/2+\alpha}{\theta}\boldsymbol{K}_l + \nu\boldsymbol{X}_l^\top\boldsymbol{X}_l| \\ &-\omega^2\log|\boldsymbol{K}_l| + d_l\omega^2\log\theta - 2\omega^2\log\left(\frac{\Gamma(d_l/2+\alpha)}{\Gamma(\alpha)}\right) \\ &+\frac{d_l/2+\alpha}{\theta}(\boldsymbol{\beta}_l^* - \tilde{\boldsymbol{\beta}}_l)^\top\boldsymbol{K}_l(\boldsymbol{\beta}_l^* - \tilde{\boldsymbol{\beta}}_l) \end{aligned}$$

Since $\theta = (d_l/2+\alpha)g_0$ it follows that

$$\begin{aligned} RHS &\leq \|\boldsymbol{X}_l\boldsymbol{\beta}_l^* - \boldsymbol{\eta}\|_2^2 + (1/g_0)(\boldsymbol{\beta}_l^* - \tilde{\boldsymbol{\beta}}_l)^\top\boldsymbol{K}_l(\boldsymbol{\beta}_l^* - \tilde{\boldsymbol{\beta}}_l) \\ &+2\omega^2\log(1/\pi_l) + \omega^2\log|\nu g_0\boldsymbol{X}_l^\top\boldsymbol{X}_l\boldsymbol{K}_l^{-1} + \boldsymbol{I}_{d_l}| \\ &+\omega^2\left(d_l\log(d_l/2+\alpha) - 2\log(\Gamma(d_l/2+\alpha)/\Gamma(\alpha))\right)\ . \end{aligned}$$

∎

References

Audibert, J.-Y. (2008). Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems 20* (J. Platt, D. Koller, Y. Singer and S. Roweis, eds.). MIT Press, Cambridge, MA, 41–48.

Barron, A. R. (1993). Universal approximation bounds for superposition of a sigmoidal function. *IEEE Transactions on Information Theory*, **39** 930–945.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, **2** 183–202. URL http://dx.doi.org/10.1137/080716542.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer, NewYork.

Berger, J. (2006). The case for objective bayesian analysis. *Bayesian Analysis*, **1** 385–402.

Berger, J. O. and Pericchi, L. (1996). The intrinsic bayes factor for model selection and prediction. *The Journal of the American Statistical Association*, **91** 109–122.

BERGER, J. O. and PERICCHI, L. R. (2001). Objective bayesian methods for model selection: Introduction and comparison (with discussion). *Model Selection*, **38** 135–207.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization.* Cambridge University Press, Cambridge.

Casella, G., Girn, F. J., Martnez, M. L. and Moreno, E. (2009). Consistency of bayesian procedures for variable selection. *The Annals of Statistics*, **37** 1207–1228.

Casella, G. and Moreno, E. (2006). Objective bayesian variable selection. *Journal of the American Statistical Association*, **101** 157–167.

Catoni, O. (1999). Universal aggregation rules with exact bias bounds. Tech. rep., Laboratoire de Probabilités et Modeles Aléatoires, Preprint 510.

Clarkson, K. (2008). Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*. 922–931.

CLYDE, M., DESIMONE, H. and PARMIGIANI, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, **91** 1197C1208.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical Science*, **19** 81–94.

Cui, W. and George, E. I. (2008). Empirical bayes vs. fully bayes variable selection. *Journal of Statistical Planning and Inference*, **138** 888–900.

Dai, D., Rigollet, P. and Zhang, T. (2012). Deviation optimal learning using greedy q-aggregation. *Ann. Statist.*, **40** 1878–1905.

Dai, D. and Zhang, T. (2011). Greedy model averaging. In *Advances in Neural Information Processing Systems 24* (J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira and K. Weinberger, eds.). 1242–1250.

Dalalyan, A. and Tsybakov, A. (2008). Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning*, **72** 39–61.

Dalalyan, A. S. and Salmon, J. (2012). Sharp oracle inequalities for aggregation of affine estimators. *The Annals of Statistics*, **40** 2327–2355.

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

De Santis, F. and Spezzaferri, F. (1997). Alternative bayes factors for model selection. *Canadian Journal of Statistics*, **25** 503–515.

DRAPER, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society*, **57** 45–97.

Eicher, T. S., Papageorgiou, C. and Raftery, A. E. (2011). Default priors and predictive performance in bayesian model averaging with application to growth determinants. *Journal of Applied Econometrics*, **26** 30–55.

Fernández, C., Ley, E. and Steel, M. F. (2001). Benchmark priors for bayesian model averaging. *Journal of Econometrics*, **100** 381–427.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, **22** 1947–1975.

Frank, M. and Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Res. Logist. Quart.*, **3** 95–110.

Gaïffas, S. and Lecué, G. (2011). Hyper-sparse optimal aggregation. *J. Mach. Learn. Res.*, **12** 1813–1833.

GEISSER, S. (1993). *Predictive Inference: An Introduction.* Chapman and Hall, London.

George, E. I. and Foster, D. P. (2000). Calibration and empirical bayes variable selection. *Biometrika*, **87** 731–747.

HJORT, N. L. and CLAESKENS, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, **98** 879–899.

JAGGI, M. (2011). Convex optimization without projection steps. `1108.1170v6`, URL `http://arxiv.org/abs/1108.1170v6`.

JONES, L. (1992). A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann. Statist.*, **20** 608–613.

JUDITSKY, A. and NEMIROVSKI, A. (2000). Functional aggregation for nonparametric regression. *Ann. Statist.*, **28** 681–712.

JUDITSKY, A., RIGOLLET, P. and TSYBAKOV, A. B. (2008). Learning by mirror averaging. *Ann. Statist.*, **36** 2183–2206.

KASS, R. and WASSERMAN, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of American Statistical Association*, **90** 928–934.

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90** 377–395.

LAURENT, B. and MASSART, P. (2000a). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, **28** 1302–1338. URL `http://dx.doi.org/10.1214/aos/1015957395`.

LAURENT, B. and MASSART, P. (2000b). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, **28** 1302–1338.

LEAMER, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* Wiley, New York.

Lecué, G. and Mendelson, S. (2009). Aggregation via empirical risk minimization. *Probab. Theory Related Fields*, **145** 591–613. URL `http://dx.doi.org/10.1007/s00440-008-0180-8`.

Lecué, G. and Mendelson, S. (2012). On the optimality of the aggregate with exponential weights for low temperatures. *Bernoulli*.

Leung, G. and Barron, A. (2006). Information theory and mixing least-squares regressions. *Information Theory, IEEE Transactions on*, **52** 3396–3410.

Ley, E. and Steel, M. F. (2009). On the effect of prior assumptions in bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, **24** 651–674.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, **103** 410–423.

Madigan, D. and Raftery, A. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of American Statistical Association*, **89** 1535–1546.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, **63** 215–232.

Marin, J. and Robert, C. (2007). *Bayesian Core: a practical approach to computational Bayesian statistics*. Springer.

Maruyama, Y. and George, E. I. (2011). Fully bayes factors with a generalized g-prior. *The Annals of Statistics*, **39** 2740–2765.

Moreno, E., Girn, F. J. and Casella, G. (2010). Consistency of objective bayes factors as the model dimension grows. *The Annals of Statistics*, **38** 1937–1952.

Nemirovski, A. (2000). Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, vol. 1738 of *Lecture Notes in Math.* Springer, Berlin, 85–277.

Nesterov, Y. and Nesterov, I. E. (2004). *Introductory lectures on convex optimization: A basic course*, vol. 87. Springer.

O'Hagan, A. (1995). Fractional bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society*, **57** 99–138.

Pérez, J. M. and Berger, J. O. (2002). Expected-posterior prior distributions for model selection. *Biometrika*, **89** 491–512.

Poljak, B. T. (1987). *Introduction to optimization.* Optimization Software.

Raftery, A., Madigan, D. and Hoeting, J. (1997). Bayesian model averaging for linear regression models. *Journal of American Statistical Association*, **92** 179–191.

Raftery, A. E., Madigan, D. and Volinsky, C. T. (1996). Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian statistics*, **5** 323–349.

Rigollet, P. (2012). Kullback–leibler aggregation and misspecified generalized linear models. *Ann. Statist.*, **40** 639–665.

Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, **39** 731–771.

Rigollet, P. and Tsybakov, A. (2012). Sparse estimation by exponential weighting. *Statistical Science (to appear). arXiv:1108.5116.*

Rockafellar, R. T. (1997). *Convex Analysis.* Princeton Landmarks.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6** 461–464.

Shalev-Shwartz, S., Srebro, N. and Zhang, T. (2010). Trading accuracy for sparsity in optimization problems with sparsity constrain ts. *Siam Journal on Optimization*, **20** 2807–2832.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* New York: Chapman and Hall.

Smith, A. and Roberts, G. (1993). Bayesian computation via the gibbs sampler and related markov chain monte carlo methods (with discussion). *Journal of the Royal Statistical Society*, **55** 3–23.

Strawderman, W. E. (1971). Proper bayes minimax estimators of the multivariate normal mean. *The Annals of Mathematical Statistics*, **42** 385C388.

Tierney, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of American Statistical Association*, **81** 82–86.

Tsybakov, A. B. (2003). Optimal rates of aggregation. In *COLT* (B. Schölkopf and M. K. Warmuth, eds.), vol. 2777 of *Lecture Notes in Computer Science*. Springer, 303–313.

Yang, Y. (1999). Model selection for nonparametric regression. *Statistica Sinica*, **9** 475–500.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*, **6** 233–243.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. (with discussion). *Bayesian Statistics*, **31** 585–603.

Vita

Dong Dai

2008-2013 Ph.D. in Statistics, Rutgers University, New Jersey, USA.

2003-2008 B.S. in Statistics, University of Science and Technology of China, Hefei, China.