Bayesian Model Averaging With Exponentiated Least Squares Loss

Dong Dai, Lei Han, Ting Yang, and Tong Zhang

Abstract—The model averaging problem is to average multiple models to achieve a prediction accuracy not much worse than that of the best single model in terms of mean-squared error. It is known that if the models are misspecified, model averaging is superior to model selection. Specifically, let n be the sample size, then the worst case regret of the former decays at a rate of $O(1/n)$, whereas the worst case regret of the latter decays at a rate of $O(1/\sqrt{n})$. The recently proposed Q-aggregation algorithm solves the model averaging problem with the optimal regret of $O(1/n)$ both in expectation and in deviation; however, it suffers from two limitations: 1) for continuous dictionary, the proposed greedy algorithm for solving Q-aggregation is not applicable and 2) the formulation of Q-aggregation appears *ad hoc* without clear intuition. This paper examines a different approach to model averaging by considering a Bayes estimator for deviation optimal model averaging by using exponentiated least squares loss. We establish a primal-dual relationship of this estimator and that of Q-aggregation and propose new algorithms that satisfactorily resolve the above-mentioned limitations of Q-aggregation.

Index Terms—Model averaging, exponentiated least squares loss, continuous dictionary, deviation optimal.

I. INTRODUCTION

THIS paper considers the model averaging problem, where the goal is to average multiple models in order to achieve improved prediction accuracy.

Let $x_1, \ldots, x_n$ be n given design points from 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 $\mathbf{f}_j = (f_j(x_1), \ldots, f_j(x_n))^\top \in \mathbb{R}^n$ 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. variables from $\mathcal{N}(0, \sigma^2)$.

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

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

We want to find an estimator $\hat{\eta}$ that mimics the function 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:

$$\text{MSE}(\hat{\eta}) \leq \min_{j=1,\ldots,M} \text{MSE}(f_j) + \Delta(n, M, \sigma^2), \quad (1)$$

where the remainder term $\Delta > 0$ should be as small as possible.

The problem of model averaging has been well-studied, and it is known (see, e.g., [2], [3]) that the smallest possible order for $\Delta(n, M, \sigma^2)$ is $\sigma^2 \log M / n$ for oracle inequalities in expectation, where "the smallest possible" is understood in the following minimax sense. There exists a dictionary $\mathcal{H} = \{f_1, \ldots, f_M\}$ such that the following lower bound holds. For any estimator $\hat{\eta}$, there exists a regression function η such that

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

for some positive constant C. It also implies that the lower bound holds not only in expectation but also with positive probability.

Although our goal is to achieve an MSE as close as that of the best model in $\mathcal{H}$, it is known (see [4, Th. 2.1]) that there exists a dictionary $\mathcal{H}$ such that any estimator $\hat{\eta}$ taking values restricted to the elements of $\mathcal{H}$ (such an estimator is referred to as a *model selection estimator*) cannot achieve an oracle inequality of form (1) with a remainder term of order smaller than $\sigma \sqrt{(\log M)/n}$; in other words, model selection is suboptimal for the purpose of competing with the best single model from a given family.

Instead of *model selection*, we can employ *model averaging* to derive oracle inequalities of form (1) that achieves the optimal regret in expectation (see the references in [4]). More recently, several work has produced optimal oracle inequalities for model averaging that not only hold in expectation but also in deviation [1], [3], [5]–[8]. In particular, the current work is closely related to the Q-aggregation estimator investigated in [1] which solves the optimal model averaging problem both in expectation and in deviation with a remainder term $\Delta(n, M, \sigma^2)$ of order $O(1/n)$; the authors also proposed a greedy algorithm GMA-0 for Q-aggregation which improves the Greedy Model Averaging (GMA) algorithm firstly proposed by [8]. Yet there are still two limitations of Q-aggregation: (1) Q-aggregation can be generalized for continuous candidates dictionary $\mathcal{H}$, but the greedy model averaging method GMA-0 can not be applied in such setting; (2) Q-aggregation can be regarded intuitively as

regression with variance penalty, but it lacks a good theoretical interpretation.

In this paper we introduce a novel method called *Bayesian Model Averaging with Exponentiated Least Squares Loss* (BMAX). We note that the previously studied exponential weighted model aggregation estimator EWMA (e.g., [4]) is the Bayes estimator under the least squares loss (posterior mean), which leads to optimal regret in expectation but is suboptimal in deviation. In contrast, the new BMAX model averaging estimator is essentially a Bayes estimator under an appropriately defined *exponentiated least squares loss*, and we will show that the Q-aggregation formulation (with Kullback-Leibler entropy) in [1] is essentially a dual representation of the newly introduced BMAX formulation, and it directly implies the optimality of the aggregate by BMAX. Computationally, the new model aggregation method BMAX can be approximately solved by a greedy model averaging algorithm and a gradient descend algorithm which is applicable to continuous dictionary. In summary, this paper establishes a natural Bayesian interpretation of Q-aggregation, and provides additional computational procedures that are applicable for the continuous dictionary setting. This relationship provides deeper understanding for model averaging procedures, and resolves the above mentioned limitations of the Q-aggregation scheme.

II. NOTATIONS

This section introduces some notations used in this paper. In the following, we denote by $Y = (y_1, \ldots, y_n)^\top$ the observation vector, $\eta = (\eta(x_1), \ldots, \eta(x_n))^\top$ the model output, and $\xi = (\xi_1, \ldots, \xi_n)^\top$ the noise vector. The underlying statistical model can be expressed as

$$Y = \eta + \xi, \tag{2}$$

with $\xi \sim N(0, \sigma^2 I_n)$. We also denote ℓ_2 norm as $\|Y\|_2 = (\sum_{i=1}^n y_i^2)^{1/2}$, and the inner product as $\langle \xi, f \rangle = \xi^\top f$.

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

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

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

Each $\lambda \in \Lambda^M$ yields a model averaging estimator as $f_\lambda = \sum_{j=1}^M \lambda_j f_j$; that is, using the vector notation $f_\lambda = (f_\lambda(x_1), \ldots, f_\lambda(x_n))^\top$ we have $f_\lambda = \sum_{j=1}^M \lambda_j f_j$. The Kullback-Leibler divergence for $\lambda, \pi \in \Lambda^M$ is defined as

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

and in the definition we use the convention $0 \cdot \log(0) = 0$.

For matrices $A, B \in \mathbb{R}^{n \times n}$, $A \geq B$ indicates that $A - B$ is positive semi-definite.

III. BAYESIAN MODEL AVERAGING WITH EXPONENTIATED LEAST SQUARES LOSS

The traditional Bayesian model averaging estimator is the exponential weighted model averaging estimator EWMA [4]

which optimizes the least squares loss. Although the estimator is optimal in expectation, it is suboptimal in deviation [1]. In this section we introduce a different Bayesian model averaging estimator called BMAX that optimizes an exponentiated least squares loss.

In order to introduce the BMAX estimator, we consider the following Bayesian framework, where we should be noted that the assumptions below are only used to derive BMAX, and these assumptions are not used in our theoretical analysis. Y is a normally distributed observation vector with mean $\mu = (\mu_1, \ldots, \mu_M)^\top$ and covariance matrix $\omega^2 I_n$:

$$Y | \mu \sim N(\mu, \omega^2 I_n), \tag{3}$$

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

$$\pi(\mu = f_j) = \pi_j. \tag{4}$$

In this setting, the posterior distribution of μ given Y is

$$
\begin{aligned}
p(\mu = f_j | Y) &= \frac{p(Y | \mu = f_j) p(\mu = f_j)}{\sum_{j=1}^M p(Y | \mu = f_j) p(\mu = f_j)} \\
&= \frac{\exp\left(-\frac{\|f_j - Y\|_2^2}{2\omega^2}\right) \pi_j}{\sum_{j=1}^M \exp\left(-\frac{\|f_j - Y\|_2^2}{2\omega^2}\right) \pi_j}.
\end{aligned}
$$

In the Bayesian decision theoretical framework considered in this paper, the quantity of interest is $\eta = \mathbb{E}Y$, and we consider a loss function $L(\psi, \mu)$ which we would like to minimize with respect to the posterior distribution. The corresponding Bayes estimator $\hat{\psi}$ minimizes the posterior expected loss from μ as follows:

$$\hat{\psi} = \underset{\psi \in \mathbb{R}^n}{\operatorname{argmin}} \mathbb{E}\left[L(\psi, \mu) | Y\right]. \tag{5}$$

It is worth pointing out that the above Bayesian framework is only used to obtain decision theoretically motivated model averaging estimators (Bayesian estimators have good theoretical properties such as admissibility, etc). In particular we do not assume that the model itself is correctly specified. That is, in this paper we allow misspecified models, where the parameters μ and ω^2 are not necessarily equal to the true mean η and the true variance σ^2 in (2), and η does not necessarily belong to the dictionary $\{f_1, \ldots, f_M\}$.

The Bayesian estimator of (5) depends on the underlying loss function $L(\cdot, \cdot)$. For example, under the standard least squares loss $L(\psi, \mu) = \|\psi - \mu\|_2^2$, the Bayes estimator is the posterior mean, which leads to the Exponential Weighted Model Aggregation (EWMA) estimator [4]:

$$\psi_{\ell_2}(\omega^2) = \frac{\sum_{j=1}^M \exp\left(-\frac{\|f_j - Y\|_2^2}{2\omega^2}\right) \pi_j f_j}{\sum_{j=1}^M \exp\left(-\frac{\|f_j - Y\|_2^2}{2\omega^2}\right) \pi_j}. \tag{6}$$

This estimator is optimal in expectation [9], [10], but suboptimal in deviation [1].

In this paper, we introduce the following *exponentiated least squares loss* motivated from the exponential moment technique

for proving large deviation tail bounds for sums of random variables:

$$L(\boldsymbol{\psi}, \boldsymbol{\mu}) = \exp\left(\frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi} - \boldsymbol{\mu}\|_2^2\right), \quad (7)$$

where the parameter $\nu \in (0, 1)$. It is easy to verify that the Bayes estimator defined by (5) with the loss function defined in (7) can be written as

$$\boldsymbol{\psi}_X(\omega^2, \nu) = \underset{\boldsymbol{\psi}\in\mathbb{R}^n}{\operatorname{argmin}} J(\boldsymbol{\psi}), \quad (8)$$

where

$$J(\boldsymbol{\psi}) = \sum_{j=1}^{M} \pi_j \exp\left(-\frac{1}{2\omega^2}\|\mathsf{f}_j - \boldsymbol{Y}\|_2^2 + \frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi} - \mathsf{f}_j\|_2^2\right). \quad (9)$$

The estimator $\boldsymbol{\psi}_X(\omega^2, \nu)$ will be referred to as the Bayesian model aggregation estimator with exponentiated least squares loss (BMAX).

To minimize $J(\boldsymbol{\psi})$, it is equivalent to minimize $\log J(\boldsymbol{\psi})$. Lemma 1 below shows the *strong convexity* and *smoothness* (under some conditions) of $\log J(\boldsymbol{\psi})$.

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

$$A_1 = \frac{1-\nu}{\omega^2}; \quad (10)$$

moreover, if the ℓ_2-norm of every f_j is bounded by a constant $L \in \mathbb{R}$:

$$\|\mathsf{f}_j\|_2 \leq L, \quad \forall j = 1, \dots, M, \quad (11)$$

we define A_2 and A_3 as

$$A_2 = \frac{1-\nu}{\omega^2} + \left(\frac{1-\nu}{\omega^2}\right)^2 L^2, \quad (12)$$

$$A_3 = \left(\frac{1-\nu}{\omega^2}\right) L^2 + \left(\frac{1-\nu}{\omega^2}\right)^2 L^4. \quad (13)$$

Lemma 1: For any $\boldsymbol{\psi} \in \mathbb{R}^n$, define the Hessian matrix of $\log J(\boldsymbol{\psi})$ as $\nabla^2 \log J(\boldsymbol{\psi}) = \frac{\partial^2 \log J(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\top}$, then we have

$$\nabla^2 \log J(\boldsymbol{\psi}) \geq A_1 \boldsymbol{I}_n. \quad (14)$$

If $\{\mathsf{f}_1, \dots, \mathsf{f}_M\}$ satisfies condition (11), then

$$\nabla^2 \log J(\boldsymbol{\psi}) \leq A_2 \boldsymbol{I}_n, \quad (15)$$

where A_1 and A_2 are defined in (10) and (12).

IV. DUAL REPRESENTATION AND Q-AGGREGATION

In this section, we will show that the *Q-aggregation* scheme of [1] with the standard Kullback-Leibler entropy solves a dual representation of the BMAX formulation defined by (8).

Given $\boldsymbol{Y}$ and $\{\mathsf{f}_1, \dots, \mathsf{f}_M\}$, Q-aggregation $\mathsf{f}_{\boldsymbol{\lambda}^Q}$ is defined as follows:

$$\mathsf{f}_{\boldsymbol{\lambda}^Q} = \sum_{j=1}^{M} \lambda_j^Q \mathsf{f}_j, \quad (16)$$

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

$$\boldsymbol{\lambda}^Q \in \underset{\boldsymbol{\lambda}\in\Lambda^M}{\operatorname{argmin}} Q(\boldsymbol{\lambda}), \quad (17)$$

$$Q(\boldsymbol{\lambda}) = \|\mathsf{f}_{\boldsymbol{\lambda}} - \boldsymbol{Y}\|_2^2 + \nu \sum_{j=1}^{M} \lambda_j \|\mathsf{f}_j - \mathsf{f}_{\boldsymbol{\lambda}}\|_2^2 + 2\omega^2 \mathcal{K}_\rho(\boldsymbol{\lambda}, \boldsymbol{\pi}), \quad (18)$$

for some $\nu \in (0, 1)$, where the ρ-entropy $\mathcal{K}_\rho(\boldsymbol{\lambda}, \boldsymbol{\pi})$ is defined as

$$\mathcal{K}_\rho(\boldsymbol{\lambda}, \boldsymbol{\pi}) = \sum_{j=1}^{M} \lambda_j \log\left(\frac{\rho(\lambda_j)}{\pi_j}\right), \quad (19)$$

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

$$\rho(t) \geq t, \quad t \log \rho(t) \text{ is convex.} \quad (20)$$

When $\rho(t) = t$, $\mathcal{K}_\rho(\boldsymbol{\lambda}, \boldsymbol{\pi})$ becomes $\mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi})$, i.e., 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, and in particular the penalty $\mathcal{K}_\rho(\boldsymbol{\lambda}, \boldsymbol{\pi})$ in (18) becomes a constant when $\boldsymbol{\pi}$ is a flat prior.

Now, to illustrate duality, we shall first introduce a function $T : \mathbb{R}^n \to \mathbb{R}$ as

$$T(\boldsymbol{h}) = -\frac{\nu}{1-\nu}\|\boldsymbol{h} - \boldsymbol{Y}\|_2^2 \\ - 2\omega^2 \log\left(\sum_{j=1}^{M} \pi_j \exp\left(-\frac{\nu}{2\omega^2}\|\mathsf{f}_j - \boldsymbol{h}\|_2^2\right)\right), \quad (21)$$

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

$$\hat{\boldsymbol{h}} = \underset{\boldsymbol{h}\in\mathbb{R}^n}{\operatorname{argmax}} T(\boldsymbol{h}). \quad (22)$$

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

$$S(\boldsymbol{\lambda}, \boldsymbol{h}) = -\frac{\nu}{1-\nu}\|\boldsymbol{h} - \boldsymbol{Y}\|_2^2 + \nu \sum_{j=1}^{M} \lambda_j \|\mathsf{f}_j - \boldsymbol{h}\|_2^2 + 2\omega^2 \mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi}). \quad (23)$$

It is not difficult to verify that for $\nu \in (0, 1)$, $S(\boldsymbol{\lambda}, \boldsymbol{h})$ is convex in $\boldsymbol{\lambda}$ and concave in $\boldsymbol{h}$. The following duality lemma states the relationship between $\hat{\boldsymbol{h}}$ and $\mathsf{f}_{\boldsymbol{\lambda}^Q}$.

Lemma 2: When $\rho(t) = t$, we have the following result

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

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

where the equality is achieved at $(\boldsymbol{\lambda}^Q, \hat{\boldsymbol{h}})$. Moreover, we have

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

where A and B are two hyper-surfaces in $\Lambda^M \times \mathbb{R}^n$ defined as

$$A = \left\{ (\boldsymbol{\lambda}, \boldsymbol{h}) \in \Lambda^M \times \mathbb{R}^n : \boldsymbol{h} = \frac{1}{\nu} Y - \frac{1-\nu}{\nu} \mathsf{f}_{\boldsymbol{\lambda}} \right\},$$

$$B = \Big\{ (\boldsymbol{\lambda}, \boldsymbol{h}) \in \Lambda^M \times \mathbb{R}^n :$$

$$\lambda_j = \frac{\exp\left(-\frac{\nu}{2\omega^2} \|\mathsf{f}_j - \boldsymbol{h}\|_2^2\right) \pi_j}{\sum_{i=1}^{M} \exp\left(-\frac{\nu}{2\omega^2} \|\mathsf{f}_i - \boldsymbol{h}\|_2^2\right) \pi_i} \Big\}. \qquad (24)$$

Lemma 2 states that, $(\boldsymbol{\lambda}^Q, \hat{\boldsymbol{h}})$ is the only joint of hyper-surfaces A and B, and the only saddle point of function $S(\boldsymbol{\lambda}, \boldsymbol{h})$ over space $\Lambda^M \times \mathbb{R}^n$.

With $T(\boldsymbol{h})$ defined as in (21), we can employ the transformation $\boldsymbol{h} = \frac{1}{\nu} Y - \frac{1-\nu}{\nu} \boldsymbol{\psi}$, and it is easy to verify that

$$T(\boldsymbol{h}) = -2\omega^2 \log\left(J(\boldsymbol{\psi})\right), \qquad (25)$$

where $J(\boldsymbol{\psi})$ is defined in (9). Since $J(\boldsymbol{\psi})$ is strictly convex, $T(\boldsymbol{h})$ is strictly concave and $\hat{\boldsymbol{h}}$ is unique. It follows that maximizing $T(\boldsymbol{h})$ is equivalent to minimizing $J(\boldsymbol{\psi})$, and thus

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

We can combine this representation with

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

from Lemma 2 to obtain $\boldsymbol{\psi}_X(\omega^2, \nu) = \mathsf{f}_{\boldsymbol{\lambda}Q}$. Therefore, we directly have the following relationship.

Theorem 1: When $\rho(t) = t$,

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

where $\boldsymbol{\psi}_X(\omega^2, \nu)$ is defined by (8) and (9), and $\mathsf{f}_{\boldsymbol{\lambda}Q}$ is defined by (16),(17) and (18).

Theorem 1 states that, when $\rho(t) = t$, $\mathcal{K}_\rho(\boldsymbol{\lambda}, \boldsymbol{\pi})$ becomes the Kullback-Leibler entropy, and Q-aggregation with the Kullback-Leibler entropy leads to an estimator $\mathsf{f}_{\boldsymbol{\lambda}Q}$ that is essentially a dual representation of the BMAX estimator $\boldsymbol{\psi}_X(\omega^2, \nu)$. It follows that, $\boldsymbol{\psi}_X(\omega^2, \nu)$ should share the same optimality (both in expectation and deviation) as $\mathsf{f}_{\boldsymbol{\lambda}Q}$ in solving the model averaging problem (optimality of $\mathsf{f}_{\boldsymbol{\lambda}Q}$ is shown in [1, Th. 3.1] with more general $\mathcal{K}_\rho(\boldsymbol{\lambda}, \boldsymbol{\pi})$, where $\rho(t)$ only needs to satisfy the condition (20)).

However, unlike the primal objective function $J(\boldsymbol{\psi})$ which is defined on $\mathbb{R}^n$, the dual objective function $Q(\boldsymbol{\lambda})$ is defined on $\mathbb{R}^M$. When M is large or infinity, the optimization of $Q(\boldsymbol{\lambda})$ is non-trivial. Although greedy algorithms are proposed in [1], they cannot handle the standard KL-divergence; instead, they can only work with the linear entropy where $\rho(t) = 1$; it gives a larger penalty than the standard KL-divergence (and thus worse resulting oracle inequality), and it cannot be generalized to handle continuous dictionaries (because in such case the linear entropy with $\rho(t) = 1$ will always be $+\infty$). Therefore, the numerical greedy procedures of [1] converge to a solution with a worse oracle bound than that of the solution for the primal formulation considered in this paper.

The following two corollaries (Corollary 1 and Corollary 2) are listed for illustration convenience. They are directly derived from Theorem 1 and the optimality of $\mathsf{f}_{\boldsymbol{\lambda}Q}$ in [1, Th. 3.1], so we omit their proofs.

Corollary 1: Assume that $\nu \in (0, 1)$ and $\omega^2 \geq \frac{\sigma^2}{\min(\nu, 1-\nu)}$. For any $\boldsymbol{\lambda} \in \Lambda^M$, the following oracle inequality holds

$$\|\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta}\|_2^2 \leq \nu \sum_{j=1}^{M} \lambda_j \|\mathsf{f}_j - \boldsymbol{\eta}\|_2^2 + (1-\nu) \|\mathsf{f}_{\boldsymbol{\lambda}} - \boldsymbol{\eta}\|_2^2$$
$$+ 2\omega^2 \mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi}\delta), \qquad (26)$$

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

$$\mathbb{E}\|\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta}\|_2^2$$
$$\leq \nu \sum_{j=1}^{M} \lambda_j \|\mathsf{f}_j - \boldsymbol{\eta}\|_2^2 + (1-\nu) \|\mathsf{f}_{\boldsymbol{\lambda}} - \boldsymbol{\eta}\|_2^2 + 2\omega^2 \mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi})$$
$$\leq \|\mathsf{f}_{\boldsymbol{\lambda}} - \boldsymbol{\eta}\|_2^2 + \nu \sum_{j=1}^{M} \lambda_j \|\mathsf{f}_j - \mathsf{f}_{\boldsymbol{\lambda}}\|_2^2 + 2\omega^2 \mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi}). \qquad (27)$$

Corollary 1 implies that when the term $\nu \sum_{j=1}^{M} \lambda_j \|\mathsf{f}_j - \mathsf{f}_{\boldsymbol{\lambda}}\|_2^2$ and the divergence term $K(\boldsymbol{\lambda}, \boldsymbol{\pi})$ are small, $\boldsymbol{\psi}_X(\omega^2, \nu)$ can compete with an arbitrary $\mathsf{f}_{\boldsymbol{\lambda}}$ in the convex hull with any $\boldsymbol{\lambda} \in \Lambda^M$. Actually, we can obtain an oracle inequality that competes with the best single model, which is the situation that $\boldsymbol{\lambda}$ is at a vertex of the simplex Λ^M:

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

$$\|\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta}\|_2^2 \leq \min_{j \in 1,\dots,M} \left\{ \|\mathsf{f}_j - \boldsymbol{\eta}\|_2^2 + 2\omega^2 \log\left(\frac{1}{\pi_j \delta}\right) \right\}, \qquad (28)$$

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

$$\mathbb{E}\|\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta}\|_2^2 \leq \min_{j \in 1,\dots,M} \left\{ \|\mathsf{f}_j - \boldsymbol{\eta}\|_2^2 + 2\omega^2 \log\left(\frac{1}{\pi_j}\right) \right\}. \qquad (29)$$

It is also worth pointing out that the condition $\omega^2 \geq \frac{\sigma^2}{\min(\nu, 1-\nu)}$ implies that ω^2 is at least greater than $2\sigma^2$ (when $\nu = 1/2$), and intuitively this inflation of noise allows the Bayes estimator to handle misspecification of the true mean $\boldsymbol{\eta}$, which is not necessarily included in the dictionary $\mathcal{H}$. Similar observations were found by [11].

Finally we note that in the Bayesian framework stated in this section, when we change the underlying loss function $L(\boldsymbol{\psi}, \boldsymbol{\mu})$ from the standard least squares loss to the exponentiated least squares loss (7), Bayes estimator changes from EWMA which is optimal only in expectation to BMAX which is optimal both in expectation and in deviation. The difference is that the least squares loss only controls the bias, while the exponentiated least squares loss controls both bias and variance (as well as higher order moments) simultaneously. This can be seen by using Taylor expansion

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

Since deviation bounds require us to control high order moments, the exponentiated least squares loss is naturally suited for obtaining deviation bounds.

It is also natural to extend Corollary 1 from discrete candidates dictionary $\mathcal{H} = \{f_1, \ldots, f_M\}$ to infinite dictionary ($M = \infty$) as well as continuous dictionary. For example, given a matrix $X \in \mathbb{R}^{n \times d}$, we may consider a continuous dictionary parameterized by vector w as $\mathcal{H}_\Omega = \{f_w : f_w = Xw \in \mathbb{R}^n\}$ where $\Omega = \{w : w \in \mathbb{R}^d\}$. Then, we have the following result.

Corollary 3: Assume $v \in (0, 1)$ and for any distribution $p(w)$ over the parameter space Ω, the divergence term $\mathcal{K}(p, \pi)$ is finite for given prior $\pi(w)$. Then, if $\omega^2 \geq \frac{\sigma^2}{\min(v, 1-v)}$, we have the oracle inequality

$$\|\hat{\psi} - \eta\|_2^2 \leq v \int_\Omega \|f_w - \eta\|_2^2 \, p(w) dw$$
$$+ (1 - v) \left\| \int_\Omega f_w p(w) dw - \eta \right\|_2^2 + 2\omega^2 \mathcal{K}(p, \pi \delta) \quad (30)$$

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

$$\mathbb{E} \|\hat{\psi} - \eta\|_2^2$$
$$\leq v \int_\Omega \|f_w - \eta\|_2^2 \, p(w) dw + (1 - v) \left\| \int_\Omega f_w p(w) dw - \eta \right\|_2^2$$
$$+ 2\omega^2 \mathcal{K}(p, \pi)$$
$$\leq \left\| \int_\Omega f_w p(w) dw - \eta \right\|_2^2 + v \int_\Omega \left\| f_w - \int_\Omega f_w p(w) dw \right\|_2^2$$
$$\times p(w) dw + 2\omega^2 \mathcal{K}(p, \pi), \quad (31)$$

where

$$\hat{\psi} = \underset{\psi \in \mathbb{R}^n}{\operatorname{argmin}} \int_\Omega \exp\left(-\frac{1}{2\omega^2} \|f_w - Y\|_2^2 \right.$$
$$\left. + \frac{1 - v}{2\omega^2} \|f_w - \psi\|_2^2 \right) \pi(w) dw.$$

When the distribution p on $w \in \mathbb{R}^d$ is concentrated around a single model, for example, p is an uniform distribution on a small ball $\{w : \|w - w_0\|_2 \leq r\}$ for some small $r > 0$, we have $\int_\Omega f_w p(w) dw = f_{w_0}$ and the term $\frac{1}{n} \int_\Omega \|f_w - f_{w_0}\|_2^2 \, p(w) dw \leq \frac{r^2}{n} \|X\|_F^2$ is small if $\frac{1}{n} \|X\|_F^2$ is bounded ($\|\cdot\|_F$ is the matrix Frobenius norm). Although a small r would lead to a larger divergence term $\mathcal{K}(p, \pi)$ and there is a tradeoff between $\int_\Omega \|f_w - f_{w_0}\|_2^2 \, p(w) dw$ and $\mathcal{K}(p, \pi)$, by dividing the number of observations n on both sides of Eq. (31), the term $\frac{1}{n} \mathcal{K}(p, \pi)$ could also be small as long as n is sufficiently large. Therefore, Corollary 3 implies that $\hat{\psi}$ can compete with a single model f_{w_0}, similar to the case of discrete dictionary.

V. ALGORITHMS TO SOLVE BMAX

In this section, we propose two algorithms, the Greedy Model Averaging (GMA-BMAX) algorithm and the Gradient Descent (GD-BMAX) algorithm, to solve BMAX. 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 geometric rate of

---

Algorithm 1 Greedy Model Averaging Algorithm (GMA-BMAX)

Input: Noisy observation Y, dictionary $\mathcal{H} = \{f_1, \ldots, f_M\}$, prior $\pi \in \Lambda^M$, parameters v, ω.
Output: Aggregate estimator $\psi^{(k)}$.
 Let $\psi^{(0)} = 0$;
  for $k = 1, 2, \ldots$ **do**
    Set $\alpha_k = \frac{2}{k+1}$;
    $J^{(k)} = \operatorname{argmin}_j \log J(\psi^{(k-1)} + \alpha_k(f_j - \psi^{(k-1)}))$;
    $\psi^{(k)} = \psi^{(k-1)} + \alpha_k(f_{J^{(k)}} - \psi^{(k-1)})$;
  end for

$O(q^k)$ for some $q \in (0, 1)$. Oracle inequalities will be derived for the estimators in the iterations for both the algorithms.

Strong convexity of $\log J(\psi)$ directly implies that the minimizer $\psi_X(\omega^2, v)$ is unique. Moreover, it implies the following proposition which shows that an estimator that approximately minimizes $\log J(\psi)$ satisfies an oracle inequality slightly worse than that of $\psi_X(\omega^2, v)$ in Corollary 1. This result suggests that we can employ appropriate numerical procedures to approximately solve (8), and Corollary 1 implies an oracle inequality for such approximate solutions.

Proposition 1: Let $\hat{\psi}$ be an ϵ-approximate minimizer of $\log J(\psi)$ for some $\epsilon > 0$ that $\log J(\hat{\psi}) \leq \min_\psi \log J(\psi) + \epsilon$. Then, we have

$$\|\hat{\psi} - \eta\|_2^2 \leq \|\psi_X(\omega^2, v) - \eta\|_2^2 + 2\sqrt{2\epsilon/A_1} \|\psi_X(\omega^2, v)$$
$$- \eta\|_2 + \frac{2\epsilon}{A_1}.$$

Next, we present the numerical algorithms to solve the BMAX problem.

A. Greedy Model Averaging Algorithm (GMA-BMAX)

The GMA-BMAX algorithm given in Algorithm 1 is a greedy algorithm that adds 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 for model averaging have appeared in [1] and [8].

The following proposition follows from the standard analysis in [12]–[14]. It shows that the estimator $\psi^{(k)}$ from Algorithm 1 converges to $\psi_X(\omega^2, v)$.

Proposition 2: For $\psi^{(k)}$ as defined in Algorithm 1 (GMA-BMAX), if $\{f_1, \ldots, f_M\}$ satisfies condition (11), then

$$\log J(\psi^{(k)}) \leq \log J(\psi_X(\omega^2, v)) + \frac{8A_3}{k + 3}. \quad (32)$$

Proposition 2 states that, after running the GMA-BMAX algorithm for k steps to obtain $\psi^{(k)}$, the corresponding objective value $\log J(\psi^{(k)})$ converges to the optimal objective value $\log J(\psi_X(\omega^2, v))$ at a rate $O(1/k)$. Combining this result with Proposition 1, we obtain the following oracle inequality, which shows that the regret of the estimator $\psi^{(k)}$ after running k steps of GMA-BMAX converges to that of $\psi_X(\omega^2, v)$ in Corollary 1 at a rate $O(1/\sqrt{k})$.

Proposition 3: Assume $\nu \in (0, 1)$ and if $\omega^2 \geq \frac{\sigma^2}{\min(\nu, 1-\nu)}$. Consider $\boldsymbol{\psi}^{(k)}$ as in Algorithm 1 (GMA-BMAX). For any $\boldsymbol{\lambda} \in \Lambda^M$, the following oracle inequality holds

$$\|\boldsymbol{\psi}^{(k)} - \boldsymbol{\eta}\|_2^2 \leq \nu \sum_{j=1}^M \lambda_j \|\mathsf{f}_j - \boldsymbol{\eta}\|_2^2 + (1 - \nu) \|\mathsf{f}_{\boldsymbol{\lambda}} - \boldsymbol{\eta}\|_2^2$$

$$+ 2\omega^2 \mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi}\delta) + 2\sqrt{\frac{16A_3}{A_1(k+3)}} \|\boldsymbol{\psi}_X(\omega^2, \nu)$$

$$- \boldsymbol{\eta}\|_2 + \frac{16A_3}{A_1(k+3)} \quad (33)$$

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

$$\mathbb{E}\|\boldsymbol{\psi}^{(k)} - \boldsymbol{\eta}\|_2^2 \leq \nu \sum_{j=1}^M \lambda_j \|\mathsf{f}_j - \boldsymbol{\eta}\|_2^2 + (1 - \nu) \|\mathsf{f}_{\boldsymbol{\lambda}} - \boldsymbol{\eta}\|_2^2$$

$$+ 2\omega^2 \mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi}) + 2\sqrt{\frac{16A_3}{A_1(k+3)}} \mathbb{E}\|\boldsymbol{\psi}_X(\omega^2, \nu)$$

$$- \boldsymbol{\eta}\|_2 + \frac{16A_3}{A_1(k+3)}. \quad (34)$$

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

$$\|\boldsymbol{\psi}^{(k)} - \boldsymbol{\eta}\|_2^2 \leq \|\mathsf{f}_j - \boldsymbol{\eta}\|_2^2 + 2\omega^2 \log\left(\frac{1}{\pi_j \delta}\right) + O(1/\sqrt{k})$$

with probability at least $1 - \delta$, and

$$\mathbb{E}\|\boldsymbol{\psi}^{(k)} - \boldsymbol{\eta}\|_2^2 \leq \|\mathsf{f}_j - \boldsymbol{\eta}\|_2^2 + 2\omega^2 \log\left(\frac{1}{\pi_j}\right) + O(1/\sqrt{k}).$$

When $k \to \infty$, $\boldsymbol{\psi}^{(k)}$ achieves the optimal deviation bound. However, it does not imply optimal deviation bound for $\boldsymbol{\psi}^{(k)}$ with small k, while the greedy algorithms described in [8] (GMA) and [1] (GMA-0) achieve optimal deviation bounds for small k when $k \geq 2$. The advantage of GMA-BMAX is that the resulting estimator $\psi^{(k)}$ competes with any $\mathsf{f}_{\boldsymbol{\lambda}}$ with $\boldsymbol{\lambda} \in \Lambda^M$ under the KL entropy, and such a result can be applied even for infinity dictionaries containing functions indexed by continuous parameters, as long as the KL divergence $\mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi})$ is well-defined (see relevant discussions in Section IV). On the other hand, the greedy estimators of [1] for the Q-aggregation scheme can only deal with an upper bound of KL divergence referred to as linear entropy (see Section IV) that is not well-defined for continuous dictionaries. This means that GMA-BMAX is more generally applicable than the corresponding greedy algorithm GMA-0 in [1].

B. Gradient Descent Algorithm (GD-BMAX)

An alternative way solving the BMAX is to use the gradient descend method. The GD-BMAX algorithm is shown in Algorithm 2. The gradient is

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

Algorithm 2 Gradient Descent Algorithm (GD-BMAX)

Input: Noisy observation Y, dictionary $\mathcal{H} = \{f_1, \ldots, f_M\}$, prior $\boldsymbol{\pi} \in \Lambda^M$, parameters ν, ω^2.
Output: Aggregate estimator $\boldsymbol{\psi}^{(k)}$.
Let $\boldsymbol{\psi}^{(0)} = 0$.
  for $k = 1, 2, \ldots$ **do**
    Choose step size $t_k \in (0, 2/A_2)$ for $k > 0$;
    Let

$$\mathsf{f}_{\boldsymbol{\lambda}^{(k-1)}} = \sum_{j=1}^M \lambda_j^{(k-1)} \mathsf{f}_j,$$

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

$$\lambda_j^{(k-1)} \propto \pi_j \exp\left(-\frac{1}{2\omega^2}\|\mathsf{f}_j - Y\|_2^2 + \frac{1-\nu}{2\omega^2}\|\boldsymbol{\psi}^{(k-1)} - \mathsf{f}_j\|_2^2\right) \quad (35)$$

$$\boldsymbol{\psi}^{(k)} = (1 - t_k \frac{1-\nu}{\omega^2})\boldsymbol{\psi}^{(k-1)} + t_k \frac{1-\nu}{\omega^2} \mathsf{f}_{\boldsymbol{\lambda}^{(k-1)}};$$
  end for

---

where $\boldsymbol{\lambda}^{(k-1)} \in \Lambda^M$ is defined as (35). In the k-th step, the update operation is

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

Therefore, Algorithm 2 is a gradient decent algorithm with step size t_k. The following proposition shows the convergence of the GD-BMAX algorithm.

Proposition 4: For $\boldsymbol{\psi}^{(k)}$ as defined in Algorithm 2, if we choose a fixed step size $t_k = s \in (0, 2/A_2)$ for $k > 0$ and $\{f_1, \ldots, f_M\}$ satisfy condition (11), then

$$\log J(\boldsymbol{\psi}^{(k)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))$$
$$\leq [1 - 2A_1(s - (A_2/2)s^2)]^k \Big(\log J(\boldsymbol{\psi}^{(0)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))\Big). \quad (36)$$

Remark 1: We may choose $t_k = s = 1/A_2$ to minimize the righthand side of (36) such that

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

Proposition 4 shows that the GD-BMAX algorithm converges at a geometric rate of $O(q^k)$ with $q = 1 - 2A_1(s - (A_2/2)s^2)$. Moreover, we have the following oracle inequality, which shows that the regret of the estimator $\boldsymbol{\psi}^{(k)}$ after running k steps of GD-BMAX converges to that of $\boldsymbol{\psi}_X(\omega^2, \nu)$ in Corollary 1 at a rate of $O(q^k)$.

Proposition 5: Assume $\nu \in (0, 1)$ and if $\omega^2 \geq \frac{\sigma^2}{\min(\nu, 1-\nu)}$, we have oracle inequality for any $\boldsymbol{\lambda} \in \Lambda^M$,

$$\|\boldsymbol{\psi}^{(k)} - \boldsymbol{\eta}\|_2^2$$
$$\leq \nu \sum_{j=1}^M \lambda_j \|\mathsf{f}_j - \boldsymbol{\eta}\|_2^2 + (1 - \nu) \|\mathsf{f}_{\boldsymbol{\lambda}} - \boldsymbol{\eta}\|_2^2$$

$$+ 2\omega^2 \mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi}) + 2\omega^2 \log(1/\delta)$$
$$+ 2\sqrt{L^2[1 - 2A_1(s - (A_2/2)s^2)]^k} \|\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta}\|_2$$
$$+ L^2[1 - 2A_1(s - (A_2/2)s^2)]^k, \tag{37}$$

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

$$\mathbb{E}\|\boldsymbol{\psi}^{(k)} - \boldsymbol{\eta}\|_2^2$$
$$\leq \nu \sum_{j=1}^{M} \lambda_j \|\mathsf{f}_j - \boldsymbol{\eta}\|_2^2 + (1 - \nu) \|\mathsf{f}_{\boldsymbol{\lambda}} - \boldsymbol{\eta}\|_2^2$$
$$+ 2\omega^2 \mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi})$$
$$+ 2\sqrt{L^2[1 - 2A_1(s - (A_2/2)s^2)]^k} \mathbb{E}\|\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta}\|_2$$
$$+ L^2[1 - 2A_1(s - (A_2/2)s^2)]^k. \tag{38}$$

From Proposition 5, if $\omega^2 \geq \frac{\sigma^2}{\min(\nu, 1 - \nu)}$, for any $j = 1, \ldots, M$ we have

$$\|\boldsymbol{\psi}^{(k)} - \boldsymbol{\eta}\|_2^2 \leq \|\mathsf{f}_j - \boldsymbol{\eta}\|_2^2 + 2\omega^2 \log\left(\frac{1}{\pi_j \delta}\right) + O(q^k),$$

with probability at least $1 - \delta$ and

$$\mathbb{E}\|\boldsymbol{\psi}^{(k)} - \boldsymbol{\eta}\|_2^2 \leq \|\mathsf{f}_j - \boldsymbol{\eta}\|_2^2 + 2\omega^2 \log\left(\frac{1}{\pi_j}\right) + O(q^k),$$

for some constant $q = 1 - 2A_1(s - (A_2/2)s^2) \in (0, 1)$.

Though the GD-BMAX algorithm does not give sparse output as the GMA-BMAX algorithm does, it has a faster geometric convergence compared to GMA-BMAX. Similar to Proposition 3, the results in Proposition 5 do not imply optimal deviation bounds of $\boldsymbol{\psi}^{(k)}$ for small k ($k < \infty$).

The GD-BMAX algorithm can be naturally applied to continuous dictionary case. When the dictionary is continuous or M is large, direct calculation of $\boldsymbol{\lambda}^{(k-1)} \in \Lambda^M$ in (35) is impractical and we can use the Markov Chain Monte Carlo (MCMC) sampling method [15] to approximate $\boldsymbol{\lambda}^{(k-1)}$ for the k-th iteration in Algorithm 2. A Metropolis-Hastings (MH) algorithm is given in Algorithm 3. The MH algorithm approximates $\mathsf{f}_{\boldsymbol{\lambda}^{(k-1)}}$ with $\boldsymbol{u}_T^{(k-1)}$, so the resulting estimator $\boldsymbol{\psi}^{(k)}$ at the k-th step in Algorithm 2 will have perturbations. Below we provide a result showing how the perturbations from approximating $\mathsf{f}_{\boldsymbol{\lambda}^{(k-1)}}$ would influence the convergence of $\log J(\boldsymbol{\psi}^{(k)})$ to $\log J(\boldsymbol{\psi}_X(\omega^2, \nu))$.

Proposition 6: Given $\boldsymbol{Y} \in \mathbb{R}^n$, for all $k > 0$, we assume that $\boldsymbol{u}_T^{(k-1)}$ in Algorithm 3 satisfies

$$\mathbb{E}[\boldsymbol{u}_T^{(k-1)} | \boldsymbol{\psi}^{(k-1)}] = \mathsf{f}_{\boldsymbol{\lambda}^{(k-1)}}, \tag{39}$$
$$\|COV[\boldsymbol{u}_T^{(k-1)} | \boldsymbol{\psi}^{(k-1)}]\|_{op} \leq s^2, \tag{40}$$

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

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

A variant of the GD-BMAX algorithm with continuous dictionary notation is also provided in Algorithm 4, in which we assume the dictionary is parameterized by w as $\mathcal{H}_\Omega = \{\mathsf{f}(w) : \mathsf{f}(w) \in \mathbb{R}^n\}$ and $\Omega = \{w : w \in \mathbb{R}^d\}$.

Algorithm 3 Metropolis-Hastings (MH) Sampler for Estimating $\mathsf{f}_{\boldsymbol{\lambda}^{(k-1)}}$ at the k-th Step of Algorithm 2

Input: Noisy observation $\boldsymbol{Y}$, dictionary $\mathcal{H} = \{\mathsf{f}_1, \ldots, \mathsf{f}_M\}$, prior $\boldsymbol{\pi} \in \Lambda^M$, parameters ν, ω^2, $(k-1)$-th step estimator $\boldsymbol{\psi}^{(k-1)}$.

Output: $\boldsymbol{u}_T^{(k-1)}$ as estimator of $\mathsf{f}_{\boldsymbol{\lambda}^{(k-1)}} = \sum_{j=1}^M \lambda_j^{(k-1)} \mathsf{f}_j$.

Initialize $j(0) = 0$;

for $t = 1, \cdots, T_0 + T$ **do**

Generate $\tilde{j} \sim q(\cdot | j^{(t-1)})$ (e.g., q can be chosen as a Gaussian distribution with mean $\mathsf{f}_{j^{(t-1)}}$);

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

$$\theta(j) = \pi_j \exp\left(-\frac{1}{2\omega^2}\|\mathsf{f}_j - \boldsymbol{Y}\|_2^2\right.$$
$$\left. + \frac{1 - \nu}{2\omega^2}\|\boldsymbol{\psi}^{(k-1)} - \mathsf{f}_j\|_2^2\right);$$

Generate a random variable $u \in [0, 1]$ and let

$$j^{(t)} = \begin{cases} \tilde{j}, & \text{if } u \leq \rho(j^{(t-1)}, \tilde{j}); \\ j^{(t-1)}, & \text{otherwise}; \end{cases}$$

end for

Calculate

$$\boldsymbol{u}_T^{(k-1)} = \frac{1}{T} \sum_{t=T_0+1}^{T_0+T} \mathsf{f}_{j^{(t)}};$$

VI. EXPERIMENTS

Although the contribution of this work is mainly theoretical, we include some simulations to illustrate the performance of the GMA-BMAX algorithm and GD-BMAX algorithm proposed for the BMAX method. We focus on the average performance of different algorithms and configurations.

The simulations will focus on discrete dictionary in order to compare the BMAX method with existing algorithms while the BMAX method can deal with continuous dictionary. Set $n = 50$ and $M = 500$. We identify a function f with a vector $(f(x_1), \ldots, f(x_n))^\top \in \mathbb{R}^n$. Let $\boldsymbol{I}_n$ denote the identity matrix of $\mathbb{R}^n$ and let $\Theta \sim \mathcal{N}(0, \boldsymbol{I}_n)$ be a random vector, and define $\{\mathsf{f}_1, \ldots, \mathsf{f}_M\}$ as

$$\begin{cases} \mathsf{f}_j = \Theta + s \cdot \boldsymbol{\zeta}_j & \text{for } 1 \leq j \leq M_1, \\ \mathsf{f}_j = \boldsymbol{\zeta}_j & \text{for } M_1 < j \leq M, \end{cases} \tag{41}$$

where $\boldsymbol{\zeta}_j \sim \mathcal{N}(0, \boldsymbol{I}_n)(j = 1, \ldots, M)$ are independent random vectors.

Let $\Delta \sim \mathcal{N}(0, \boldsymbol{I}_n)$ be a random vector. The regression function is defined by $\boldsymbol{\eta} = \mathsf{f}_1 + 0.5\Delta$. Note that typically 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 independent of $\{\mathsf{f}_1, \ldots, \mathsf{f}_M\}$ and $\sigma = 2$.

Algorithm 4 Gradient Descent Algorithm With Continuous Dictionary

---

Input: Noisy observation Y, continuous dictionary $\mathcal{H}_\Omega$, prior $\boldsymbol{\pi}_\Omega$, parameters ν, ω^2.

Output: Aggregate estimator $\boldsymbol{\psi}^{(k)}$.

  Let $\boldsymbol{\psi}^{(0)} = 0$.

  for $k = 1, 2, \ldots$ **do**

    Choose step size $t_k \in (0, 2/A_2)$ for $k > 0$;

    Initialize $w^{(0)}$;

    for $t = 1, \cdots, T_0 + T$ **do**

      Generate $\tilde{w} \sim q(\cdot | w^{(t-1)})$;

      Compute

$$\gamma(w^{(t-1)}, \tilde{w}) = \min\left(\frac{q(w^{(t-1)} | \tilde{w})\theta(\tilde{w})}{q(\tilde{w} | w^{(t-1)})\theta(w^{(t-1)})}, 1 \right),$$

      where

$$\theta(w) = \pi(w) \exp\left(-\frac{1}{2\omega^2} \|\mathsf{f}(w) - Y\|_2^2 \right.$$
$$\left. + \frac{1-\nu}{2\omega^2} \|\boldsymbol{\psi}^{(k-1)} - \mathsf{f}(w)\|_2^2 \right);$$

      Generate a random variable $u \in [0, 1]$ and let

$$w^{(t)} = \begin{cases} \tilde{w}, & \text{if } u \leq \gamma(w^{(t-1)}, \tilde{w}); \\ w^{(t-1)}, & \text{otherwise}; \end{cases}$$

    end for

    Calculate

$$\boldsymbol{u}_T^{(k-1)} = \frac{1}{T} \sum_{t=T_0+1}^{T_0+T} \mathsf{f}(w^{(t)});$$

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

  end for

We define the oracle model (OM) f_{k^*}, where $k^* = \text{argmin}_j \text{MSE}(f_j)$. 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:

$$R(\hat{\eta}) = \text{MSE}(\hat{\eta}) - \text{MSE}(f_{k^*}). \quad (42)$$

Since the target is $\boldsymbol{\eta} = \mathsf{f}_1 + 0.5\Delta$, and f_1 and Δ are random Gaussian vectors, the oracle model is likely f_1 (but it may not be 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 100 replications.

One method compared is the STAR algorithm of [5], which is optimal both in expectation and in deviation under the uniform prior. Mathematically, suppose f_{k_1} is the empirical

risk minimizer among functions in $\mathcal{H}$, where

$$k_1 = \underset{j}{\text{argmin}} \, \widehat{\text{MSE}}(f_j), \quad (43)$$

the STAR estimator f^* is defined as

$$f^* = (1 - \alpha^*)f_{k_1} + \alpha^* f_{k_2}, \quad (44)$$

where

$$(\alpha^*, k_2) = \underset{\alpha, j}{\text{argmin}} \, \widehat{\text{MSE}}\big((1-\alpha)f_{k_1} + \alpha f_j\big), \quad \alpha \in (0, 1). \quad (45)$$

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

$$\boldsymbol{\lambda}^{\text{PROJ}} \in \underset{\boldsymbol{\lambda} \in \Lambda^M}{\text{argmin}} \, \widehat{\text{MSE}}(\mathsf{f}_{\boldsymbol{\lambda}}), \quad (46)$$

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 Y onto the convex hull of the f_j's.

We compare the exponential weighted model averaging method denoted as EWMA. Q-aggregation with linear entropy (when $\rho(t) = 1$) is also compared and will be solved by GMA-0 from [1] (see Algorithm 5 below).

Algorithm 5 GMA-0 Algorithm

Input: Noisy observation Y, dictionary $\mathcal{H} = \{f_1, \ldots, f_M\}$, prior $\boldsymbol{\pi} \in \Lambda^M$, parameters ν, β.

Output: Aggregate estimator $\mathsf{f}_{\boldsymbol{\lambda}^{(k)}}$.

  Let $\boldsymbol{\lambda}^{(0)} = 0$, $\mathsf{f}_{\boldsymbol{\lambda}^{(0)}} = 0$;

  for $k = 1, 2, \ldots$ **do**

    Set $\alpha_k = \frac{2}{k+1}$;

    $J^{(k)} = \text{argmin}_j \, Q(\boldsymbol{\lambda}^{(k-1)} + \alpha_k(\mathsf{e}^{(j)} - \boldsymbol{\lambda}^{(k-1)}))$;

    $\boldsymbol{\lambda}^{(k)} = \boldsymbol{\lambda}^{(k-1)} + \alpha_k(\mathsf{e}^{(J^{(k)})} - \boldsymbol{\lambda}^{(k-1)})$;

  end for

We adopt flat priors $\pi = 1/M$ ($j = 1, \ldots, M$) for simplicity. From the definition of $Q(\boldsymbol{\lambda})$ (18), 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 (46) by setting $\nu = 0$, so $\boldsymbol{\lambda}^{\text{PROJ}}$ is approximated by GMA-0 with $\nu = 0$ and 200 iterations, and the projection algorithm is denoted by PROJ.

We evaluate two versions of the GD-BMAX algorithm that one is exactly the algorithm depicted in Algorithm 2 and the other is to approximate $\mathsf{f}_{\boldsymbol{\lambda}^{(k-1)}}$ in Algorithm 2 with the MH sampler in Algorithm 3. We denote the latter variant as the GD-MH-BMAX algorithm. For the MH sampler, we use Gaussian distribution for $q(\cdot | \cdot)$ and set $T_0 = T = 500$. The GMA-BMAX, GD-BMAX GD-MH-BMAX and GMA-0 algorithms are run for K iterations up to $K = 150$, with $\nu = 1/2$. Parameter ω for GMA-BMAX, GD-BMAX, GD-MH-BMAX and EWMA is tuned by 10-fold cross-validation. Regrets of all algorithms defined in (42) are reported for comparisons.

In the following, we consider two scenarios. The first situation is when the bases are not very correlated; in such case GMA-0 can perform better than GMA-BMAX, GD-BMAX and GD-MH-BMAX because the former (which employs

TABLE I

PERFORMANCE COMPARISON ($s = 1$ AND $M_1 = 50$)

	STAR	PROJ	EWMA
	0.3895±0.3816	0.3953±0.2680	0.2823±0.5718

	$k = 1$	$k = 5$	$k = 15$	$k = 60$	$k = 100$	$k = 150$
GMA-BMAX	0.5726±0.7536	0.4196±0.4339	0.3480±0.3606	0.2840±0.3213	0.2745±0.3180	0.2690±0.3175
GD-BMAX	1.6187±0.3313	0.7231±0.2305	0.3003±0.3098	0.2603±0.3161	0.2602±0.3161	0.2602±0.3161
GD-MH-BMAX	1.6192±0.3311	0.7253±0.2328	0.3061±0.3127	0.2645±0.3107	0.2705±0.3237	0.2583±0.3058
GMA-0	0.3742±0.7804	0.2916±0.3691	0.2567±0.3291	0.2546±0.3289	0.2548±0.3292	0.2550±0.3287

TABLE II

CUMULATIVE FREQUENCY OF REGRET ($s = 1$, $M_1 = 50$ AND $k = 150$)

Upper Boundary	0.00	0.33	0.66	0.98	1.31	1.64	1.97	2.29
GMA-BMAX	17	69	85	95	100	100	100	100
GD-BMAX	13	70	85	96	100	100	100	100
GD-MH-BMAX	14	68	85	97	100	100	100	100
GMA-0	21	72	84	96	100	100	100	100
EWMA	39	78	82	84	88	95	99	100

linear entropy) produces sparser estimators. The second situation is when the bases are highly correlated; in such case GMA-BMAX, GD-BMAX and GD-MH-BMAX are superior than GMA-0 because the former algorithms (which employ strongly convex KL-entropy) give the clustered basis functions similar weights while the GMA-0 tends to select one from the clustered basis functions, which may lead to model selection error. The correlated bases situation occurs in the continuous dictionary setting.

A. Experiment 1: When $s = 1$ and $M_1 = 50$, Bases Are Not Very Correlated

Table I compares the commonly used estimators, i.e., STAR, PROJ and EWMA, with GMA-BMAX, GD-BMAX, GD-MH-BMAX, and GMA-0. The regrets are reported using the format of "mean ± standard deviation". Table II reports the cumulative frequency of GMA-BMAX, GD-BMAX, GD-MH-BMAX, GMA-0 and EWMA with 100 replicates and fixed iteration k. For each entry, we summarize the number of replicates with regrets which are smaller than or equal to the upper boundary value.

The results in Table I indicate that GMA-0 achieves the best performance as iteration k increases. GMA-0 outperforms STAR, EWMA and PROJ after as small as $k = 15$ iterations, which still gives a relatively sparse averaged model. This is consistent with [1, Ths. 4.1 and 4.2] which shows that GMA-0 has optimal bounds for small k ($k \geq 2$).

GMA-BMAX, GD-BMAX and GD-MH-BMAX prefer dense model that assigns similar weights to similar candidates. It is easy to verify that $\boldsymbol{\psi}_X(\omega^2, \nu) = \mathsf{f}_\lambda$ with $\boldsymbol{\lambda} \in \Lambda^M$ defined as

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

and thus two similar candidates f_i and f_j will have similar weights.

In contrast, GMA-0 prefers sparse model by selecting candidates less related to the estimator from previous iteration. Specifically, with flat prior $\boldsymbol{\pi}$, the choice of $J^{(k)}$ in GMA-0

algorithm can be further simplified to

$$J^{(k)} = \underset{j}{\operatorname{argmin}} \left\{\|\mathsf{f}_j - Y\|_2^2 - (1 - \nu)(1 - \alpha_k)\|\mathsf{f}_{\boldsymbol{\lambda}^{(k-1)}} - \mathsf{f}_j\|_2^2\right\},$$
(47)

and thus at each iteration k in the GMA-0 algorithm, the estimator f_j is preferred if it is close to Y while it is less correlated to the current aggregate estimator $\mathsf{f}_{\boldsymbol{\lambda}^{(k-1)}}$ (because the minimization requires $\|\mathsf{f}_{\boldsymbol{\lambda}^{(k-1)}} - \mathsf{f}_j\|_2^2$ to be large while $\|\mathsf{f}_j - Y\|_2^2$ being small).

In Experiment 1, the first 50 candidates $\{\mathsf{f}_1, \ldots, \mathsf{f}_M\}$ are closer to the truth $\boldsymbol{\eta} = \mathsf{f}_1 + 0.5\Delta$ than other candidate f_j for $j > 50$, yet they are not very correlated when $s = 1$. Sparsity is preferred when correlations are not strong among the predicting features (in our experiment, the first 50 candidates), and GMA-0 is to output sparser estimator than GMA-BMAX, GD-BMAX and GD-MH-BMAX. Therefore, we would expect GMA-0 achieving smaller regret than GMA-BMAX, GD-BMAX and GD-MH-BMAX under this situation. Although GMA-BMAX, GD-BMAX and GD-MH-BMAX are worse than GMA-0 when the bases are not very correlated, they beat EWMA, STAR and PROJ when the iteration k is large enough.

Figure 1 compares the regrets of GMA-BMAX, GD-BMAX, GD-MH-BMAX, GMA-0 and EWMA. (a-e) illustrate the histograms of the regrets with 100 replicates. The corresponding cumulative frequencies are presented in Table II. Since Proposition 3 indicates that the optimal deviation bound is obtained by $k \to \infty$, we pick $k = 150$ for GMA-BMAX, GD-BMAX, GD-MH-BMAX and GMA-0 in order to make fair comparison. As the histograms show, although EWMA has the most replicates in which the regrets are close to zero, the distribution of the EWMA estimator is the most dispersive with many extreme values compared to other methods. The performance is consistent with [1], [9], and [10] which state that the EWMA estimator is optimal in expectation but sub-optimal in deviation. Therefore, we would expect GMA-BMAX, GD-BMAX, GD-MH-BMAX and GMA-0 enjoy more concentrated distribution than EWMA because they are also optimal in deviation. (f) shows the convergence of GMA-BMAX, GD-BMAX, GD-MH-BMAX and GMA-0. We observe that GD-BMAX and GD-MH-BMAX show faster convergence compared to GMA-BMAX, which is consistent with Proposition 5. Moreover, the GD-MH-BMAX algorithm approximates GD-BMAX well with small perturbations, which is consistent with Proposition 6.
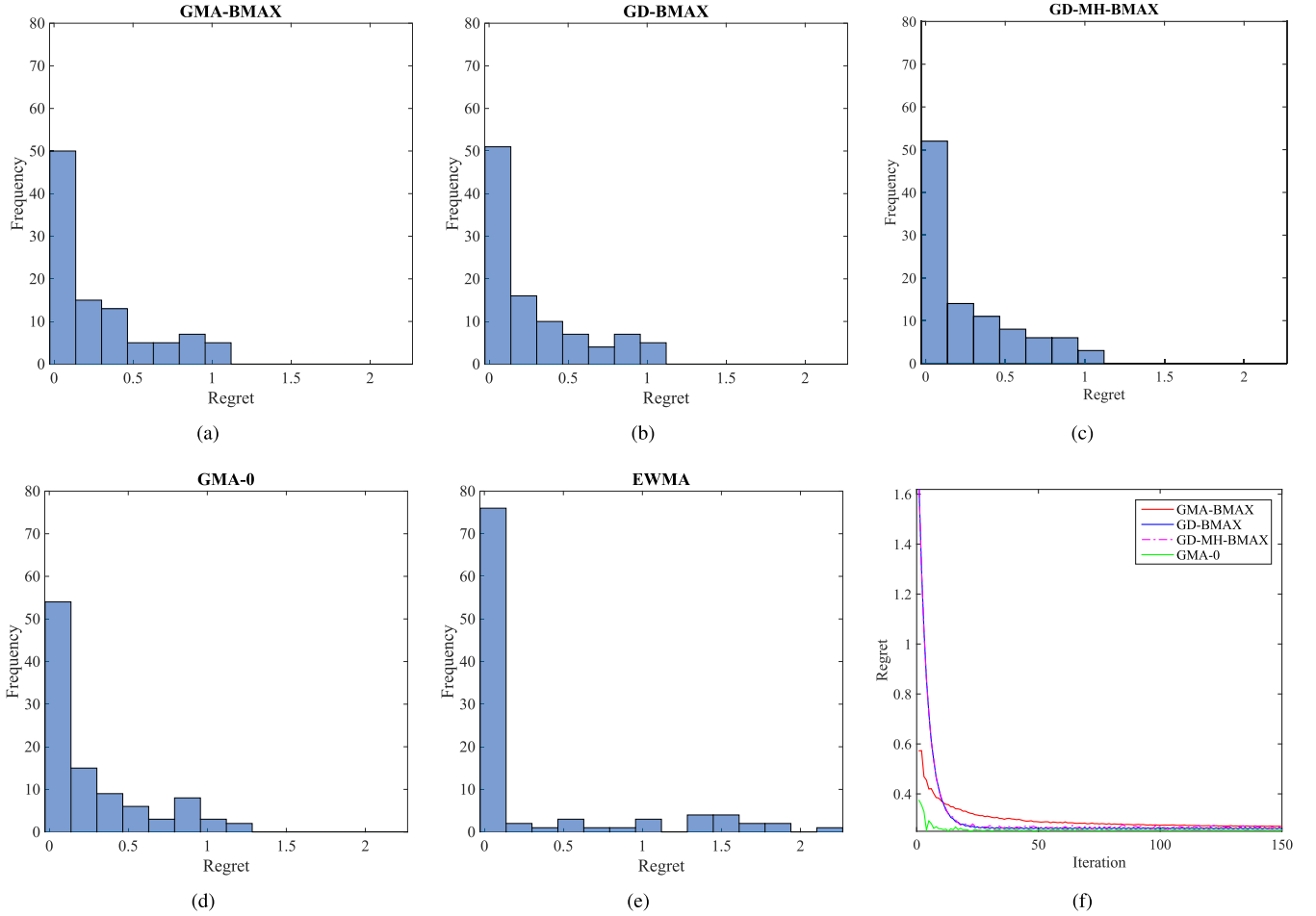
Fig. 1. (a-e) show the histograms of regrets for GMA-BMAX, GD-BMAX, GD-MH-BMAX, GMA-0 (with $k = 150$) and EWMA; (f) reports the results of regrets $R(\psi^{(k)})$ versus iterations k under the case $s = 1$ and $M_1 = 50$.

TABLE III

PERFORMANCE COMPARISON($s = \sigma/\|\zeta_j\|$ AND $M_1 = M$)

STAR	PROJ	EWMA
0.0587±0.0338	0.0495±0.0322	0.0452±0.0307

	$k = 1$	$k = 5$	$k = 15$	$k = 60$	$k = 100$	$k = 150$
GMA-BMAX	0.0929±0.0381	0.0503±0.0300	0.0431±0.0285	0.0408±0.0282	0.0407±0.0283	0.0406±0.0283
GD-BMAX	0.7817±0.2207	0.2525±0.1041	0.0501±0.0317	0.0406±0.0284	0.0406±0.0284	0.0406±0.0284
GD-MH-BMAX	0.7414±0.2126	0.1996±0.0876	0.0435±0.0291	0.0393±0.0273	0.0382±0.0273	0.0384±0.0284
GMA-0	0.0904±0.0380	0.0650±0.0379	0.0630±0.0371	0.0621±0.0364	0.0622±0.0364	0.0620±0.0364

Note that GMA-BMAX and GMA-0 produce different estimators after the first iteration ($k = 1$). GMA-BMAX selects $j \in \{1, \ldots, M\}$ that minimizes $\log J(f_j)$, while GMA-0 selects $j \in \{1, \ldots, M\}$ that minimizes $Q(f_j)$ and the output of the first stage is actually the empirical risk minimizer f_{k_1} where $k_1 = \arg\min_j \widehat{\mathrm{MSE}}(f_j)$. Moreover, since sparse model is preferred in this scenario, GMA-0 has smaller regret than GMA-BMAX, GD-BMAX and GD-MH-BMAX, and all of them converge fast within a few iterations.

B. Experiment 2: When $s = \sigma/\|\zeta_j\|$ and $M_1 = M$, Bases Are All Highly Correlated

In Experiment 2, we define regression function as $\eta = \Theta + 0.5\Delta$ which is slightly different from Experiment 1.

The results in Table III indicate that GMA-BMAX, GD-BMAX and GD-MH-BMAX perform better than GMA-0 as iteration k increases, and GMA-BMAX, GD-BMAX and GD-MH-BMAX also beat STAR, PROJ and EWMA when k is large enough.

In this experiment, all of the candidates $\{f_1, \ldots, f_M\}$ are close to the truth $\eta = \Theta + 0.5\Delta$, and they are highly correlated when s is small. GMA-BMAX, GD-BMAX and GD-MH-BMAX tend to assign similar weights to similar candidates, while GMA-0 tends to exclude other correlated candidates once one has been selected. Therefore, GMA-BMAX, GD-BMAX and GD-MH-BMAX will average over those candidates with similar weights resulting less variance (also less bias due to the design), while GMA-0 will have
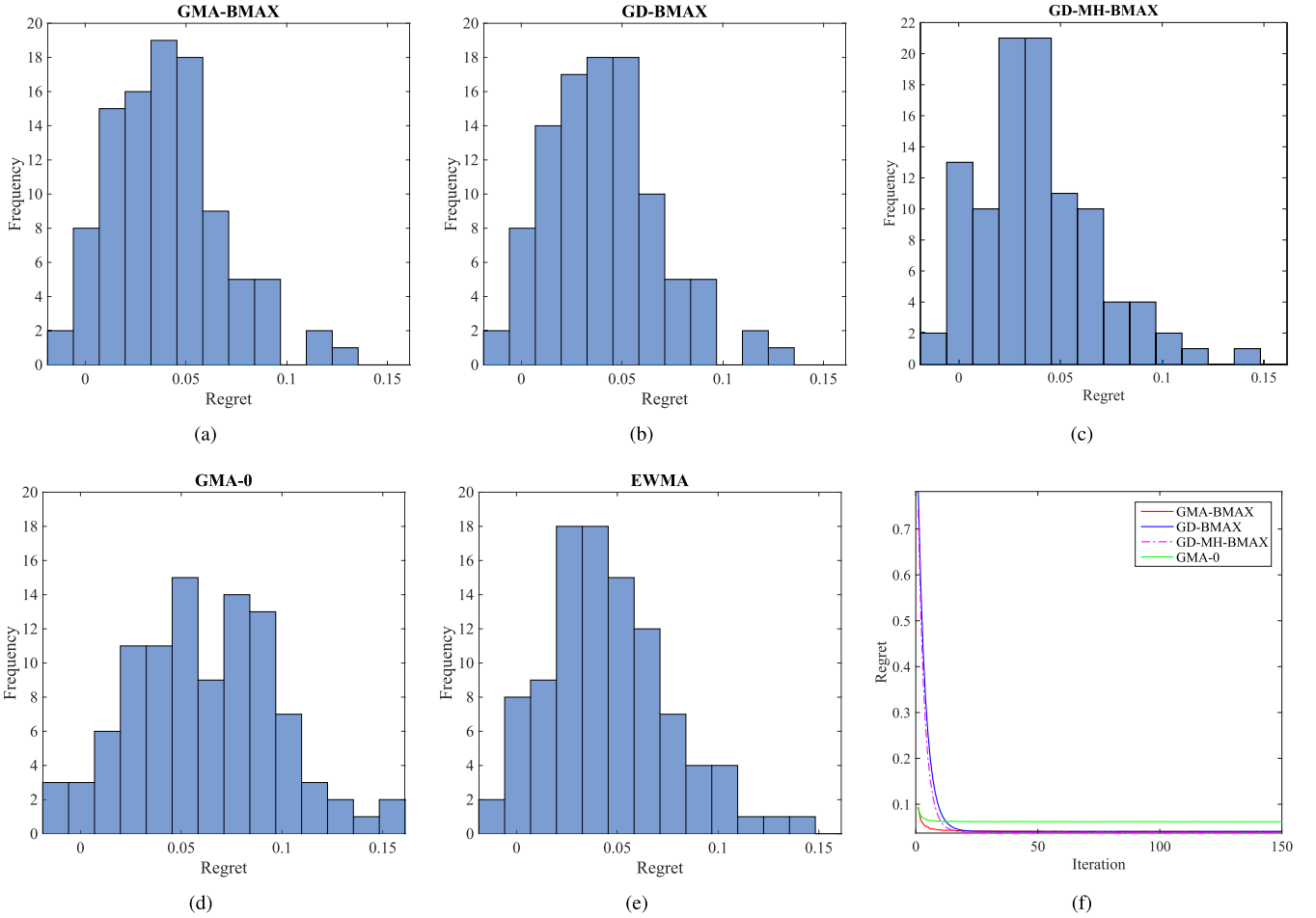
Fig. 2. (a-e) show the histograms of regrets for GMA-BMAX, GD-BMAX, GD-MH-BMAX, GMA-0 (with $k = 150$) and EWMA; (f) reports the result of regrets $R(\boldsymbol{\psi}^{(k)})$ versus iterations k under the case $s = \sigma/\|\boldsymbol{\zeta}_j\|$ and $M_1 = M$.

high variance because it only selects one of the candidates. Moreover, similar to GMA-BMAX, GD-BMAX and GD-MH-BMAX, EWMA also prefer dense model by assigning similar weights to similar candidates. However, the EWMA estimator assigns weights defined as

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

while the weights in GMA-BMAX, GD-BMAX and GD-MH-BMAX are defined as

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

where $\boldsymbol{\psi}_X(\omega^2, \nu) = \mathsf{f}_\lambda$ with $\lambda \in \Lambda^M$. Note that for all j, $\|\mathsf{f}_j - \boldsymbol{Y}\|_2^2$ are roughly equal to each other under this scenario. That is, the EWMA estimator becomes the average of all candidates. However, the GMA-BMAX, GD-BMAX and GD-MH-BMAX estimators are still weighed averages of all bases, and the weights are adjusted by the extra term $\|\boldsymbol{\psi}_X(\omega^2, \nu) - \mathsf{f}_j\|_2^2$. Therefore, we would hope GMA-BMAX, GD-BMAX and GD-MH-BMAX have smaller variances than EWMA.

Figure 2 compares the regrets of GMA-BMAX, GD-BMAX, GD-MH-BMAX, GMA-0 and EWMA.

TABLE IV
CUMULATIVE FREQUENCY OF REGRET ($s = \sigma/\|\boldsymbol{\zeta}_j\|$, $M_1 = M$ AND $k = 150$)

Upper Boundary	0.000	0.026	0.051	0.077	0.103	0.128	0.154	0.180
GMA-BMAX	5	31	69	89	97	99	100	100
GD-BMAX	6	31	69	88	97	99	100	100
GD-MH-BMAX	5	36	71	90	97	99	100	100
GMA-0	4	17	40	67	89	95	99	100
EWMA	5	26	59	86	96	98	100	100

(a-e) summarize the histograms of the regrets, and the corresponding cumulative frequencies are represented in Table IV. GMA-BMAX, GD-BMAX and GD-MH-BMAX have the most concentrated distributions, because they are optimal both in expectation and in deviation. (f) illustrates the convergence of GMA-BMAX, GD-BMAX, GD-MH-BMAX and GMA-0. All these methods converge within a few iterations. As expected, GMA-BMAX, GD-BMAX and GD-MH-BMAX achieve lower regret than GMA-0 when the basis functions are correlated.

VII. CONCLUSION

This paper introduces a new formulation for deviation optimal model averaging which we refer to as BMAX. It is motivated by Bayesian theoretical considerations with an

appropriately defined exponentiated least squares loss. Moreover we established a primal-dual relationship of this estimator and the Q-aggregation scheme (with KL entropy) by [1]. This relationship not only establishes a natural Bayesian interpretation for Q-aggregation but also leads to new numerical algorithms for model aggregation that are suitable for the continuous dictionary setting where some basis functions are highly correlated. The new formulation and its relationship to Q-aggregation provides deeper understanding of deviation optimal model averaging procedures.

APPENDIX
PROOFS

A. Proof of Lemma 1

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

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

It follows that

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

and

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

Then we have

$$\nabla^2 \log J(\boldsymbol{\psi})$$
$$= \frac{(\nabla^2 J(\boldsymbol{\psi}))J(\boldsymbol{\psi}) - (\nabla J(\boldsymbol{\psi}))(\nabla J(\boldsymbol{\psi}))^\top}{J^2(\boldsymbol{\psi})}$$
$$= \sum_{j=1}^M \lambda_j\left(\left(\frac{1-\nu}{\omega^2}\right)^2 (\boldsymbol{\psi} - f_j)(\boldsymbol{\psi} - f_j)^\top + \left(\frac{1-\nu}{\omega^2}\right)\boldsymbol{I}_n\right)$$
$$\quad - \left(\frac{1-\nu}{\omega^2}\right)^2 (\boldsymbol{\psi} - f_{\boldsymbol{\lambda}})(\boldsymbol{\psi} - f_{\boldsymbol{\lambda}})^\top$$
$$= \left(\frac{1-\nu}{\omega^2}\right)\boldsymbol{I}_n + \sum_{j=1}^M \lambda_j\left(\frac{1-\nu}{\omega^2}\right)^2 (f_{\boldsymbol{\lambda}} - f_j)(f_{\boldsymbol{\lambda}} - f_j)^\top.$$

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

With the assumption that $\|f_j\|_2 \leq L$ for all j, we have

$$\sum_{j=1}^M \lambda_j(f_{\boldsymbol{\lambda}} - f_j)(f_{\boldsymbol{\lambda}} - f_j)^\top = \sum_{j=1}^M \lambda_j f_j f_j^\top - f_{\boldsymbol{\lambda}} f_{\boldsymbol{\lambda}}^\top$$
$$\leq \sum_{j=1}^M \lambda_j f_j f_j^\top \leq \sum_{j=1}^M \lambda_j L^2 \boldsymbol{I}_n = L^2 \boldsymbol{I}_n.$$

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

B. Proof of Lemma 2

Lemma 3: For any $\boldsymbol{\lambda} \in \Lambda^M$, real numbers $\{x_j\}_{j=1}^M$ and some constant $a > 0$, we have

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

where the equality is obtained when $(x_j/a) - \log(\lambda_j/\pi_j)$ is a constant for $1 \leq j \leq M$.

Proof: The result follows directly from Jensen's Inequality as

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

Now by setting $x_j = -\nu\|f_j - \boldsymbol{h}\|_2^2$ and $a = 2\omega^2$ in Lemma 3, we obtain

$$\min_{\boldsymbol{\lambda}\in\Lambda^M}\left(\nu\sum_{j=1}^M \lambda_j\|f_j - \boldsymbol{h}\|_2^2 + 2\omega^2\mathcal{K}(\boldsymbol{\lambda}, \boldsymbol{\pi})\right) - \frac{\nu}{1-\nu}\|\boldsymbol{h} - \boldsymbol{Y}\|_2^2$$
$$= -\frac{\nu}{1-\nu}\|\boldsymbol{h} - \boldsymbol{Y}\|_2^2 - 2\omega^2 \log\left(\sum_{j=1}^M \pi_j e^{-\nu\|f_j - \boldsymbol{h}\|_2^2/2\omega^2}\right),$$

which implies that

$$T(\boldsymbol{h}) = \min_{\boldsymbol{\lambda}\in\Lambda^M} S(\boldsymbol{\lambda}, \boldsymbol{h}).$$

In addition, it is easy to verify that

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

where the minimum is achieved at $\boldsymbol{h} = f_{\boldsymbol{\lambda}}$.

Now let $\hat{\boldsymbol{h}}$ be the maximizer of $T(\boldsymbol{h})$ in (21), then by setting the derivative of (21) to zero, it is easy to observe that there exists a corresponding $\hat{\boldsymbol{\lambda}}$ so that $(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{h}}) \in A \cap B$. This means that $A \cap B \neq \emptyset$.

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

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

The third inequality is the well-known weak duality (e.g., [16, Lemma 36.1]).

Also we have

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

We thus 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 $Q(\lambda^0) = T(h^0)$. Since $(\lambda^0, h^0) \in A \cap B$, we have

$$\begin{cases} h^0 = \frac{1}{\nu}Y - \frac{1-\nu}{\nu}f_{\lambda^0}, \\ \lambda_j^0 = \frac{\exp\left(-\frac{\nu}{2\omega^2}\|f_j - h^0\|_2^2\right)\pi_j}{\sum_{i=1}^M \exp\left(-\frac{\nu}{2\omega^2}\|f_i - h^0\|_2^2\right)\pi_i}. \end{cases}$$

It follows that for all j:

$$\sum_{i=1}^M \exp\left(-\frac{\nu}{2\omega^2}\|f_i - h^0\|_2^2\right)\pi_i = \frac{\exp\left(-\frac{\nu}{2\omega^2}\|f_j - h^0\|_2^2\right)\pi_j}{\lambda_j^0},$$

which implies that

$$\log\left(\sum_{i=1}^M \exp\left(-\frac{\nu}{2\omega^2}\|f_i - h^0\|_2^2\right)\pi_i\right)$$
$$= -\frac{\nu}{2\omega^2}\|f_j - h^0\|_2^2 - \log(\lambda_j^0/\pi_j)$$
$$= \sum_{i=1}^M \lambda_i^0\left(-\frac{\nu}{2\omega^2}\|f_i - h^0\|_2^2 - \log(\lambda_i^0/\pi_i)\right),$$

where the second equation is from summing up two sides of the first equation with weight λ_i^0 over $i = 1, \ldots, M$.

Plug back into $T(h^0)$, we obtain

$$T(h^0)$$
$$= -\frac{\nu}{1-\nu}\|h^0 - Y\|_2^2$$
$$\quad - 2\omega^2\left[\sum_{i=1}^M \lambda_i^0\left(-\frac{\nu}{2\omega^2}\|f_i - h^0\|_2^2 - \log(\lambda_i^0/\pi_i)\right)\right]$$
$$= -\frac{\nu}{1-\nu}\|h^0 - Y\|_2^2 + \nu\sum_{i=1}^M \lambda_i^0\|f_i - h^0\|_2^2 + 2\omega^2 \mathcal{K}(\lambda^0, \pi)$$
$$= \|f_{\lambda^0} - Y\|_2^2 + \nu\sum_{i=1}^M \lambda_i^0\|f_i - f_{\lambda^0}\|_2^2 + 2\omega^2 \mathcal{K}(\lambda^0, \pi)$$
$$= Q(\lambda^0),$$

where the third equality is obtained by plugging in $h^0 = \frac{1}{\nu}Y - \frac{1-\nu}{\nu}f_{\lambda^0}$. Therefore,

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

Since $Q(\cdot)$ is strictly convex and $T(\cdot)$ is strictly concave, we have $h^0 = \hat{h}$ is the unique solution of $\max_h T(h)$, and $\lambda^0 = \lambda$ is the unique solution of $\min_\lambda Q(\lambda)$. Using $h^0 = \frac{1}{\nu}Y - \frac{1-\nu}{\nu}f_{\lambda^0}$, we have

$$\hat{h} = \frac{1}{\nu}Y - \frac{1-\nu}{\nu}f_{\lambda^Q}.$$

This proves that $A \cap B$ contains the unique point $(\lambda^Q, \hat{h})$. ∎

C. Proof of Proposition 1

The strong convexity of $\log J(\cdot)$ in (14) implies that

$$\|\hat{\psi} - \psi_X(\omega^2, \nu)\|_2^2 \le \frac{2}{A_1}\left(\log J(\hat{\psi}) - \log J(\psi_X(\omega^2, \nu))\right)$$
$$\le 2\epsilon/A_1.$$

Now plug the above inequality into the following equation

$$\|\hat{\psi} - \eta\|_2^2 = \|\psi_X(\omega^2, \nu) - \eta\|_2^2 + 2\|\hat{\psi} - \psi_X(\omega^2, \nu)\|_2$$
$$\times \|\psi_X(\omega^2, \nu) - \eta\|_2 + \|\hat{\psi} - \psi_X(\omega^2, \nu)\|_2^2,$$

we obtain the desired bound. ∎

D. Proof of Proposition 2

From definition, $\psi_X(\omega^2, \nu) = f_\lambda$ with $\lambda \in \Lambda^M$ defined as

$$\lambda_j \propto \pi_j \exp\left(-\frac{1}{2\omega^2}\|f_j - Y\|_2^2 + \frac{1-\nu}{2\omega^2}\|\psi_X(\omega^2, \nu) - f_j\|_2^2\right).$$

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

$$\log J(\psi^{(k)})$$
$$= \log J\left(\psi^{(k-1)} + \alpha_k(f_{J^{(k)}} - \psi^{(k-1)})\right)$$
$$\le \log J\left(\psi^{(k-1)} + \alpha_k(f_j - \psi^{(k-1)})\right)$$
$$\le \log J(\psi^{(k-1)}) + \alpha_k(f_j - \psi^{(k-1)})^\top \frac{\nabla J(\psi^{(k-1)})}{J(\psi^{(k-1)})} + 2\alpha_k^2 A_3,$$

where the first inequality comes from definition, the second inequality is from Taylor expansion at $\psi^{(k-1)}$ and (15) in Lemma 1 with the fact that $\|f_j - \psi^{(k-1)}\|_2^2 \le 4L^2$.

We multiply the above inequality by λ_j and sum over j to obtain

$$\log J(\psi^{(k)}) \le \log J(\psi^{(k-1)}) + \alpha_k \sum_{j=1}^M \lambda_j(f_j - \psi^{(k-1)})^\top$$
$$\times \frac{\nabla J(\psi^{(k-1)})}{J(\psi^{(k-1)})} + 2\alpha_k^2 A_3$$
$$= \log J(\psi^{(k-1)}) + \alpha_k(\psi_X(\omega^2, \nu) - \psi^{(k-1)})^\top$$
$$\times \frac{\nabla J(\psi^{(k-1)})}{J(\psi^{(k-1)})} + 2\alpha_k^2 A_3$$
$$\le \log J(\psi^{(k-1)}) + \alpha_k(\log J(\psi_X(\omega^2, \nu))$$
$$\times - \log J(\psi^{(k-1)})) + 2\alpha_k^2 A_3,$$

where the last inequality follows from the convexity of $\log J(\psi)$.

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

$$\delta_k \le (1 - \alpha_k)\delta_{k-1} + 2\alpha_k^2 A_3.$$

We now bound δ_0. If we let $\mu_j \propto \pi_j \exp\left(-\frac{1}{2\omega^2}\|f_j - Y\|_2^2\right)$ such that $\sum_{j=1}^M \mu_j = 1$, then

$$\delta_0 = \log J(\psi^{(0)}) - \log J(\psi_X(\omega^2, \nu))$$
$$= \log \sum_j \mu_j \exp\left(\frac{1-\nu}{2\omega^2}\|\psi^{(0)} - f_j\|_2^2\right)$$
$$\quad - \log \sum_j \mu_j \exp\left(\frac{1-\nu}{2\omega^2}\|\psi_X(\omega^2, \nu) - f_j\|_2^2\right)$$
$$\le \log\left(\sum_{j=1}^M \mu_j \exp\left(\frac{1-\nu}{2\omega^2}\|\psi^{(0)} - f_j\|_2^2\right)\right)$$
$$\le \frac{1-\nu}{2\omega^2}L^2 \le 2A_3. \tag{48}$$

The claim thus hold for δ_0. By mathematical induction, if $\delta_{k-1} \leq \frac{8A_3}{k+2}$, then

$$\delta_k \leq (1 - \alpha_k)\delta_{k-1} + 2\alpha_k^2 A_3$$

$$\leq (1 - 2/(k+1))\frac{8A_3}{k+2} + 2(2/(k+1))^2 A_3 \leq \frac{8A_3}{k+3}.$$

This proves the desired bound. ∎

E. Proof of Proposition 3

We have

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

$$\leq \frac{2}{A_1}\frac{8A_3}{k+3} = \frac{16A_3}{A_1(k+3)},$$

where the first inequality comes from Taylor expansion at point $\boldsymbol{\psi}_X(\omega^2, \nu)$, with using (14) in Lemma 1 and $\nabla J(\boldsymbol{\psi}_2)$; and the second inequality is from Proposition 2. It follows that

$$\|\boldsymbol{\psi}^{(k)} - \boldsymbol{\eta}\|_2^2$$

$$= \|(\boldsymbol{\psi}^{(k)} - \boldsymbol{\psi}_X(\omega^2, \nu)) + (\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta})\|_2^2$$

$$\leq \|\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta}\|_2^2 + 2\|\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta}\|_2\|\boldsymbol{\psi}^{(k)}$$

$$- \boldsymbol{\psi}_X(\omega^2, \nu)\|_2 + \|\boldsymbol{\psi}^{(k)} - \boldsymbol{\psi}_X(\omega^2, \nu)\|_2^2$$

$$\leq \|\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta}\|_2^2 + 2\sqrt{\frac{16A_3}{A_1(k+3)}}\|\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta}\|_2$$

$$+ \frac{16A_3}{A_1(k+3)}.$$

∎

F. Proof of Proposition 4

The update operation implies

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

Then,

$$\log J(\boldsymbol{\psi}^{(k)})$$

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

$$\leq \log J(\boldsymbol{\psi}^{(k-1)}) - t_k\|\nabla \log J(\boldsymbol{\psi}^{(k-1)}))\|_2^2$$

$$+ (A_2/2)t_k^2\|\nabla \log J(\boldsymbol{\psi}^{(k-1)}))\|_2^2$$

$$= \log J(\boldsymbol{\psi}^{(k-1)}) - (t_k - (A_2/2)t_k^2)\|\nabla \log J(\boldsymbol{\psi}^{(k-1)}))\|_2^2,$$

where the inequality is from (15). By subtracting $\log J(\boldsymbol{\psi}_X(\omega^2, \nu))$ by each side, we have

$$\log J(\boldsymbol{\psi}^{(k)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))$$

$$\leq \log J(\boldsymbol{\psi}^{(k-1)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))$$

$$- (t_k - (A_2/2)t_k^2)\|\nabla \log J(\boldsymbol{\psi}^{(k-1)}))\|_2^2. \qquad (49)$$

Also from (15) we have

$$\|\nabla \log J(\boldsymbol{\psi}^{(k-1)}))\|_2^2$$

$$\geq 2A_1\left(\log J(\boldsymbol{\psi}^{(k-1)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))\right). \quad (50)$$

Choose fixed step size $t_k = s \in (0, 2/A_2)$ for any $k > 0$. Combining (49) and (50) results

$$\log J(\boldsymbol{\psi}^{(k)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))$$

$$\leq [1 - 2A_1(s - (A_2/2)s^2)]\left(\log J(\boldsymbol{\psi}^{(k-1)})\right.$$

$$\left. - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))\right).$$

It follows that

$$\log J(\boldsymbol{\psi}^{(k)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))$$

$$\leq [1 - 2A_1(s - (A_2/2)s^2)]^k\left(\log J(\boldsymbol{\psi}^{(0)})\right.$$

$$\left. - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))\right).$$

∎

G. Proof of Proposition 5

We choose $t_k = s$ as in Remark 1. Then,

$$\|\boldsymbol{\psi}^{(k)} - \boldsymbol{\psi}_X(\omega^2, \nu)\|_2^2$$

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

$$\leq \frac{2}{A_1}(1 - A_1/A_2)^k \log J(\boldsymbol{\psi}^{(0)})$$

$$\leq \frac{2}{A_1}(1 - A_1/A_2)^k\frac{1-\nu}{2\omega^2}L^2 = L^2(1 - A_1/A_2)^k,$$

where the first inequality comes from Taylor expansion at point $\boldsymbol{\psi}_X(\omega^2, \nu)$, with using (14) in Lemma 1 and $\nabla \log J(\boldsymbol{\psi}_X(\omega^2, \nu)) = 0$; the second inequality is from Proposition 4; and the third inequality is from assumption (11) resulting $\log J(\boldsymbol{\psi}^{(0)}) \leq \frac{1-\nu}{2\omega^2}L^2$. It follows that

$$\|\boldsymbol{\psi}^{(k)} - \boldsymbol{\eta}\|_2^2$$

$$= \|(\boldsymbol{\psi}^{(k)} - \boldsymbol{\psi}_X(\omega^2, \nu)) + (\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta})\|_2^2$$

$$\leq \|\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta}\|_2^2 + 2\|\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta}\|_2\|\boldsymbol{\psi}^{(k)}$$

$$- \boldsymbol{\psi}_X(\omega^2, \nu)\|_2 + \|\boldsymbol{\psi}^{(k)} - \boldsymbol{\psi}_X(\omega^2, \nu)\|_2^2$$

$$\leq \|\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta}\|_2^2 + 2\sqrt{L^2(1 - A_1/A_2)^k}\|\boldsymbol{\psi}_X(\omega^2, \nu) - \boldsymbol{\eta}\|_2$$

$$+ L^2(1 - A_1/A_2)^k.$$

∎

H. Proof of Proposition 6

Given Y, the expectation is with respect to the randomness from the MH algorithm. For $k > 0$, $\boldsymbol{u}_T^{(k-1)}$ from Algorithm 3 is an estimator of $\mathbf{f}_{\boldsymbol{\lambda}^{(k-1)}} = \sum_{j=1}^M \lambda_j^{(k-1)}\mathbf{f}_j$. 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 $\upsilon^{(k-1)} = \frac{1-\nu}{\omega^2}(\boldsymbol{\psi}^{(k-1)} - \boldsymbol{u}_T^{(k-1)})$, then we have

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

and

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

It follows that

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

where the inequality is from (15). Then, by subtracting $\log J(\boldsymbol{\psi}_X(\omega^2, \nu))$ from each side of the above equation and taking expectation conditioned on $\boldsymbol{\psi}^{(k-1)}$, we have

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

where $\delta_k = \log J(\boldsymbol{\psi}^{(k)}) - \log J(\boldsymbol{\psi}_X(\omega^2, \nu))$ and $t_k = s = 1/A_2$ as in Remark 1. Combining the above inequality with

$$\|\nabla \log J(\boldsymbol{\psi}^{(k-1)}))\|_2^2 \geq 2A_1 \Big(\log J(\boldsymbol{\psi}^{(k-1)})$$
$$- \log J(\boldsymbol{\psi}_X(\omega^2, \nu)) \Big), \quad (51)$$

which is from (15), we have

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

Therefore, it follows that

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

and we have

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

∎

ACKNOWLEDGMENT

REFERENCES

[1] D. Dai, P. Rigollet, and T. Zhang, "Deviation optimal learning using greedy Q-aggregation," *Ann. Statist.*, vol. 40, no. 3, pp. 1878–1905, 2012.

[2] A. B. Tsybakov, "Optimal rates of aggregation," in *Learning Theory and Kernel Machines* (Lecture Notes in Computer Science), vol. 2777, B. Schölkopf and M. K. Warmuth, Eds. Berlin, Germany: Springer, 2003, pp. 303–313.

[3] P. Rigollet, "Kullback–Leibler aggregation and misspecified generalized linear models," *Ann. Statist.*, vol. 40, no. 2, pp. 639–665, 2012.

[4] P. Rigollet and A. B. Tsybakov. (2012). "Sparse estimation by exponential weighting." [Online]. Available: https://arxiv.org/abs/1108.5116

[5] J.-Y. Audibert, "Progressive mixture rules are deviation suboptimal," in *Advances in Neural Information Processing Systems 20*, J. C. Platt, D. Koller, Y. Singer, and S. Roweis, Ed. Cambridge, MA, USA: MIT Press, 2008, pp. 41–48.

[6] G. Lecué and S. Mendelson, "Aggregation via empirical risk minimization," *Probab. Theory Rel. Fields*, vol. 145, nos. 3–4, pp. 591–613, 2009.

[7] S. Gaïffas and G. Lecué, "Hyper-sparse optimal aggregation," *J. Mach. Learn. Res.*, vol. 12, pp. 1813–1833, Jan. 2011.

[8] D. Dai and T. Zhang, "Greedy model averaging," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, Eds. Cambridge, MA, USA: MIT Press, pp. 1242–1250 2011.

[9] A. S. Dalalyan and A. B. Tsybakov, "Aggregation by exponential weighting and sharp oracle inequalities," in *Learning Theory* (Lecture Notes in Computer Science), vol. 4539. Berlin, Germany: Springer, 2007, pp. 97–111.

[10] A. Dalalyan and A. B. Tsybakov, "Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity," *Mach. Learn.*, vol. 72, nos. 1–2, pp. 39–61, Aug. 2008.

[11] G. Leung and A. R. Barron, "Information theory and mixing least-squares regressions," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3396–3410, Aug. 2006.

[12] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval Res. Logistics Quart.*, vol. 3, nos. 1–2, pp. 95–110, 1956.

[13] L. K. Jones, "A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training," *Ann. Statist.*, vol. 20, no. 1, pp. 608–613, 1992.

[14] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 930–945, May 1993.

[15] J.-M. Marin and C. Robert, *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Berlin, Germany: Springer, 2007.

[16] R. T. Rockafellar, *Convex Analysis* (Princeton Landmarks in Mathematics and Physics). Princeton, NJ, USA: Princeton Univ. Press, 1997.

Dong Dai received the PhD degree from the Department of Statistics, Rutgers University, in 2013. He was with the IMS Health, PA, USA. His research interests include statistical methods, mathematics, data mining and machine learning methods. He has extensive experience of statistical and data mining techniques, including regression/classification, decision trees, neural network, experiment design, time series, Bayesian approaches, support vector machine, model selection, and ensemble learning, etc. He has deep understanding of statistical concepts and methods, with publications in peer reviewed journals, and perform reviews for journals in statistics, machine learning, data mining and healthcare.

Lei Han received the PhD degree from the School of Electronics Engineering and Computer Science (EECS), Peking University, in 2014. He is currently a senior researcher in Tencent AI Lab, Shenzhen, China. He worked as a research assistant professor at Mississippi State University, and worked as postdoctoral research associate at Rutgers University and Hong Kong Baptist University respectively. His research interests include artificial intelligence, machine learning, and data mining. He is especially interested in statistical methods, reinforcement learning, transfer learning, and optimization problems.

Ting Yang received the B.S. degree in statistics from Peking University in 2011. She is currently pursuing Ph.D. degree at Rutgers University. Her research interests include statistical computing, nonparametric methods and subsampling.

Tong Zhang received the B.A. degree in mathematics and computer science from Cornell University and the Ph.D. degree in Computer Science from Stanford University. He is a machine learning researcher, and the executive director of Tencent AI Lab. Previously, he was a professor at Rutgers university, and worked at IBM, Yahoo, and Baidu. His research interests include machine learning algorithms and theory, statistical methods for big data and their applications. His research has been supported by many grants from funding agencies such as NSF and NIH. He is a fellow of ASA and IMS, and he has been in the editorial boards of leading machine learning journals and program committees of top machine learning conferences.