A Nonparametric Maximum Likelihood Approach to Mixture of Regressions

Hansheng Jiang hansheng_jiang@berkeley.edu
Adityanand Guntuboyina aditya@stat.berkeley.edu
University of California, Berkeley

Abstract

Mixture of regression models are useful for regression analysis in heterogeneous populations where a single regression model may not be appropriate for the entire population. We study the nonparametric maximum likelihood estimator (NPMLE) for fitting these models. The NPMLE is based on convex optimization and does not require prior specification of the number of mixture components. We establish existence of the NPMLE and prove finite-sample parametric (up to logarithmic multiplicative factors) Hellinger error bounds for the predicted density functions. We also provide an effective procedure for computing the NPMLE without ad-hoc discretization and prove a theoretical convergence rate under certain assumptions. Numerical experiments on simulated data for both discrete and non-discrete mixing distributions demonstrate the remarkable performance of our approach. We also illustrate the approach on two real data sets.
Keywords: Conditional Gradient Method (CGM), finite-sample Hellinger error, mixture of regressions, nonparametric maximum likelihood estimator (NPMLE), parametric rate

1 Introduction

1.1 Mixture of regressions

Given a univariate response variable Y and a p-dimensional regressor variable X, the usual homoscedastic Gaussian linear regression model assumes that the conditional distribution of Y given $X = x$ is normal with mean $x^\top \beta$ and variance σ^2 for some $\beta \in \mathbb{R}^p$ and $\sigma^2 > 0$. In other words, the conditional density of $Y|X = x$ is given by

$$y \mapsto \frac{1}{\sigma}\phi\left(\frac{y - x^\top \beta}{\sigma}\right).$$

Here ϕ is the standard normal density function, and $N(0, 1)$ is the standard normal distribution. In contrast, the mixture of linear regressions model assumes that the conditional density of Y given $X = x$ is given by the mixture density

$$y \mapsto \int \frac{1}{\sigma}\phi\left(\frac{y - x^\top \beta}{\sigma}\right) \mathrm{d}G(\beta)$$

for some probability measure G on $\mathbb{R}^p$ and $\sigma^2 > 0$. Equivalently, given $X = x$, we can write

$$Y = x^\top \beta + \sigma Z \qquad \text{where } \beta \sim G \text{ and } Z \sim N(0, 1) \text{ are independent.} \tag{1.1}$$

Because the regression coefficient β is assumed to be random, the mixture of linear regressions model is also known as the random coefficient regression model in related literature.

The mixture of regressions model is a prominent mixture model in statistics and has a lengthy history (Quandt, 1958; De Veaux, 1989). It is also shown to be related to phase retrieval (Netrapalli et al., 2013; Balakrishnan et al., 2017). The mixture of regressions model naturally arises in various fields including pharmacokinetics (Lai and Shih, 2003), and marketing (Wedel and Kamakura, 2012). In population pharmacokinetics, different coefficients represents how different subjects react to drug treatments. In marketing and business, the mixture of regressions model is used to study consumer heterogeneity in order to analyze demand and future sales. Mixture of regressions models have been very popular in applications due to their modeling simplicity and effectiveness for establishing the relationship between responses and regressors from a heterogeneous population.

In this paper, we study the estimation problem under the mixture of regressions model from independent observations $(x_1, y_1), \ldots, (x_n, y_n)$ drawn according to the model (1.1). We consider both fixed design and random design settings of $x_1, \ldots, x_n$. The two unspecified terms in (1.1) are the probability measure G and the scale parameter σ. We assume that G is either completely unspecified or only known to be contained in a ball centered at the origin. In contrast, we assume that the scale parameter σ is known, or a range of σ is known. We focus on linear regressions in the main context of this paper. Nevertheless, as we will show later, our computational procedure applies generally to any nonlinear regression models, and our theoretical prediction error analysis also holds beyond linear regression models.

1.2 Related work

Model (1.1) encompasses two well-known models. (1) If G is assumed to be supported on only k points, model (1.1) becomes the finite mixture of linear regressions model with k components, also known as the hierarchical mixture of experts model (Jordan and Jacobs, 1994) in the machine learning community. (2) If $x_1, \ldots, x_n$ are identical, then $y_1, \ldots, y_n$ are independently and identically distributed, and model (1.1) is reduced to the Gaussian location mixture model. The Gaussian location mixture model, including finite mixtures of Gaussians, has broad applications in clustering and discriminant analysis (Cai et al., 2019).

The finite-mixture modeling approach comes with a priori problem of choosing k, which is long considered challenging due to its nonregularity (Kasahara and Shimotsu, 2015). Moreover, the finite-mixture model does not consider non-discrete probability measure G, and is therefore prone to misspecification. In terms of computation, the expectation-maximization algorithm is commonly used to estimate the finite mixture of linear regressions (Faria and Soromenho, 2010) via a likelihood-maximization approach. Despite its widespread popularity and broad applicability, the expectation-maximization algorithm is known to suffer from convergence to non-global optimums and sensitivity to initialization, and a complete theoretical understanding of expectation-maximization algorithms remains largely elusive (Balakrishnan et al., 2017). Some works from the machine learning community have proposed and analyzed alternative algorithms for the finite mixture of linear regressions models with parameter recovery guarantees under certain regularity assumptions, where the primary focus is on finite-component mixtures with relatively low signal-to-noise ratio (Li and Liang, 2018; Kwon et al., 2019).

In contrast to the extensive studies on estimation of finite mixtures and estimation under parametric assumptions of G, nonparametric treatment for mixture of regressions models is relatively rare in the literature. In terms of linear regressions, the settings in Beran et al. (1996) are mostly relevant to us, where they used characteristic function of the density function of y_i based on inverse Radon transform. Beran and Millar (1994) proposed an minimum distance estimator that minimizes the distance between the empirical distribution and a predefined nonparametric family of distributions, where the distance is any statistical distance that metrizes weak convergence. The estimation method in Beran and Millar (1994) is defined under a broader multivariate setting, but the numerical algorithm is computationally expensive and plagued by local minimas. Hoderlein et al. (2010) proposed an estimator by combining

kernel density estimation with Radon transform, but their method needs strong Sobolev smoothness assumption on the density function of β. None of these works have addressed the estimation of mixture of regressions with a principled likelihood maximization approach as presented in our work.

1.3 Main results

Following the discussion above, a natural question arises. Does there exist a nonparametric estimator that not only has satisfying numerical performances but also comes with certain theoretical guarantees in general? Aiming to give an affirmative answer to this question, we study the nonparametric maximum likelihood estimator (NPMLE) for the mixture of regressions model. There is a lengthy history of using NPMLEs for mixture models, to name a few, ranging from the early works on estimating mixture of distributions by Kiefer and Wolfowitz (1956) and Laird (1978) to related theories in optimal design reviewed in Silvey (1980) (for an excellent survey on classical works, see, e.g., Lindsay (1995)).

The nonparametric maximum likelihood estimator of the mixture of regressions model refers to any maximizer in the following optimization problem,

$$\hat{G} \in \arg \max_{G \text{ supported on } K} \sum_{i=1}^{n} \log f_{x_i}^{G}(y_i), \tag{1.2}$$

Here

$$f_{x_i}^{G}(y_i) = \frac{1}{\sigma} \int \phi\left(\frac{y_i - x_i^\top \beta}{\sigma}\right) \mathrm{d}G(\beta), i = 1, \ldots, n \tag{1.3}$$

is the conditional density function of y_i given x_i. Here G is a probability measure, and K is a set in $\mathbb{R}^p$. Since G is not assumed to be finitely supported or follow any parametric form in the optimization problem here, this estimation method is nonparametric in nature.

It is commonly assumed that K is a compact set, for simplicity, K can be taken as a ball centered at the origin with some radius R, $R > 0$. Under the compactness assumption, the existence of the NPMLE has been previously established (see, e.g., Lindsay (1983)). It is also known that there exists an NPMLE that is supported on at most n points in K. However, it is not previously known whether the NPMLE still exists without the compactness assumption. In Section 2.1, we extend the existence results to the completely unspecified case, i.e., $K = \mathbb{R}^p$. Though the NPMLE $\hat{G}$ as defined in (1.2) might not be unique, the vector $(f_{x_1}^{\hat{G}}(y_1), \ldots, f_{x_n}^{\hat{G}}(y_n))$ is unique for any $\hat{G}$. This is a simple argument following the strong concavity of the logarithm function in (1.2).

To compute $\hat{G}$, we develop an effective computational procedure based on the Conditional Gradient Method (CGM). This computational procedure works for any specification of K and is applicable when the linear regression model $x^\top \beta$ is replaced by other sorts of parametric functions $r(x, \beta)$. The Conditional Gradient Method is guaranteed to strictly increase the likelihood over the iterations before convergence. Unlike previous algorithms (Wu, 1983; Böhning, 1986) that rely on ad-hoc parameter space discretization, CGM can adaptively select new support points from the set K during the execution of the algorithm. The new support points are found by solving a subproblem at each iteration. We provide systematic algorithmic analysis of CGM and prove the global convergence rate.

Give x, we can use $f_x^{\hat{G}}(y)$ as an estimator of $f_x^{G^*}(y)$, where G^* denotes the true distribution of β. The estimator $f_x^{\hat{G}}(y)$ can be directly applied to construct prediction intervals of y given the regressor x, which is of great importance in applications, for example in the analysis of longitudinal data.

To study the theoretical properties of NPMLEs, we investigate the discrepancy between $f_x^{\hat{G}}(y)$ and $f_x^{G^*}(y)$. We use the squared Hellinger distance $\mathfrak{H}(f_x^{\hat{G}}, f_x^{G^*})$ to quantify of the error in predicting y for regressor x, where

$$\mathfrak{H}^2\left(f_x^{\hat{G}}, f_x^{G^*}\right) = \int \left\{(f_x^{\hat{G}}(y))^{1/2} - (f_x^{G^*}(y))^{1/2}\right\}^2 \mathrm{d}y. \tag{1.4}$$

One can average the squared Hellinger distance in (1.4) over x to get an overall measure of prediction error. In the fixed design setting, it is natural to average over $x_i, i = 1, \ldots, n$, which leads to the loss function in (1.5)

$$\mathfrak{H}_n^2\left(f^{\hat{G}}, f^{G^*}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathfrak{H}^2\left(f_{x_i}^{\hat{G}}, f_{x_i}^{G^*}\right). \tag{1.5}$$

In the random design setting, we can take an average of Hellinger distance with respect to the distribution of x assuming $x \sim \mu$ for some generating distribution μ. We present a detailed analysis of the tail of $\mathfrak{H}_n^2\left(f^{\hat{G}}, f^{G^*}\right)$, as well as the error expectation $\mathbb{E}\mathfrak{H}_n^2\left(f^{\hat{G}}, f^{G^*}\right)$ and $\mathbb{E}\bar{\mathfrak{H}}_n^2\left(f^{\hat{G}}, f^{G^*}\right)$ for fixed design and random design respectively. In particular, we prove that the prediction error $\mathbb{E}\mathfrak{H}_n^2\left(f^{\hat{G}}, f^{G^*}\right)$ is $\tilde{O}(n^{-1})$ under only the compactness assumption on K. The error bound is finite-sample and hold even when k grows with n. Built upon the analysis for fixed design, we move to random design and present a finite-sample bound on the prediction error under random design. Adopting the notion of strong identifiability in Beran and Millar (1994), we further show that the distance between $\hat{G}$ and G^* converges to zero in probability under mild conditions, where the distance can be chosen as any distance on probability measure space that metrizes weak convergence.

Our proofs of Hellinger accuracy align with the line of works on empirical processes for characterizing the performance of maximum likelihood estimators (van der Vaart et al., 1996). Previous work on the Hellinger accuracy of NPMLEs are attributed to Ghosal and Van Der Vaart (2001) and Zhang (2009) on univariate Gaussian location mixtures and recently the generalization to multi-dimensional Gaussian location mixtures by Saha and Guntuboyina (2020). However, our models on the mixture of regressions contains non-identical regressors for each data point, which is critically different from the mixture of distributions where are data points are identically distributed. Our theoretical contributions include proposing an appropriate pseudometric for the family of density functions in model (1.1) and proving sharp covering number bounds under the proposed pseudometric. The transfer to another pseudometric whose definition relies on both the regressors and response variables is crucial for the analysis of regression models. Based on the new covering number bounds, we accomplish the final proof of the Hellinger accuracy by tail arguments similar to Zhang (2009) and Saha and Guntuboyina (2020).

1.4 Notation and Organization

Let G be a probability measure of the p-dimensional Euclidean space $\mathbb{R}^p$. $f_x^G(y)$ is the conditional density functions of the univariate response y given a p-dimensional regressor x. The true mixing probability measure is denoted by G^*, and we introduce the shorthand notation $f_x^*(y)$ to indicate $f_x^{G^*}(y)$. Let $\{(x_i, y_i)\}_{i=1}^n$ be n data points. The n-dimensional vector $\mathrm{f}^G = (f_{x_1}^G(y_1), \ldots, f_{x_n}^G(y_n))^\top$ is called a mixture likelihood vector. In the special case when $G = \delta(\beta)$, i.e. when G is a point mass probability measure, $\mathrm{f}^\beta = (f_{x_1}^\beta(y_1), \ldots, f_{x_n}^\beta(y_n))^\top$ is called an atomic likelihood vector. For a likelihood vector f, we use $\mathrm{f}(i)$ to represent the i-th component of f for $i = 1, 2, \ldots, n$.

The rest of this paper is organized as follows. In Section 2, we prove the existence of the NPMLEs for $K = \mathbb{R}^p$, which previous works (Lindsay, 1983; Böhning, 2000) do not directly apply. We also provide insights and methods about computing the NPMLEs efficiently based on modern optimization methods. In Section 3, we talk about the nearly parametric rate of the prediction error in terms of the Hellinger distance. In Section 4 and Section 5, we show the numerical experiments on simulated data and real data respectively. We conclude the paper with some discussions and future works in Section 6. The appendices include all the technical proofs of our theoretical results.

2 Existence and computation of NPMLEs

2.1 Existence

In this section, we establish the existence of the nonparametric maximum likelihood estimators (NPMLEs) for the mixture of regressions model. In Theorem 2.1, we show that the NPMLEs exist and there exists an NPMLE that is a convex combination of at most n points in K, when the support set K of the mixing probability measure is assumed to be the whole space $\mathbb{R}^p$ or a compact set.

Theorem 2.1. *For model* (1.1) *and the NPMLE defined as in* (1.2)*, given data* $\{(x_i, y_i)\}_{i=1}^n$*, if* $K = \mathbb{R}^p$ *or* K *is a compact set in* $\mathbb{R}^p$*, there exists an NPMLE that is supported on at most* n *points in* K*.*

When K is compact, the existence proof is fairly standard in the literature (see, e.g., Lindsay (1983)). Though most existing results are stated for non-regressor cases, the proof can be extended to the mixture of regressions model straightforwardly. In the unbounded case, we focus on the scenario that generates the full NPMLE without any constraints i.e. $K = \mathbb{R}^p$. This scenario is most common when dealing with real data sets, where we usually do not have direct information on the range of regression parameters. For other cases where K is neither $\mathbb{R}^p$ nor bounded, for a typical example, if K is a linear subspace of $\mathbb{R}^p$ of dimension d_k, the model can be directly reparametrized to a model on $\mathbb{R}^{d_k}$. In practice, one can usually assume that the norm of parameters are not too large so that K can be taken as a ball centered at the origin.

We stress that the existence problem in the mixture of regressions model is different from the one lies in the estimation of mixtures of distributions for the unbounded case. In the Gaussian location mixture model, the support set K is readily reduced to a compact region that contains all the observed data $y_1, \ldots, y_n$. Intuitively, this reduction is correct because any location parameter that is far away from all data points $y_i, i = 1, 2, \ldots, n$ is unlikely to contribute to the maximization of the likelihood. More specifically, for multivariate Gaussian location mixture model with unknown means $\theta_1^*, \ldots, \theta_n^*$, Feng and Dicker (2018, Proposition 1) showed that NPMLEs can be found in $\text{conv}(\hat{\theta}_1, \ldots, \hat{\theta}_n\})$ where $\hat{\theta}_i$ is a unimodal maximizer of the density function for data point y_i over all points in K. However, similar reduction to compact regions is not valid for mixture of regressions models. To see this, we first notice that the analog of the location parameter in the Gaussian location mixture model is $x_i^\top \beta$ in the mixture of regressions model (1.1). However, for every i, there exists β with arbitrarily large norm and relatively small value $\|y_i - x_i^\top \beta\|$ at the same time, as long as β is almost perpendicular to x_i. Therefore, the support of NPMLEs may not be restricted in a bounded region around the origin beforehand. Moreover, the geometric relationships of n regressors $x_1, \ldots, x_n$ are not presumably assumed, which greatly complicates the analysis here.

Apart from establishing the existence, Theorem 2.1 also reveals the discrete nature of the solution. The discreteness turns out to be useful in developing our computational procedure, as we show later in this section. Since the likelihood function is evaluated at n data points $\{(x_i, y_i)\}_{i=1}^n$, we first transform the infinite-dimensional maximization problems over all probability measures into a n-dimensional optimization problem that is easier to deal with. To show how the transformation is done, the notation of likelihood vectors defined at the end of Section 1 is used throughout. We define two sets $\mathcal{P}_K, \mathcal{Q}_K$ composed of likelihood vectors, $\mathcal{P}_K = \{\mathrm{f}^\beta : \beta \in K\}, \mathcal{Q}_K = \{\mathrm{f}^G : G \text{ is any probability measure supported on } K\}$, and write them for brevity as $\mathcal{P}, \mathcal{Q}$ respectively when $K = \mathbb{R}^p$.

2.2 Computation of NPMLEs via iterative approximations

Computing NPMLEs, even in the non-regressor and univariate response case, is challenging. Towards the computation of NPMLEs, a number of algorithms have been proposed in the literature. Most of the existing approaches focus on the problem of mixtures of Gaussian distributions, i.e., non-regressor case.

We now review some recent efforts in this direction and explain why these approaches are not favored or even not applicable in our mixture of regressions model.

The fact that one NPMLE is supported on n points implies a direct connection to finite mixtures. One may find one NPMLE by solving a finite mixture model with n components, in contrast to the common k-component mixture model where k is usually much smaller than n. However, likelihood maximization in finite mixtures model is no long convex and thus hard to solve accurately. When the Gaussian distribution is one dimensional, Jiang and Zhang (2009) employed the expectation-maximization algorithm to iteratively update the weights and support points of n mixtures, and use the last iteration as the approximation of NPMLEs. The expectation-maximization algorithm is dominating in solving non-convex problems for various statistical problems, but it is also known to be burdened by slow convergence and sensitivity to initialization. In order to implement the expectation-maximization algorithm, Jiang and Zhang (2009) fixed a uniform grid within $[\min_{1\le i\le n} y_i, \max_{1\le i\le n} y_i]$ as support points and used the expectation-maximization algorithm to update the weights. Koenker and Mizera (2014) also used the discretization approach, and Koenker and Mizera (2014) utilize modern iterative convex optimization algorithm (specifically the interior point method) on the dual problem. Koenker and Mizera (2014) reported significant increase in speed and applicability, by comparison to the expectation-maximization algorithm based methods. Dicker and Zhao (2016) explored the approach of Koenker and Mizera (2014) in more details, and they recommend $\lfloor n^{1/2} \rfloor$ to be the number of grid points. Dicker and Zhao (2016) theoretically showed that with appropriately fine grid points, the approximated NPMLE has a squared Hellinger accuracy of $O((\log n)^2/n)$ under fixed design when the true mixing measure has compact support. For multiple dimensional Gaussian location mixtures model, Feng and Dicker (2018) recommended taking a regular grid within the compact region containing all data points.

Despite the recent advances in discretization based methods, gridding is generally problematic in multiple dimensions since the number of grid points grows exponentially as the dimension increases. Moreover, gridding would be inapplicable when the support K of the mixing measure cannot be reduced to a compact region safely, as in our mixture of regressions case. Other than gridding, another ad-hoc approximation approach is the "exemplar" method (Böhning, 2000; Lashkari and Golland, 2008; Saha and Guntuboyina, 2020). The exemplar method suggests a more daring approach that replaces a fine grid with n data points, which makes the computation more feasible for multiple dimensions. This approach is reasonable when the number of data points is sufficient but not too large. However, in order to apply the exemplar method in the mixture of regressions problem, one need to figure out what points to be chosen in the parameter space of β as the analog of data points in the estimation of mixture distributions. It is unclear how such candidate support points can be chosen directly from the data points $\{(x_i, y_i)\}_{i=1}^n$.

A direct consequence of Theorem 2.1 is stated as follows.

Corollary 2.1. *For any NPMLE defined as define* (1.2)*, the likelihood vector that it maps to is the unique optimal solution to* (2.1) *defined as follows.*

$$\begin{aligned} &\textit{maximize} \quad L(\mathrm{f}) = \frac{1}{n}\sum_{i=1}^{n} \log \mathrm{f}(i) \\ &\textit{subject to} \quad \mathrm{f} \in \mathrm{conv}(\mathcal{P}_K) \end{aligned} \tag{2.1}$$

where K represents either $\mathbb{R}^p$ or a compact set in $\mathbb{R}^p$.

Our computation procedure deals directly with the formulation in (2.1). We show that the Conditional Gradient Method (CGM) combining with a subproblem *oracle* can solve (2.1) with global convergence guarantee. We summarize the computation procedures in Algorithm 1, and Algorithm 2 is used to solve the re-optimization step in Algorithm 1. Conditional gradient method (CGM), also known

as the Frank-Wolfe method (Frank and Wolfe, 1956), is an iterative algorithm for constrained convex optimization, and has recently regained attention for its efficiency in solving modern large-scale data analysis problems (Jaggi, 2013). Originally from the literature in optimal design, vertex direction method (VDM, Wu (1978)) and vertex exchange method (VEM, Böhning (1986)) are two algorithms for computing mixture of distributions (Böhning, 2000). The two algorithms can be both viewed as applications of CGM and its *away-step* variant. However, VDM and VEM were previously considered to be numerically slow and limited, mostly due to the following two reasons. First, existing implementations use gridding strategies to discretize the parameter space of β. The fineness of the grid directly affects the accuracy of the final solution. When the set K is large or even unbounded, or when the dimension p gets larger, the gridding step severely limits the performances of VDM/VEM. As a result, Böhning (2000, Chapter 3.4) suggests using VEM only as a preliminary step for initializing the expectation-maximization algorithm. Second, at each iteration, after finding a new support point, VDM/VEM only moves from current iterate along the direction of the new atomic likelihood vector or the difference of new and some certain old atomic likelihood vector. As we will show later, fully optimization over all the likelihood vectors obtained in previous iterations can actually be solved efficiently via our Algorithm 2 or other common convex optimization algorithms, and therefore effectively increase the progress at each iteration. The existing algorithm that is probably the closet to ours is Mallet (1986), but the implementation in Mallet (1986) still depends on discretization. Nevertheless, the idea of finding new support points through the gradient characterization has long existed in the statistics literature (Lindsay, 1995, Chapter 5), and we stress our contribution lies in systematic algorithmic analysis and successfully leveraging modern optimization tools.

Algorithm 1: Conditional gradient method for NPMLE

Data: $\{(x_i, y_i)\}_{i=1}^n$

Input: σ, K

Result: $\mathrm{f}^{(T)} = \sum_{j=1}^{N_T} \pi_j^{(T)} \mathrm{g}_j^{(T)}$, $\sum_{j=1}^{N_T} \pi_j^{(T)} = 1$, $\pi_j^{(T)} > 0$, and number of iterations T

Initialization: likelihood vector $\mathrm{f}^{(0)} = \mathrm{f}^{\beta_0}$, active set $\mathcal{A}^{(0)} = \{\mathrm{f}^{\beta_0}\}$,

while *stopping criterion not met* **do**

  Approximately solving linear optimization subproblem: Find $\tilde{\mathrm{g}}^{(t)} \in \mathcal{P}_K$ s.t.

$$\langle \tilde{\mathrm{g}}^{(t)}, \nabla L(\mathrm{f}^{(t)}) \rangle \geq \max_{\mathrm{g} \in \mathcal{P}_K} \langle \mathrm{g}, \nabla L(\mathrm{f}^{(t)}) \rangle - \epsilon_s = \max_{\mathrm{g} \in \mathcal{A}} \sum_{i=1}^{n} \mathrm{g}(i)/\mathrm{f}^{(t)}(i) - \epsilon_s$$

  Adding the new vector: $\mathcal{A}^{(t+1)} = \mathcal{A}^{(t)} \cup \{\tilde{\mathrm{g}}^{(t)}\}$

  Re-optimization: $\mathrm{f}^{(t+1)} := \arg\max_{\mathrm{f} \in \mathrm{conv}(\mathcal{A}^{(t+1)})} L(\mathrm{f})$; `/* Use Algorithm 2 */`

  Updating active set: $\mathcal{A}^{(t+1)} = \{\mathrm{g}_j^{(t+1)} | \pi_j^{(t+1)} > 0\}$ for $\mathrm{f}^{(t+1)} = \sum_{i=1}^{N_{t+1}} \pi_j^{(t+1)} \mathrm{g}_j^{(t+1)}$

end

Instead of trying to discretize the parameter space at the beginning, we adaptively add one support points at every iteration. The new support point is obtained via solving a linear optimization subproblem approximately, and each support point maps to a n-dimensional atomic likelihood vector in the active set. We then re-optimize the objective function over the convex hull of all the atomic likelihood vectors in the active set to get the new iterate. In the previous theories on the VDM/VEM algorithms, only the convergence over the grid points is established. We study the algorithm in more details, and provide sharp algorithm convergence rates for CGM. Specifically, we show that CGM converges to the global maximum with rate $O(1/T)$ under mild conditions, where T is the number of iterations.

Algorithm 2: Classic conditional gradient method for re-optimization step

Data: $\{(x_i, y_i)\}_{i=1}^n$

Input: $\mathcal{A} = \{\mathrm{g}_j | j = 1, \ldots, N\}$, $\mathrm{f}^{(0)} = \sum_{j=1}^N \pi^{(0)} \mathrm{g}_j$

Result: $\mathrm{f}^{\mathrm{opt}} = \sum_{j=1}^N \pi_j \mathrm{g}_j$, $\sum_{j=1}^N \pi_j = 1$, $\pi_j > 0$

while $\mathrm{f}^{(l)}$ *not converge* **do**

  Exactly solving linear optimization subproblem: Find index j_l in $\{1, \ldots, N\}$ s.t.

$$\langle \mathrm{g}_{j_l}, \nabla L(\mathrm{f}^{(l)}) \rangle = \max_{\mathrm{g} \in \mathcal{A}} \langle \mathrm{g}, \nabla L(\mathrm{f}^{(j)}) \rangle = \max_{\mathrm{g} \in \mathcal{A}} \sum_{i=1}^n \mathrm{g}(i) / \mathrm{f}^{(j)}(i)$$

  Update: For $\gamma_l = \frac{2}{l+2}$, $\mathrm{f}^{(l+1)} = (1 - \gamma_l)\mathrm{h}^{(l)} + \gamma_l \mathrm{g}_{j_l}$ and $\pi_j^{(l+1)} = (1 - \gamma_l)\pi_j^{(l)} + \gamma_l \mathbf{1}\{j = j_l\}$, $j = 1, \ldots, N$

  Set: $l \leftarrow l + 1$

end

Now we discuss the convergence guarantee of Algorithm 1 in Theorem 2.2, and the mild assumptions are stated in Assumption 2.1.

Assumption 2.1. *There exists nonnegative scalar scalar e_s close to 0 and some positive δ such that for all iterations $t = 0, 1, \ldots, T$,*

$$\langle \tilde{\mathrm{g}}, \nabla L(\mathrm{f}^{(t)}) \rangle \geq \max_{\mathrm{g} \in \mathcal{P}_K} \langle \mathrm{g}, \nabla L(\mathrm{f}^{(t)}) \rangle - e_t, \tag{2.2}$$

and

$$\min_i \hat{\mathrm{f}}(i) \geq \delta, \min_i \tilde{\mathrm{g}}^{(t)}(i) \geq \delta. \tag{2.3}$$

In Assumption 2.1, (2.2) specifies the additive error level when approximately solving the linear optimization subproblem. While (2.3) might appear nonstandard in the literature of CGM, it holds trivially when K is compact. Moreover, it is reasonable for our specific problem even if K is unbounded for the following reasons: both $L(\cdot)$ and $\langle \cdot, \nabla L(\mathrm{f}^{(j)}) \rangle$ are monotonic with respect to each component of f, thus $\hat{\mathrm{f}}$ and $\tilde{\mathrm{g}}$ are unlikely to have very small components while $\hat{\mathrm{f}}$ is the maximizer of $L(\cdot)$ and $\tilde{\mathrm{g}}$ approximately maximizes $\langle \cdot, \nabla L(\mathrm{f}^{(j)}) \rangle$.

Theorem 2.2. *Under Assumption 2.1, $\gamma_t = \frac{2}{t+2}$ and $e_t = 2\gamma_t \epsilon / \delta^2$ for some $\epsilon > 0$, the iterate $\mathrm{f}^{(t)}$ satisfies*

$$L(\hat{\mathrm{f}}) - L(\mathrm{f}^{(t)}) \leq \frac{8}{t+2}(1 + \epsilon), \quad t = 0, 1, \ldots, T. \tag{2.4}$$

where $L(\hat{\mathrm{f}})$ is the maximum value of $L(\cdot)$ over $\mathcal{Q}_K$.

Theorem 2.2 quantifies the gap between the global optimum and current iterate at step t. We utilize the standard analysis (see, e.g., Jaggi (2013)) of first-order optimization algorithms and give a proof of Theorem 2.2 in Appendix B. The righthand side of (2.4) shows the accuracy of solution to the linear optimization subproblem and closeness to 0 of the density functions has impact on the convergence speed.

Proposition 2.1. *Under Assumption 2.1, the following inequalities hold.*

$$L(\hat{\mathrm{f}}) - L(\mathrm{f}^{(t)}) \leq \max_{\mathrm{g} \in \mathcal{P}_K} \langle \nabla L(\mathrm{f}^{(t)}), \mathrm{g} - \mathrm{f}^{(t)} \rangle \leq \langle \nabla L(\mathrm{f}^{(t)}), \tilde{\mathrm{g}} - \mathrm{f}^{(t)} \rangle + e_t. \tag{2.5}$$

Next, we give some detailed discussions on how CGM is tailored for our mixture of regressions model.

Linear optimization subproblem. By linearity, $\max_{\mathrm{g}\in\mathrm{conv}(\mathcal{P}_K)}\langle \mathrm{g}, \nabla L(\mathrm{f}^{(t)})\rangle = \max_{\mathrm{g}\in\mathcal{P}_K}\langle \mathrm{g}, \nabla L(\mathrm{f}^{(t)})\rangle$. If $\mathcal{P}_K$ were a finite set, this linear optimization subproblem becomes very easy to solve, as one can simply enumerate all points in $\mathcal{P}_K$. However, $\mathcal{P}_K$ contains uncountable points for typical choice of K, either a compact region in $\mathbb{R}^p$ or the whole space $\mathbb{R}^p$. Instead, we translate this optimization problem over $\mathcal{P}_K$ into a continuous optimization problem about variable β, i.e. $\max_{\beta\in K}\sum_{i=1}^n \frac{1}{\mathrm{f}^{(t)}(i)}\cdot\phi((y_i - x_i^\top\beta)/\sigma)$. Though this problem is non-convex with respect to β, it is relatively easy to get an approximate solution because the dimension is only p which does not scale with n. Moreover, it suffices to perform local search and approximately solve this subproblem to ensure the overall convergence of the algorithm, as stated in Theorem 2.2. In practice we find that the least square solution $(X^\top X)^{-1}X^\top y$ turns out to be a good initialization when solving the subproblem. Similar to most non-convex optimization problems, different initialization may lead to different local maximums. Therefore, we suggest using multiple random initialization and choosing the best solution from the multiple runs. For fast computing speed, we adopt the off-the-shelf solver from the Python `scipy` package. We use a specific Python solver to call the Powell's conjugate direction method (Powell, 1964). Powell's method is an effective method for finding local optimum unconstrained optimization. R function with similar functionality is available at R package `mize` (Melville, 2019). In case where K is not $\mathbb{R}^p$ and the returned solution from the non-convex solver is outside K, we restart this step by a new random initialization until a solution within set K is found.

The subproblem solving step is the only step in Algorithm 1 that directly evaluates the intrinsic structure of f as a function of β. On the other hand, all other steps only deal with the high-level computation based on n-dimensional vectors. It is not hard to see that, we can replace $x^\top\beta$ by any regression function $r(x,\beta)$ here in order to solve a general class of mixture of regressions model. More specifically, the optimization problem at this step becomes $\max_{\beta\in K}\sum_{i=1}^n \frac{1}{\mathrm{f}^{(t)}(i)}\cdot\phi((y_i - r(x,\beta))/\sigma)$. Later in Section 4, our simulations show that this approach indeed works for nonlinear regression models too.

Re-optimization step. At iteration t, the re-optimization step amounts to maximizing a concave objective over a convex hull of at most t points in $\mathbb{R}^n$. Since the convex hull is generated from finite points, this convex optimization problem is easy to solve numerically. One can use several standard convex optimization approaches here. We point out that many other methods, for example interior point method and first-order methods, can also be used. Such rich tools for convex optimization are available via the `Rmosek` package (Friberg, 2019) in R. For the sake of clarity, we implement a classic CGM for the re-optimization step as described in Algorithm 2. Algorithm 2 is very similar to Algorithm 1, and the major difference is that the linear optimization step in Algorithm 2 can be solved exactly by enumeration. Noticeably, the re-optimization step guarantees that the CGM iteration keeps increasing the likelihood. We use the classic CGM with canonical pre-defined step-sizes $\gamma_l = \frac{2}{l+2}$ to get the optimal solution over the convex hull of the current active set. Since this re-optimization step is conducted over a set supported on finite sets, it is very efficient to solve.

Active set update. Since we only add one atomic likelihood vector into the active set at each iteration, the cardinality of $\mathcal{A}^{(t)}$ is at most t. However, as noted in Jaggi (2013) , we observe that the re-optimization step enables $\mathrm{f}^{(t)}$ to be a relatively sparse representation of vectors in $\mathcal{A}^{(t)}$. Thus the cardinality of $\mathcal{A}^{(t)}$ is effectively controlled after each update, and does not scale linearly with respect to number of iterations in practice.

Stopping criterion. The convexity of the optimization problem (2.1) leads to a computable bound on the optimality gap of the current iterate. As stated in Proposition 2.1, we can calculate $\langle \nabla L(\mathrm{f}^{(t)}), \tilde{\mathrm{g}} - \mathrm{f}^{(t)}\rangle$ at each iteration and stop the algorithm when this scalar is smaller than a certain threshold. In practice, the number of iterations can also be chosen manually when new iterates stop making significant increase in the likelihood objective. We will discuss these in more details in our simulations.

3 Bounds on prediction error

In this section, we prove bounds for the prediction error of the NPMLE $\hat{G}$ (defined in (1.2)). According to our mixture of regressions model, the conditional distribution of the response variable y for a regressor x has the density $y \mapsto f_x^{G^*}(y)$ which will be estimated by $y \mapsto f_x^{\hat{G}}(y)$. We shall use the squared Hellinger distance (1.4) to measure the discrepancy between the densities $f_x^{G^*}(\cdot)$ and $f_x^{\hat{G}}(\cdot)$. To get our overall loss function, we separately study the two cases of fixed design and random design. In the fixed design setting which we study first, we take the average of (1.4) over the possible values $x_1, \ldots, x_n$ of x to obtain the loss function $\mathfrak{H}_n^2(f^{\hat{G}}, f^{G^*})$ that is defined in (1.5). In the following theorem, we shall give a bound $\mathfrak{H}_n^2(f^{\hat{G}}, f^{G^*})$ that holds in high probability and in expectation. The bound is parametric up to logarithmic multiplicative factors in n.

Theorem 3.1 (Fixed Design Prediction Error Result)**.** *Consider data $(x_1, y_1), \ldots, (x_n, y_n)$ with $n \geq 3$ where $x_1, \ldots, x_n$ are fixed design points and $y_1, \ldots, y_n$ are independent with y_i having the density*

$$y \mapsto f_{x_i}^{G^*}(y) = \int \frac{1}{\sigma} \phi\left(\frac{y - x_i^\top \beta}{\sigma}\right) \mathrm{d}G^*(\beta) \qquad \text{for } i = 1, \ldots, n. \tag{3.1}$$

Assume that $\max_{1 \leq i \leq n} \|x_i\| \leq B$ and $G^* \{\beta \in \mathbb{R}^p : \|\beta\| \leq R\} = 1$ for some $B > 0$ and $R > 0$. Let $\epsilon_n = \epsilon_n(B, R, \sigma)$ be defined via

$$\epsilon_n^2 := n^{-1} \max\left(\left(\mathrm{Log}\ \frac{n}{\sqrt{\sigma}}\right)^{p+1}, \left(\frac{RB}{\sigma}\right)^p \left(\mathrm{Log}\ \left\{\frac{n}{\sqrt{\sigma}}\left(\frac{\sigma}{RB}\right)^p\right\}\right)^{\frac{p}{2}+1}\right), \tag{3.2}$$

where we used $\mathrm{Log}\ x := \max(\log x, 1)$. *Then there exists a constant C_p depending only on p such that*

$$\mathbb{P}\left\{\mathfrak{H}_n(f^{\hat{G}}, f^{G^*}) \geq t\epsilon_n \sqrt{C_p}\right\} \leq \exp(-nt^2\epsilon_n^2) \qquad \text{for every } t \geq 1, \tag{3.3}$$

and

$$\mathbb{E}\mathfrak{H}_n^2(f^{\hat{G}}, f^{G^*}) \leq C_p \epsilon_n^2. \tag{3.4}$$

The error ϵ_n defined via (3.2) clearly satisfies $\epsilon_n^2 = O(n^{-1}(\log n)^{p+1})$ as $n \to \infty$ (keeping R, B, σ fixed) and thus ϵ_n^2 gives the usual parametric rate n^{-1} up to a logarithmic multiplicative factor.

For the proof of Theorem 3.1, we apply a framework of developing a large deviation equality for the Helinger distances, and also prove some new metric entropy results in aid of the proof. The function class $\mathcal{M}_R$ lying at the heart of the proof of Theorem 3.1 is

$$\mathcal{M}_R = \{f_x^G(y) : \text{ any probability measure } G \text{ supported on } \mathrm{B}_p(0, R)\}, \tag{3.5}$$

where $\mathrm{B}_p(0, R) := \{\beta \in \mathbb{R}^p : \|\beta\| \leq R\}$. We critically use a covering of $\mathcal{M}_R$ under an appropriately defined metric to construct inequalities. The proved bound relies on the size of the covering number, whose logarithm is known as the metric entropy. Intuitively, the metric entropy quantifies the complexity of the density function class, and also captures the difficulty level of the estimation problem.

Mixture density function classes were also investigated in the proofs of prediction error in the Gaussian mixture distribution case (Jiang and Zhang, 2009; Saha and Guntuboyina, 2020). In our regression case, as defined in (3.7), every density function $f_x^G(y)$ in the class $\mathcal{M}_R$ is a function of both the regressor variable $x \in \mathbb{R}^p$ and the response variable $y \in \mathbb{R}$. In contrast, no regressor variable was included in the distribution case of Jiang and Zhang (2009) and Saha and Guntuboyina (2020). Therefore, the generalization of metric entropy results from the Gaussian mixture distribution to the mixture of regressions is

non-trivial. To tackle the difficulty of generalizing to the regression case, we design an appropriate metric over this new function classes and prove sharp bounds on the metric entropy. Specifically, we define the metric as the L_∞ metric of density functions over a set containing $\{x_i \times [-R\|x_i\| - M, R\|x_i\| + M]\}_{i=1}^n$ in proving Theorem 3.1. These new results on the metric entropy are critical in the proof of Theorem 3.1 and subsequent results explained later in Section 3. The main quantity ϵ_n^2 in the upper bound depends on the Lipschitz constant $\mathfrak{L}$ and the largest norm of parameters R, which is crucially different from Jiang and Zhang (2009) and Saha and Guntuboyina (2020).

We make a few comments on the two assumptions in Theorem 3.1. The first assumption requires the design points $\{x_i\}_{i=1}^n$ are bounded, which is quite standard in the literature. The second assumption requires that G is also bounded. Although this bounded assumption is not necessary in our existence results and computation procedures in Section 2, it turns out boundedness of G is indispensable to controlling the complexity of the density function class $\mathcal{M}_R$ defined in (3.5).

In the end of this section on fixed design, we point out that our proof strategies for prediction error work with more complicated regression models than linear regressions. In the following Theorem 3.2, we generalize Theorem 3.1 through replacing the linear function $x^\top \beta$ by a function $r(x, \beta)$. The key reason that $r(x, \beta)$ needs to be polynomial with respect to β comes from the assumptions necessary to the metric entropy bound in Theorem E.1.

For any x, we denote $\mathfrak{L}(x_i)$ as a Lipschitz constant for $r(x, \beta)$ as a function of β. For each x, let $\mathfrak{L}(x)$ be defined as

$$\mathfrak{L}(x) := \sup_{\beta_1, \beta_2 \in K : \beta_1 \neq \beta_2} \frac{|r(x, \beta_1) - r(x, \beta_2)|}{\|\beta_1 - \beta_2\|} \tag{3.6}$$

Therefore, $\sup_{1 \leq i \leq n} \mathfrak{L}(x_i)$ controls how much the regression function changes as the regression parameter changes under fixed design $\{x_i\}_{i=1}^n$. For example, in the linear regressions, we can simply choose $\|x_i\|$ as the Lipschitz constant of function $x_i^\top \beta$. In other words, Theorem 3.1 belongs to the $\zeta = 1$ case of Theorem 3.2.

Theorem 3.2. *Assuming $r(x, \beta)$ is polynomial function such that $r(0, \beta) = 0$ for all β, and the degree of $r(x, \beta)$ is at most ζ with respect to β. We assume that the maximum of n Lipschitz constants $\mathfrak{L}(x_i)$ is bounded and denoted by $\mathfrak{L} = \max_{1 \leq i \leq n} \mathfrak{L}(x_i)$.*

Then the same results in (3.3) *and* (3.4) *hold with ϵ_n^2 being replaced by*

$$\epsilon_n^2 = n^{-1} \left(\zeta^p \left(\text{Log } \frac{n}{\sqrt{\sigma} \zeta^p} \right)^{p+1}, \left(\frac{\zeta R \mathfrak{L}}{\sigma} \right)^p \left(\text{Log } \left\{ \frac{n}{\sqrt{\sigma}} \left(\frac{\sigma}{\zeta R \mathfrak{L}} \right)^p \right\} \right)^{\frac{p}{2}+1} \right).$$

3.1 Random design

In this section, we consider the random design where $x_1, \ldots, x_n$ are generated from an unknown probability distribution μ. Given data points $\{(x_i, y_i)\}_{i=1}^n$, the nonparametric maximum likelihood estimator is defined in the same way as in the fixed design. Note that μ is not necessary in neither the definition nor the computation of nonparametric maximum likelihood estimators. To evaluate the performance of nonparametric maximum likelihood estimators under random design, we take the average of Hellinger distance over the distribution μ. The averaged prediction error is defined as

$$\bar{\mathfrak{H}}_n \left(\tilde{f}, f^* \right) = \int \mathfrak{H} \left(\tilde{f}_x, f_x^* \right) \mathrm{d}\mu(x).$$

We choose to work with the Hellinger distance rather than squared Hellinger distance due to technical reasons. Our first main result under random design is Theorem 3.3. We show a parametric bound (up

to logarithmic multiplicative factors) on $\bar{\mathfrak{H}}_n^2\left(f^{\hat{G}}, f^*\right)$ that holds with high probability under some mild assumptions. The assumptions on G^* and x_i in Theorem 3.3 are very similar to those in Theorem 3.1.

Theorem 3.3. *Consider data $(x_1, y_1), \ldots, (x_n, y_n)$ with $n \geq 3$ where $x_1, x_2, \ldots, x_n$ are independently generated from some probability measure μ, and $y_1, \ldots, y_n$ are independent with y_i having the density*

$$y_i \mid x_i \mapsto f_{x_i}^*(y_i) = \int \frac{1}{\sigma}\phi\left(\frac{y_i - x_i^\top \beta}{\sigma}\right) \mathrm{d}G^*(\beta) \qquad \text{for } i = 1, \ldots, n. \tag{3.7}$$

Assume that $\mu\{x \in \mathbb{R}^p : \|x\| \leq B\} = 1$ and $G^*\{\beta \in \mathbb{R}^p : \|\beta\| \leq R\} = 1$ for some $B > 0$ and $R > 0$. There exist universal constants C_p and C_p' depending only on p so that

$$\int \mathfrak{H}(f^{\hat{G}}, f^*)\mathrm{d}\mu(x) \leq C_p\epsilon_n + \frac{\rho_p(B, R, \sigma)}{\sqrt{n}} + \frac{2(\log n)^{1/2}}{\sqrt{n}} \tag{3.8}$$

with probability at least $1 - 2n^{-1}$, where ϵ_n is defined as in (3.2) *and*

$$\begin{aligned}&\rho_p(B, R, \sigma)\\ &= \max((e\sigma)^{-1/4})/2, 1) \int_0^{\min(1, 2(e\sigma)^{-1/4})} \sqrt{1 + C_p'\left(1 + \frac{2RB}{\{2\log(48\sigma^{-1}\epsilon^{-4})\}^{1/2}\sigma}\right)^p \left(\log\frac{16}{\sigma\epsilon^4}\right)^{p+1}}\,\mathrm{d}\epsilon.\end{aligned}$$

Theorem 3.3 is a nearly parametric bound on the prediction error under random design. The proof of Theorem 3.3 relies on the bound on expected supremum of empirical processes using the bracketing number. To control the bracketing integral $\rho(B, R, p)$ in (3.8), we again use the metric entropy results established in Appendix E.

Next, we shall show that nonparametric maximum likelihood estimators are weakly consistent in the sense that the . We invoke a slightly stricter assumption on the distribution μ in Theorem 3.4.

Theorem 3.4. *For the mixture of linear regressions model, let $\hat{G}_n$ be the NPMLE solution from n sample points $\{(x_i, y_i)\}_{i=1}^n$. Let μ be the probability measure of x_i, i.e. x_i are independently distributed as μ, and the support of probability measure μ is S_0. Assuming that S_0 is compact and contains an open set in $\mathbb{R}^p$, then*

$$d(\hat{G}_n, G^*) \to 0 \text{ in probability.}$$

Here d is the distance metric induced by the L_2 norm on the characteristic functions.

To prove Theorem 3.4, we adopt techniques from Beran and Millar (1994) by first showing strong identifiability in Lemma D.2 and then considering the characteristic function of the ground truth and that of the estimation. Our previous result in Theorem 3.3 also provides key ingredients for proving Theorem 3.4.

4 Numerical experiments in simulated settings

4.1 Overview of simulations

Since the simulation results are best visualized when regressors are 2-dimensional with a dimension being constant, we will illustrate our computational procedures under such settings. The remainder of this section is divided into three parts. Section 4.2 contains simulations on the case that G^* is supported on finite points, which we refer to as discrete mixtures. The fitted predictive distribution

$f_x^{\hat{G}}(y)$ is compared to the ground truth, and demonstrates remarkable accuracy when compared to the ground truth. Section 4.3 contains experiments on the case that G^* is not a discrete probability measure. Specifically, we let G^* be a mixture of two Gaussians. The visualization of data points when G^* is mixture of two Gaussians is close to the case when G is supported on only two points. Therefore one may be tempted to fit with two-component mixtures using the expectation-maximization algorithm. However, we show that such misspecification leads to deficiency in terms of predictions error. In Section 4.4, we go beyond linear regression and implement our computational procedure on the mixture of nonlinear regression models, including polynomial functions and exponential functions.

4.2 Discrete mixture

Data set. Given parameters $(J, \beta^1, \ldots, \beta^J, \pi_1, \ldots, \pi_J, \sigma)$, each data point (x_i, y_i) is generated as follows: (i) a random number z_i is generated from Unif$(0, 1)$, and then the j-th regression model is selected if $z_i \in [\sum_{l=1}^{j-1} \pi_l, \sum_{l=1}^{j} \pi_l]$. (ii) x_i is randomly generated from Unif$[-1, 3]$. (iii) $y_i = \beta_1^j + \beta_2^j x_i + \epsilon_{ij}$, where $\epsilon_{ij} \sim N(0, \sigma^2)$ is noise. 300 or 500 data points are generated for each simulation.

Implementation. CGM is run for $10 \sim 20$ iterations with known σ. The initialization β^0 is set as the ordinary least square solution.

Outputs. CGM returns a fitted mixture of linear regressions with parameters

$$(\hat{J}, \hat{\beta}^1, \ldots, \hat{\beta}^{\hat{J}}, \hat{\pi}_1, \ldots, \hat{\pi}_{\hat{J}}, \sigma).$$

Plots of fitted mixtures. The fitted components with mixing proportions of at least $p = 0.02$ are called *dominant components*. Only dominant components are plotted and only 3 digits of the corresponding mixing proportions are kept. The width and darkness of each line in the plots are proportional to its mixing proportion respectively. The i-th data point is colored the same as component j if the posterior probability is the largest among all dominant components (which, in most cases, is also the largest among all components since non-dominant components have very low mixing proportions), namely

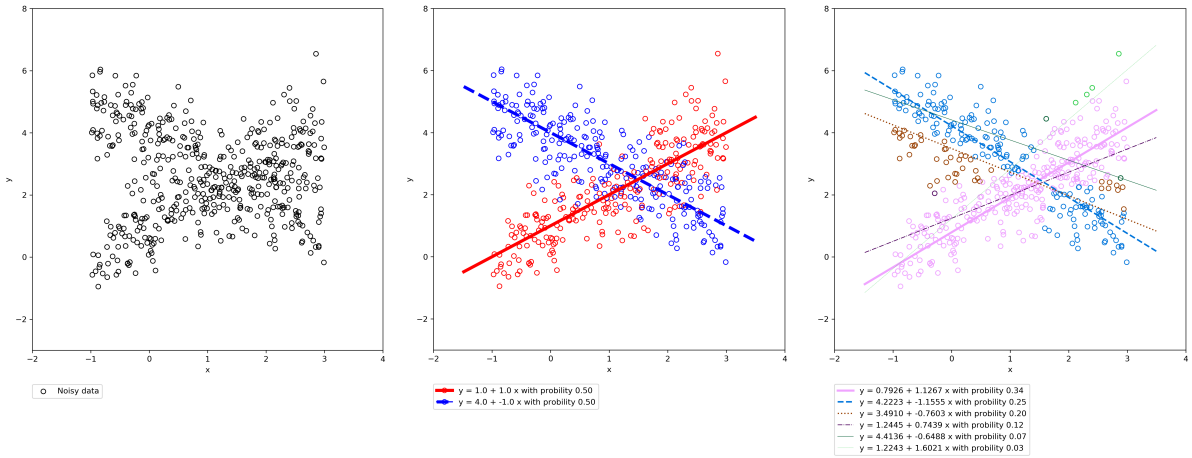
$$j \in \arg \max_{1 \le l \le \hat{J}, \hat{\pi}_l \ge p} \hat{\pi}_l \phi \left(\frac{y_i - (\hat{\beta}_1^l + \hat{\beta}_2^l x_i)}{\sigma} \right). \tag{4.1}$$

The density functions of y at fixed x values for both true mixtures and fitted mixtures are also plotted.
(1) Two-component concurrent mixture: $\sigma = 0.8$. Figure 1 shows two components with equal proportions.
(2) Two-component parallel mixture: $\sigma = 0.3$. Results are in Figure 2.
(3) Three-component mixture: $\sigma = 0.5$. Results are in Figure 3.

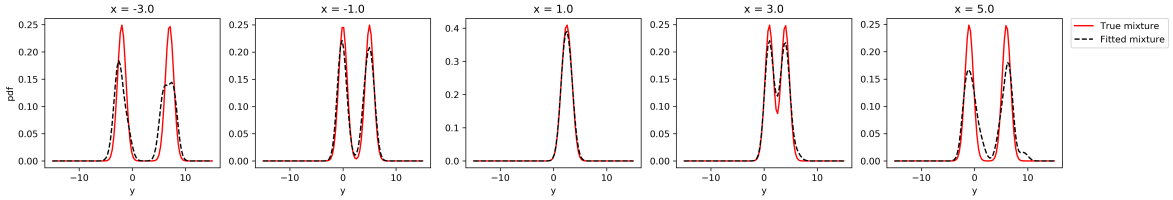
4.3 Non-discrete mixture

Data set. Given the probability measure function $G(\beta)$ and σ, each data point is generated as follows: (i) x_i is randomly sampled from Unif$(-3, 3)$. (ii) β^i is randomly sampled from $G(\beta)$. (iii) $y_i = \beta_1^i + \beta_2^i x_i + \epsilon_i$, where $\epsilon_{ij} \sim N(0, \sigma^2)$ is noise and $\sigma = 0.5$. We set $G(\beta)$ as a mixture of two Gaussians,

$$\begin{aligned}
\gamma_1 \sim & N\left(\begin{pmatrix} 1 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 0.5 & 0.2 \\ 0.2 & 0.3 \end{pmatrix} \right) \\
\gamma_2 \sim & N\left(\begin{pmatrix} 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 0.5 & 0.2 \\ 0.2 & 0.3 \end{pmatrix} \right) \\
\beta = & (1 - \Delta) \cdot \gamma_1 + \Delta \cdot \gamma_2
\end{aligned}$$

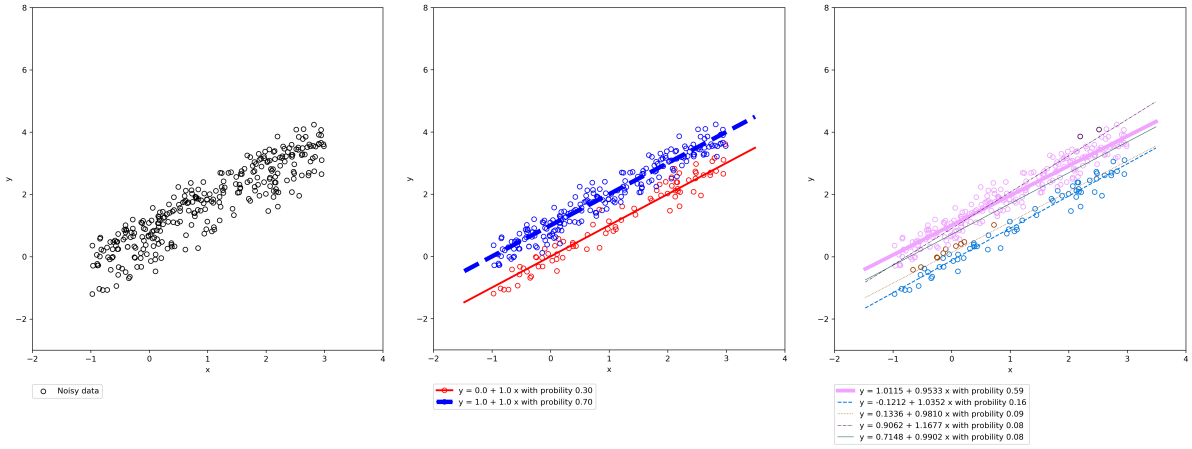


(a) Left: Noisy data; Middle: True mixture; Right: Fitted mixture

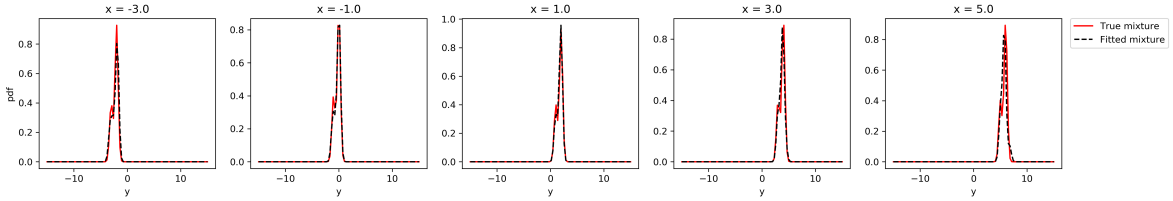


(b) True and fitted probability density functions (pdf) of y at different x's

Figure 1: Two-component concurrent mixture of linear regressions

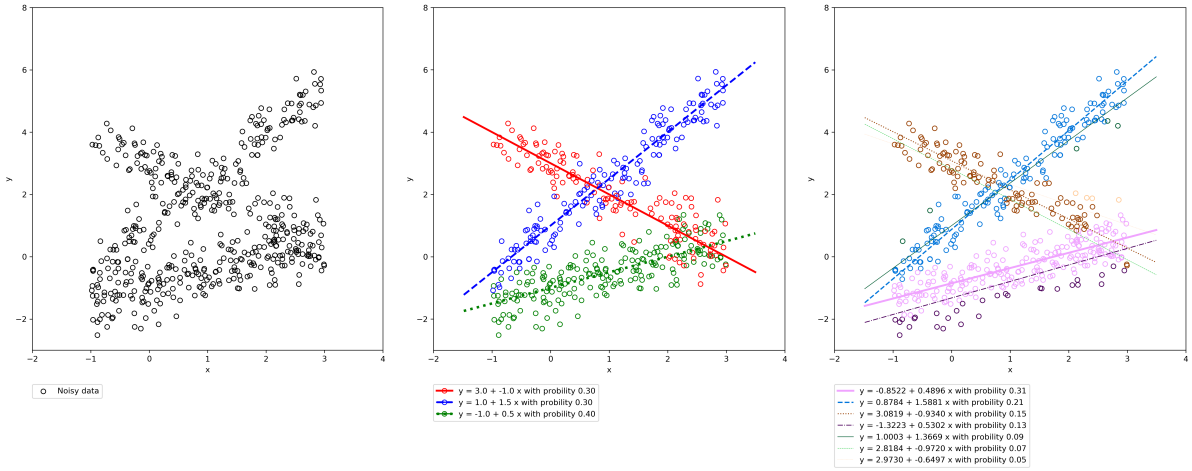


(a) Left: Noisy data; Middle: True mixture; Right: Fitted mixture

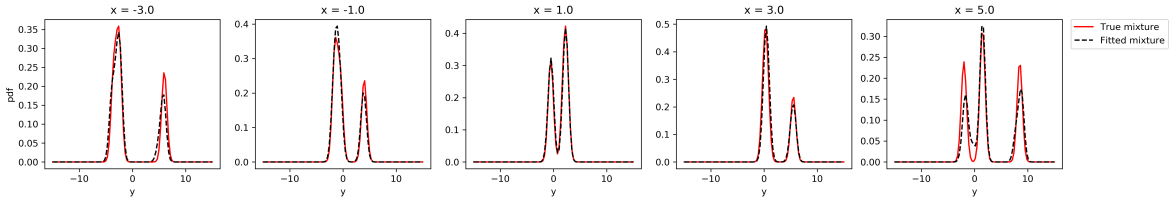


(b) True and fitted probability density functions (pdf) of y at different x's

Figure 2: Two-component parallel mixture of linear regressions



(a) Left: Noisy data; Middle: True mixture; Right: Fitted mixture



(b) True and fitted probability density functions (pdf) of y at different x's

Figure 3: Three-component mixture of linear regressions

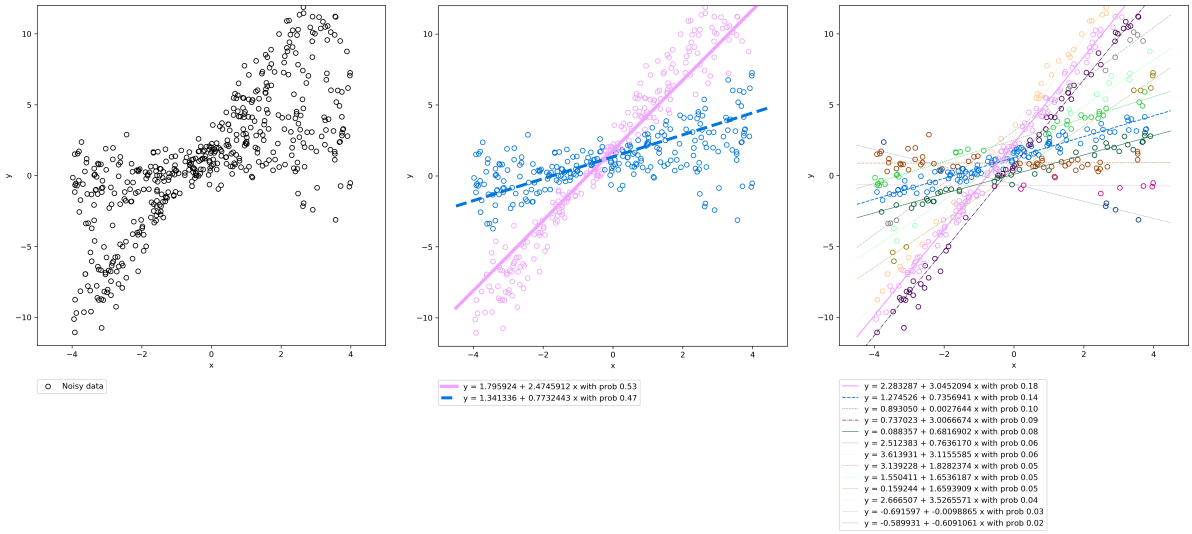
where $\Delta \in \{0, 1\}$ with $\mathbb{P}(\Delta = 1) = 0.5$.

Implementation. In this case, we recommend using randomized initialization when solving the non-convex subproblem for better convergence performance. More specifically, at iteration t of CGM, the subproblem on line 3 of Algorithm 1 is solved with initialization at g^{β_t} with β_t randomly sampled from Unif$[-10, 10]$. We choose x from $(-4, 4)$ as well as one value 5.0 outside the interval $(-4, 4)$ to evalute their generalization accuracy. Our implementation of the expectation-maximization algorithm follows the standard form in Faria and Soromenho (2010). To be fair in the comparsion, we feed the expectation-maximization algorithm with the true σ value, and β is also initialized randomly by sampling uniformly from $[-10, 10]^2$ in the expectation-maximization algorithm.

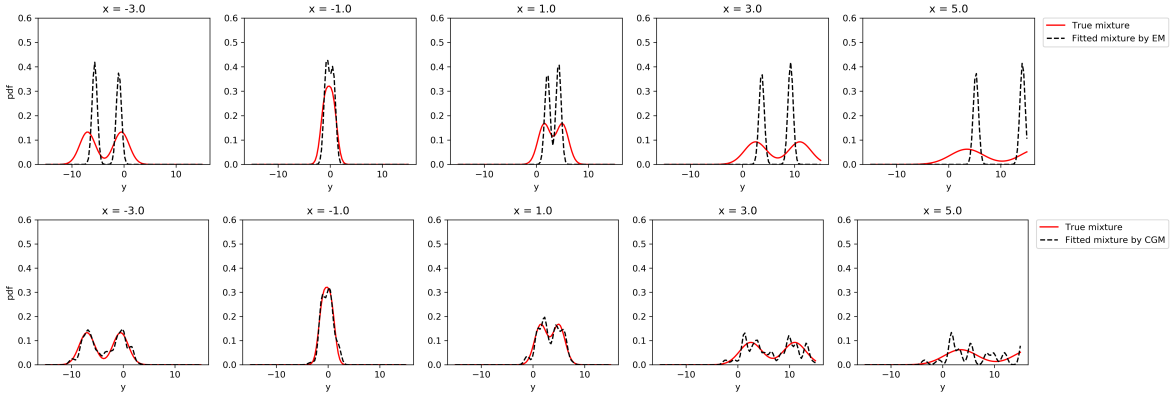
Plots. The results are shown in Figure 4. It is remarkable that G is actually not compact here, but the algorithm achieves excellent accuracy. Specifically, in Figure 4(b), the fitted density function is much more close to the ground truth compared to the two-component mixture fitted by the expectation-maximization algorithm.

4.4 Nonlinear regression

Data set. (a) Polynomial regressions: Given the probability measure function $G(\beta)$ and σ, each data point is generated as follows: (i) x_i is randomly sampled from Unif$(-1, 1.5)$. (ii) β^i is randomly sampled from $G(\beta)$. (iii) $y_i = \beta_1^i(\beta_1^i + \beta_2^i) + 0.5\beta_1^i\beta_2^i x_i + 0.5\beta_1^i x_i^2 + \epsilon_i$, where $\epsilon_{ij} \sim N(0, \sigma^2)$ is noise and $\sigma = 0.5$. We set G as a two-component mixture here with $\beta = (0.5, 2)$ with probability 0.5 and $\beta = (1, 2.5)$ with probability 0.5. (b) Exponential regression: Given the probability measure function $G(\beta)$ and σ, each data point is generated as follows: (i) x_i is randomly sampled from Unif$(-1, 3)$. (ii) β^i is randomly sampled from $G(\beta)$. (iii) $y_i = \beta_1^i + \exp(-\beta_2^i x)$, where $\epsilon_{ij} \sim N(0, \sigma^2)$ is noise and $\sigma = 0.5$. We set



(a) Left: Noisy data; Middle: Fitted mixture by two-component EM; Right: Fitted mixture by CGM



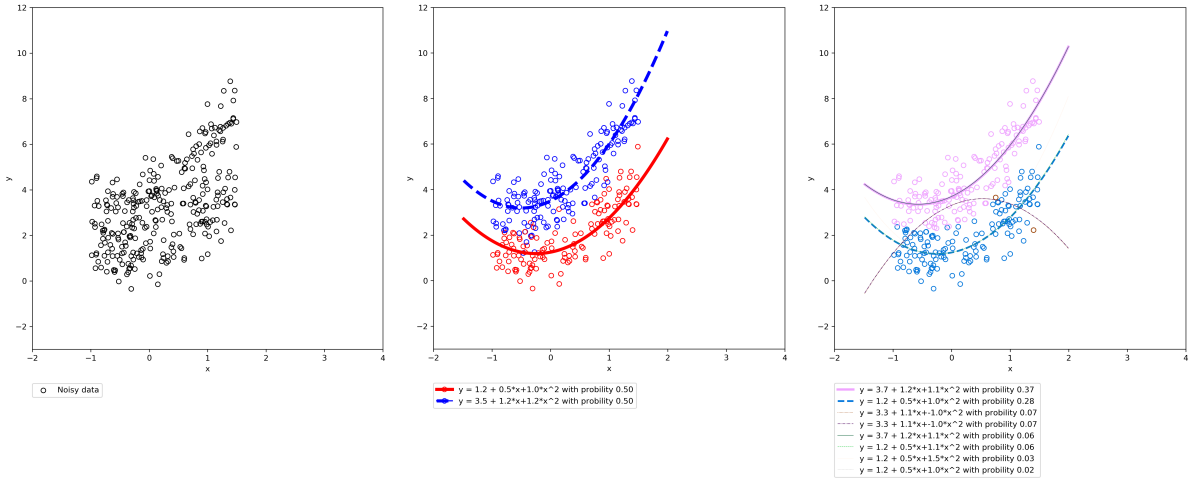
(b) True and fitted probability density functions (pdf) of y at different x's: Two-component EM (top) and CGM (bottom)

Figure 4: Non-discrete mixture

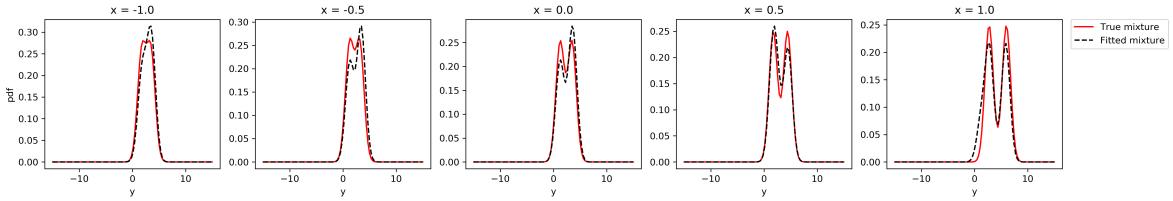
G as a two-component mixture here with $\beta = (-0.5, 1)$ with probability 0.5 and $\beta = (-1.5, 1.5)$ with probability 0.5.

Implementation. We use randomized initialization when solving the non-convex subproblem for better convergence performance. More specifically, at iteration t of CGM, the subproblem on line 3 is solved with initialization at g^{β_t} with β_t randomly sampled from Unif$[-10, 10]$.

Plots. The results are shown in Figure 4 and 6.



(a) Left: Noisy data; Middle: True mixture; Right: Fitted mixture



(b) True and fitted probability density functions at different x's

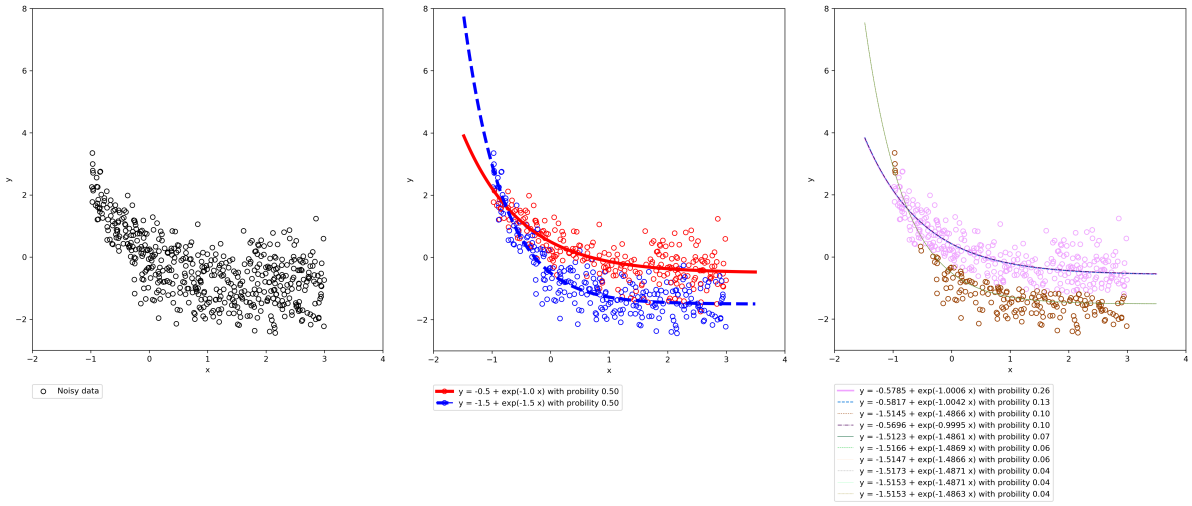
Figure 5: Mixture of polynomial regressions

In our numerical experiments, the number of simulations T (10 to 100) is usually much smaller than the number of data points n (300 to 500), so the number of support points never exceed the theoretical characterization upper bound n. However, an index limitation step (Mallet, 1986, Section 4.2) can be enforced to replace any iteration with more than n support points by a new iteration with at most n support points.

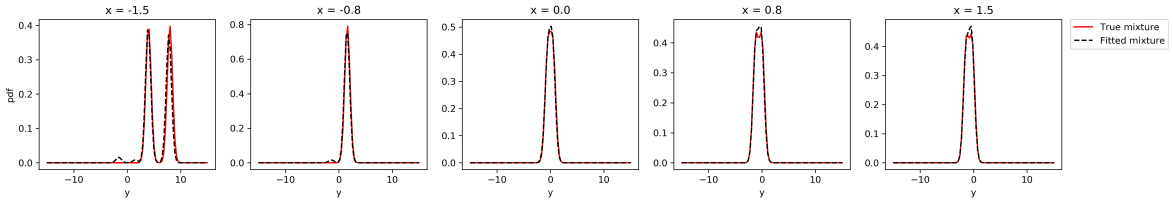
5 Numerical experiments on real data

5.1 Tone perception data

The tone perception data originally comes from experiments by Cohen (1980) and has been analyzed in the statistics literature by De Veaux (1989), Viele and Tong (2002) and Yao and Song (2015). A trained musician is presented with a pure fundamental tone plus a series of stretched overtones, and the experiment is repeated a few times with different tones. The regressor value x is the stretching ratio of the overtone to the fundamental tone. The musician is then asked to tune an adjustable tone to the octave above the fundamental tone. The response value y is the ratio of the adjusted tone to the fundamental. The experiment is designed to verify if either of the two existing theories of music



(a) Left: Noisy data; Middle: True mixture; Right: Fitted mixture.



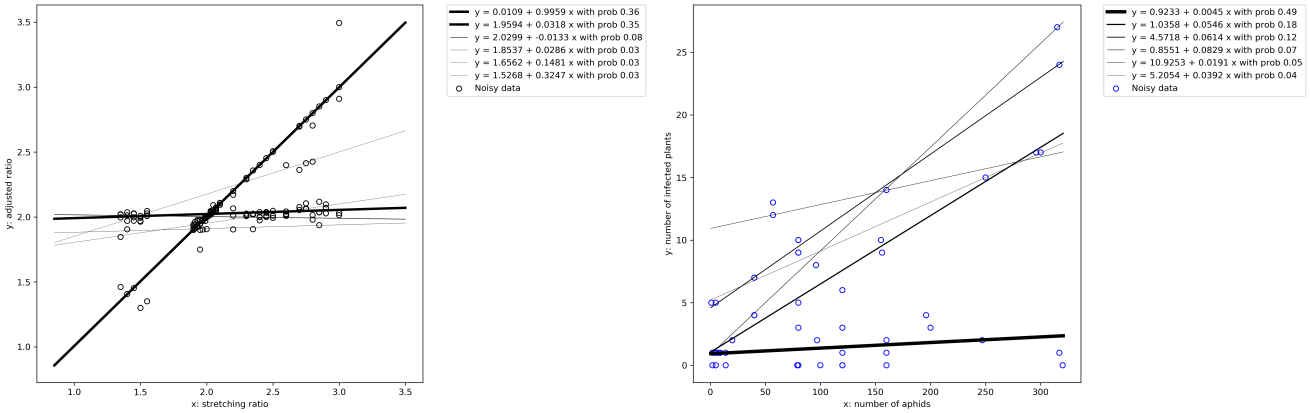
(b) True and fitted probability density functions at different x's

Figure 6: Mixture of exponential regressions

perception is valid: one theory states that the adjusted tone is at ratio $2:1$ to the fundamental tone, while the other theory states that the adjusted tone will be equal to the overtone. In the language of linear regression, two existing theories correspond to $y = 2$ and $y = x$ respectively. The data is available in R package `mixtools` (Benaglia et al., 2009) containing 150 data points from one musician. The result of fitting data with mixture of linear regressionss is shown in Figure 7a. The noise level σ is chosen as 0.2.

5.2 Aphids data

The aphids data was firstly analyzed in Boiteau et al. (1998). The data comes from the study on spreading of tobacco plants virus by aphids. The regressor value x is the number of aphids that were released in a closed chamber containing 12 infected and 69 healthy tobacco plants. The response value y is the number of infected plants amongst those previously healthy after 24 hours. The data is currently available in R package `mixreg` (Boiteau et al., 1998) containing 51 independent experiments. The result of fitting data with mixture of linear regressionss is shown in Figure 7b. The noise level σ is chosen as 0.2.



(a) Fitting music tone perception data

(b) Fitting aphids data

Figure 7: Real data experiments

6 Conclusions

We propose to fit mixture of linear regressions with the nonparametric maximum likelihood estimators. Concretely, we provide both algorithmic advances and detailed theoretical analysis. It is of interest to establish the error bound for other sorts of mixture of regressions model as well, such as multivariate linear regressions and logistic regressions. In multivariate linear regression, the response variable is generalized to multiple dimensions. We expect our proof techniques and computing methods can still be useful in that domain. Our study focuses on the regime where p is small compared to n. More structural assumptions of G are probably needed to make the nonparametric maximum likelihood estimator work in the high dimensional case. In that case, it is interesting to propose and analyze the appropriate modifications of nonparametric maximum likelihood estimators under some sparsity constraints on β. We leave these as important future works.

References

Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

Tatiana Benaglia, Didier Chauveau, David R. Hunter, and Derek Young. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29, 2009. URL `http://www.jstatsoft.org/v32/i06/`.

Rudi Beran and P Warwick Millar. Minimum distance estimation in random coefficient regression models. *The Annals of Statistics*, 22(4):1976–1992, 1994.

Rudolf Beran, Andrey Feuerverger, and Peter Hall. On nonparametric estimation of intercept and slope distributions in random coefficient regression. *The Annals of Statistics*, 24(6):2569–2592, 1996.

Dimitri P Bertsekas, Angelia Nedi, and Asuman E Ozdaglar. Convex analysis and optimization. 2003.

D Böhning. A vertex-exchange-method in d-optimal design theory. *Metrika*, 33(1):337–347, 1986.

Dankmar Böhning. *Computer-assisted analysis of mixtures and applications*. Taylor & Francis Group, 2000.

G Boiteau, M Singh, RP Singh, GCC Tai, and TR Turner. Rate of spread of pvy n by alatemyzus persicae (sulzer) from infected to healthy plants under laboratory conditions. *Potato Research*, 41(4): 335–344, 1998.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

T Tony Cai, Jing Ma, and Linjun Zhang. Chime: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267, 2019.

E Cohen. Inharmonic tone perception. *Unpublished Ph. D. Dissertation, Stanford University*, 1980.

Richard D De Veaux. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8(3): 227–245, 1989.

Lee H Dicker and Sihai D Zhao. High-dimensional classification via nonparametric empirical bayes and maximum likelihood inference. *Biometrika*, 103(1):21–34, 2016.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Susana Faria and Gilda Soromenho. Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2):201–225, 2010.

Long Feng and Lee H Dicker. Approximate nonparametric maximum likelihood for mixture models: A convex optimization approach to fitting arbitrary multivariate mixing distributions. *Computational Statistics & Data Analysis*, 122:80–91, 2018.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Henrik A. Friberg. *The R to MOSEK Optimization Interface*, 2019. URL `https://cran.r-project.org/web/packages/Rmosek/index.html`. R package version 1.3.5).

Subhashis Ghosal and Aad Van Der Vaart. Posterior convergence rates of dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35(2):697–723, 2007.

Subhashis Ghosal and Aad W Van Der Vaart. Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29(5):1233–1263, 2001.

Stefan Hoderlein, Jussi Klemelä, and Enno Mammen. Analyzing the random coefficient model nonparametrically. *Econometric Theory*, 26(3):804–837, 2010.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of The 30th International Conference on Machine Learning*, volume 28, pages 427–435. Curran, 2013.

Wenhua Jiang and Cun-Hui Zhang. General maximum likelihood empirical bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684, 2009.

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

Hiroyuki Kasahara and Katsumi Shimotsu. Testing the number of components in normal mixture regression models. *Journal of the American Statistical Association*, 110(512):1632–1645, 2015.

Jack Kiefer and Jacob Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906, 1956.

Roger Koenker and Ivan Mizera. Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association*, 109(506):674–685, 2014.

Jeongyeol Kwon, Wei Qian, Constantine Caramanis, Yudong Chen, and Damek Davis. Global convergence of the em algorithm for mixtures of two component linear regression. In *Conference on Learning Theory*, pages 2055–2110, 2019.

Tze Leung Lai and Mei-Chiung Shih. Nonparametric estimation in nonlinear mixed effects models. *Biometrika*, 90(1):1–13, 2003.

Nan Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.

Danial Lashkari and Polina Golland. Convex clustering with exemplar-based models. In *Advances in neural information processing systems*, pages 825–832, 2008.

Yuanzhi Li and Yingyu Liang. Learning mixtures of linear regressions with nearly optimal complexity. In *Conference On Learning Theory*, pages 1125–1144, 2018.

Bruce G Lindsay. The geometry of mixture likelihoods: a general theory. *The Annals of Statistics*, pages 86–94, 1983.

Bruce G Lindsay. Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–163. JSTOR, 1995.

AA Mallet. A maximum likelihood estimation method for random coefficient regression models. *Biometrika*, 73(3):645–656, 1986.

James Melville. *mize: Unconstrained Numerical Optimization Algorithms*, 2019. URL `https://cran.r-project.org/web/packages/mize/`. R package version 0.2.3.

Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pages 2796–2804, 2013.

Kalyanapuram Rangachari Parthasarathy. *Probability measures on metric spaces*, volume 352. American Mathematical Soc., 2005.

Michael JD Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The computer journal*, 7(2):155–162, 1964.

Richard E Quandt. The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the american statistical association*, 53(284):873–880, 1958.

Sujayam Saha and Adityanand Guntuboyina. On the nonparametric maximum likelihood estimator for gaussian location mixture densities with application to gaussian denoising. *Annals of Statistics*, 48 (2):738–762, 2020.

S.D. Silvey. *Optimal design: an introduction to the theory for parameter estimation*, volume 1. Springer Science & Business Media, 1980.

Aad van der Vaart, AW van der Vaart, Adrianus Willem van der Vaart, and Jon Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media, 1996.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Kert Viele and Barbara Tong. Modeling with mixtures of linear regressions. *Statistics and Computing*, 12(4):315–330, 2002.

Michel Wedel and Wagner A Kamakura. *Market segmentation: Conceptual and methodological foundations*, volume 8. Springer Science & Business Media, 2012.

CF Jeff Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, pages 95–103, 1983.

Chien-Fu Wu. Some algorithmic aspects of the theory of optimal designs. *The Annals of Statistics*, pages 1286–1301, 1978.

Weixin Yao and Weixing Song. Mixtures of linear regression with measurement errors. *Communications in Statistics-Theory and Methods*, 44(8):1602–1614, 2015.

Cun-Hui Zhang. Generalized maximum likelihood estimation of normal mixture densities. *Statistica Sinica*, pages 1297–1318, 2009.

Appendix A Existence of NPMLEs

This section is organized as follows. We first introduce and prove an important result that deals with general set K in Proposition A.1, which also fully proves the existence in the case that K is compact. Based on Proposition A.1, we prove the existence of nonparametric maximum likelihood estimators when K is $\mathbb{R}^p$ in Theorem 2.1. To avoid ambiguity of notation here, we shall use $\mathrm{f}(i)$ to denote the i-th component of vector f instead of the more common notation f_i, and save subscripts of f to differentiate vectors. Let us briefly recall the notation for sets of likelihood vectors. The mixture likelihood vector $(f^G_{x_1}(y_1), \ldots, f^G_{x_n}(y_n))^\top$ refers to f^G. Specially, when $G = \delta(\beta)$, the atomic likelihood vector $(f^\beta_{x_1}(y_1), \ldots, f^\beta_{x_n}(y_n))^\top$ refers to f^β. We further denote set $\mathcal{P}_K = \{\mathrm{f}^\beta : \beta \in K\}, \mathcal{Q}_K = \{\mathrm{f}^G : G \text{ is any probability measure supported on } K\}$, and write them for brevity as $\mathcal{P}, \mathcal{Q}$ respectively when $K = \mathbb{R}^p$.

In searching of the NPMLEs, we need to solve the following convex optimization problem.

$$\hat{\mathrm{f}} \in \arg\max_{\mathrm{f} \in \mathcal{Q}_K} \frac{1}{n} \sum_{i=1}^n \log \mathrm{f}(i), \tag{A.1}$$

where $\mathrm{f}(i)$ denotes the i-th element of vector f.

We define the following auxiliary optimization problem with the some objective function as in (A.1) but a different constraint set $\mathrm{conv}(\mathrm{cl}(\mathcal{P}_K))$, namely

$$\hat{\mathrm{f}} \in \arg\max_{\mathrm{f} \in \mathrm{conv}(\mathrm{cl}(\mathcal{P}_K))} \frac{1}{n} \sum_{i=1}^n \log \mathrm{f}(i). \tag{A.2}$$

The auxiliary optimization problem (A.2) is of interest because the constraint set in the original problem (A.1) is actually a subset of the constraint set in the auxiliary problem (A.2), as proved in Proposition A.1. More importantly, Carathéodory theorem can be directly employed to characterize the solution to (A.2).

Proposition A.1. *Given data $\{(x_i, y_i)\}_{i=1}^n$ and any set $K \subseteq \mathbb{R}^p$, the solution to* (A.1) *can be expressed as a convex combination of at most n points in* $\mathrm{cl}(\mathcal{P}_K)$.

Proof of Proposition A.1. Suppose G is a probability measure on $\mathbb{R}^p$, by Parthasarathy (2005, Theorem 6.3), there exist measures $\{\mu_m\}_{m=1}^\infty$ with finite supports, i.e. $\mu_m = \sum_{m_j} \pi_{m_j} \delta(\beta_{m_j})$, such that $\mu_m \xrightarrow{w} G$ as $m \to \infty$. For $i = 1, \ldots, n$, $f^{\mu_m}_{x_i}(y_i) \to f^G_{x_i}(y_i)$, and thus $\mathrm{f}^{\mu_m} \to \mathrm{f}^G$ as $m \to \infty$. Since $\mathrm{f}^{\mu_m} = \sum_{m_j} \pi_{m_j} \mathrm{f}^{\beta_{m_j}} \in \mathrm{conv}(\mathcal{P}_K)$ for $m = 1, 2, \ldots$, we have

$$\mathrm{f}^G = \lim_{m \to \infty} \mathrm{f}^{\mu_m} \in \mathrm{cl}(\{\mathrm{f}^{\mu_m}\}_{m=1}^\infty) \subseteq \mathrm{cl}(\mathrm{conv}(\mathcal{P}_K)).$$

Next we establish a basic fact for $\mathcal{P}_K$ that $\mathrm{conv}(\mathrm{cl}(\mathcal{P}_K)) = \mathrm{cl}(\mathrm{conv}(\mathcal{P}_K))$. $\mathcal{P}_K \subseteq [0, \frac{1}{\sqrt{2\pi}\sigma}]^n \subseteq \mathbb{R}^n$ is bounded, thus $\mathrm{cl}(\mathcal{P}_K)$ is compact. Since the convex hull of a compact set in Euclidean space is also compact (Bertsekas et al., 2003, Proposition 1.3.2), $\mathrm{conv}(\mathrm{cl}(\mathcal{P}_K))$ is compact. The closure of a compact set is itself, thus$\mathrm{conv}(\mathrm{cl}(\mathcal{P}_K)) = \mathrm{cl}(\mathrm{conv}(\mathrm{cl}(\mathcal{P}_K)))$. On the other hand, for any set $\mathcal{P}_K$ in $\mathbb{R}^p$, we have $\mathcal{P}_K \subseteq \mathrm{conv}(\mathcal{P}_K)$, so $\mathrm{cl}(\mathcal{P}_K) \subseteq \mathrm{cl}(\mathrm{conv}(\mathcal{P}_K))$. $\mathrm{cl}(\mathrm{conv}(\mathcal{P}_K))$ is convex, and $\mathrm{conv}(\mathrm{cl}(\mathcal{P}_K))$ is the smallest convex set that contains $\mathrm{cl}(\mathcal{P}_K)$, so

$$\mathrm{conv}(\mathrm{cl}(\mathcal{P}_K)) \subseteq \mathrm{cl}(\mathrm{conv}(\mathcal{P}_K)).$$

Moreover, $\mathrm{cl}(\mathrm{conv}(\mathcal{P}_K)) \subseteq \mathrm{cl}(\mathrm{conv}(\mathrm{cl}(\mathcal{P}_K)))$ since $\mathcal{P}_K \subseteq \mathrm{cl}(\mathcal{P}_K)$, we have

$$\mathrm{conv}(\mathrm{cl}(\mathcal{P}_K)) = \mathrm{cl}(\mathrm{conv}(\mathrm{cl}(\mathcal{P}_K))) = \mathrm{cl}(\mathrm{conv}(\mathcal{P}_K))$$

where the inclusion becomes equality.

Therefore,

$$\mathcal{Q}_K = \{\mathrm{f}^G : G \text{ is any probability measure on } \mathbb{R}^p\} \subseteq \mathrm{cl}(\mathrm{conv}(\mathcal{P}_K)) = \mathrm{conv}(\mathrm{cl}(\mathcal{P}_K)).$$

The optimization problem (A.1) is to maximize $L(\mathrm{f})$ over $\mathcal{Q}_K$, but $\mathcal{Q}_K$ is not closed because vectors with zero elements are not in $\mathcal{Q}_K$. Instead, we will consider maximizing $L(\mathrm{f})$ over a larger set, the compact set $\mathrm{conv}(\mathrm{cl}(\mathcal{P}_K))$, and then prove that there is an optimal solution to this new problem on the larger set $\mathrm{conv}(\mathrm{cl}(\mathcal{P}_K))$ that also lies in $\mathcal{Q}_K$. Since $L(\mathrm{f})$ is a continuous function over a compact region, there exists $\hat{\mathrm{f}} \in \mathrm{conv}(\mathrm{cl}(\mathcal{P}_K))$ that $L(\hat{\mathrm{f}})$ achieves maximum over $\mathrm{conv}(\mathrm{cl}(\mathcal{P}_K))$. $\hat{\mathrm{f}}$ is unique because $L(\cdot)$ is strictly concave. If $\hat{\mathrm{f}}$ is an interior point of $\mathrm{conv}(\mathrm{cl}(\mathcal{P}_K))$, then by the optimal condition, $\nabla L(\hat{\mathrm{f}}) = 0$, which is impossible because $\nabla L(\hat{\mathrm{f}})^\top = (1/\hat{\mathrm{f}}(1), \ldots, 1/\hat{\mathrm{f}}(n))$. Therefore, we know $\hat{\mathrm{f}}$ lies in the boundary of set $\mathrm{conv}(\mathrm{cl}(\mathcal{P}_K))$.

By Carathéodory's theorem (see, e.g., Silvey (1980, Appendix 2)), any boundary point of $\mathrm{conv}(\mathrm{cl}(\mathcal{P}_K))$ can be expressed as a convex combination of at most n points in $\mathrm{cl}(\mathcal{P}_K)$. Specifically, $\hat{\mathrm{f}}$ can be expressed as a convex combination of at most n points in $\mathrm{cl}(\mathcal{P}_K)$.

□

When K is compact thus $\mathcal{P}_K$ is also compact and $\mathrm{cl}(\mathcal{P}_K) = \mathcal{P}_K$, Proposition A.1 directly implies that the solution to (A.1) can be expressed as a convex combination of at most n points in $\mathcal{P}_K$. However, when K is not compact, Proposition A.1 is useful but not ideal enough in the sense that points in $\mathrm{cl}(\mathcal{P}_K) \setminus \mathcal{P}_K$ are defined by the limit of a series of atomic likelihood vectors and cannot be directly expressed as f^β for a certain β. From the computational perspective, these boundary points make existing algorithms inviable. To see why $\mathcal{P}_K$ may not be compact, all elements of f are positive for any $\mathrm{f} \in \mathcal{P}_K$ by definition, no mater how small they can be; while $\mathrm{cl}(\mathcal{P}_K)$ contain vectors with elements equal to 0. One way to resolve this problem is by showing that we can actually replace vectors in $\mathrm{cl}(\mathcal{P}_K)$ with some vector in $\mathcal{P}_K$, without decreasing the objective function. Our proof of Theorem 2.1 is based on this idea and reveals the structural properties of points in $\mathrm{cl}(\mathcal{P}_K)$ when K is $\mathbb{R}^p$.

Proof of Theorem 2.1. We only need to prove $K = \mathbb{R}^p$ case since when K is a compact set we have $\mathrm{cl}(\mathcal{P}_K) = \mathcal{P}_K$.

By Proposition A.1, there exists a maximizer over $\mathrm{conv}(\mathrm{cl}(\mathcal{P}))$

$$\hat{\mathrm{f}} = \sum_{j=1}^{N} \pi_j \mathrm{g}_j,$$

where $N \leq n$, $\pi_j > 0$, $\sum_{j=1}^N \pi_j = 1$, and $\mathrm{g}_j \in \mathrm{cl}(\mathcal{P})$ for $j = 1, \ldots, N$.

Given any $j \in \{1, \ldots, N\}$, we denote g_j by g for simplicity in notation. Without loss of generality, we also assume that the positive elements of g are $\mathrm{g}(1), \ldots, \mathrm{g}(s)$. By definition of closure, there exists a vector series $\{\mathrm{g}^{\beta_l}\}_{l=1}^\infty$ such that

$$\mathrm{g} = \lim_{l \to \infty} \mathrm{g}^{\beta_l}.$$

Denote $X_s = [x_1, \ldots, x_s]$, and define $\alpha_l = X_s^\top \beta_l \in \mathbb{R}^s$ for $l = 1, 2, \ldots$. Because $\mathrm{g}(1), \ldots, \mathrm{g}(s)$ are positive, there exist positive number ϵ and N_ϵ such that the first k elements of g^{β_l} are larger than ϵ for all $l > N_\epsilon$. Since $\phi((y_i - \alpha_l(i))/\sigma) > \epsilon$ for all $i = 1, \ldots, s$ and $l > N_\epsilon$, we have $\{\alpha_l\}_{l > N_\epsilon}$ is bounded. By Bolzano-Weierstrass property, there exists a convergent subsequence $\{\alpha_{j_l}\}$ and we denote the limit of the subsequence by $\alpha = \lim_{l \to \infty} \alpha_{j_l}$.

Pick any generalized inverse $X_s^g \in \mathbb{R}^{p \times s}$ of matrix $X_s^\top$, then

$$\tilde{\beta}_l = X_s^g \alpha_l$$

satisfies that

$$X_s^\top \tilde{\beta}_l = X_s^\top X_s^g \alpha_l = X_s^\top X_s^g X_s^\top \beta_l = X_s^\top \beta_l = \alpha_l.$$

Thus for subsequence $\{\tilde{\beta}_{j_l}\} = \{X_s^g \alpha_{j_l}\} \to X_s^g \alpha$ as $l \to \infty$, and we define $\tilde{\beta} = X_s^g \alpha$.

For $i = 1, \ldots, s$

$$\mathrm{g}(i) = \lim_{l\to\infty} \mathrm{g}^{\beta_l}(i) = \lim_{l\to\infty} \mathrm{g}^{\beta_{j_l}}(i) = \lim_{l\to\infty} \mathrm{g}^{\tilde{\beta}_{j_l}}(i) = \mathrm{g}^{\tilde{\beta}}(i).$$

For all $j = 1, \ldots, N$, we can find a corresponding $\tilde{\beta}_j$ such that $\mathrm{f}^{\tilde{\beta}_j}(i) = \mathrm{g}_j(i)$ for all positive element i's of g_j. If g_j contains 0 element, say $\mathrm{g}_j(i_0) = 0$, we have $\mathrm{f}^{\tilde{\beta}_j}(i_0) > 0 = \mathrm{g}_j(i_0)$, thus $L(\hat{\mathrm{f}}) < L(\hat{\mathrm{f}} - \pi_j \mathrm{g}_j + \pi_j \mathrm{f}^{\tilde{\beta}_j})$, which contradicts with the optimality of g. As a result, all elements of g_j are positive, and $\mathrm{g}_j = \mathrm{f}^{\tilde{\beta}_j} \in \mathcal{P}$.

Define a discrete measure over $\mathbb{R}^p$ as

$$\hat{G} = \sum_{j=1}^{N} \pi_j \delta(\tilde{\beta}_j),$$

then $\hat{\mathrm{f}} = \mathrm{f}^{\hat{G}} \in \mathcal{Q}$, and $L(\mathrm{f}^{\hat{G}})$ achieves the maximum over $\mathcal{Q}$. □

Now we prove a corollary of Theorem 2.1, which is very useful in the computation of NPMLEs.

Proof of Corollary 2.1. By Theorem 2.1, the optimal solution to problem (A.1), denoted by $\mathrm{f}^{\hat{G}}$, can be expressed as a convex combination of at most n points in $\mathcal{P}_K$, which belongs to $\mathrm{conv}(\mathcal{P}_K)$. Moreover, vectors in $\mathrm{conv}(\mathcal{P}_K)$ correspond to discrete probability measures, so $\mathrm{conv}(\mathcal{P}_K) \subseteq \mathcal{Q}_K$. Thus $\mathrm{f}^{\hat{G}}$ must also be an optimal solution to (2.1). On the other hand, the objective function in (2.1) is strictly concave, so $\mathrm{f}^{\hat{G}}$ is the unique optimal solution. □

Appendix B Algorithm convergence

In this section, we present the convergence guarantee for our computational procedures, specifically Algorithm 1. The convergence proof of Algorithm 2 is standard in existing listerature (see, e.g., Jaggi (2013)).

Proof of Proposition 2.1. By concavity of $L(\cdot)$, we have

$$L(\hat{\mathrm{f}}) - L(\mathrm{f}^{(t)}) \le \langle \nabla L(\mathrm{f}^{(t)}), \hat{\mathrm{f}} - \mathrm{f}^{(t)} \rangle. \tag{B.1}$$

Since f is contained in $\mathcal{P}_K$, we have

$$\langle \nabla L(\mathrm{f}^{(t)}), \hat{\mathrm{f}} - \mathrm{f}^{(t)} \rangle \le \max_{\mathrm{g}\in\mathcal{P}_K} \langle \nabla L(\mathrm{f}^{(t)}), \mathrm{g} - \mathrm{f}^{(t)} \rangle. \tag{B.2}$$

By equation (2.2) we have

$$\max_{\mathrm{g}\in\mathcal{P}_K} \langle \mathrm{g}, \nabla L(\mathrm{f}^{(t)}) \rangle - e_t \le \langle \tilde{\mathrm{g}}, \nabla L(\mathrm{f}^{(t)}) \rangle.$$

Adding the three equations above together, proof is completed.

□

Proof of Theorem 2.2. First we show that $L(\mathrm{f})$ has Lipschitz gradient under Assumption 2.3. For any f, g with $\min_i \mathrm{f}(i) \geq \delta$ and $\min_i \mathrm{g}(i) \geq \delta$,

$$\|\nabla L(\mathrm{f}) - \nabla L(\mathrm{g})\|^2 = \frac{1}{n^2} \sum_{i=1}^{n} \frac{|\mathrm{f}(i) - \mathrm{g}(i)|^2}{\mathrm{f}(i)^2 \mathrm{g}(i)^2} \leq \frac{1}{\delta^2 n^2} \sum_{i=1}^{n} |\mathrm{f}(i) - \mathrm{g}(i)|^2 = \frac{1}{\delta^4 n^2} \|\mathrm{f} - \mathrm{g}\|^2.$$

Thus L has Lipschitz continuous gradient with constant $\frac{1}{\delta^2 n}$.

Denote $\ell(t) = L(\mathrm{f} + t(\mathrm{g} - \mathrm{f}))$, then

$$\ell'(0) - \ell'(t) = \langle \nabla L(\mathrm{f} + t(\mathrm{g} - \mathrm{f})) - \nabla L(\mathrm{f}), \mathrm{f} - \mathrm{g} \rangle \leq t \frac{1}{\delta^2 n} \|\mathrm{f} - \mathrm{g}\|^2,$$

where the second inequality is by Cauchy-Schwarz.

By integral formula,

$$\begin{aligned} L(\mathrm{g}) = \ell(1) = \ell(0) + \int_0^1 \ell'(t) \mathrm{d}t &= \ell(0) + \ell'(0) + \int_0^1 (\ell'(t) - \ell'(0) \mathrm{d}t \\ &\geq \ell(0) + \ell'(0) - \frac{1}{2\delta^2 n} \|\mathrm{f} - \mathrm{g}\|^2 \\ &= L(\mathrm{f}) + \langle L(\mathrm{f}), \mathrm{g} - \mathrm{f} \rangle - \frac{1}{2\delta^2 n} \|\mathrm{f} - \mathrm{g}\|^2 \\ &= L(\mathrm{f}) + \langle L(\mathrm{f}), \mathrm{g} - \mathrm{f} \rangle - \frac{2}{\delta^2} \end{aligned}$$

where the last step is because $\|\mathrm{f} - \mathrm{g}\|^2 \leq 4n$.

After deriving some basic properties of function $L(\mathrm{f})$, now we get back to the iterations in the CGM algorithm.

$$\begin{aligned} L(\mathrm{f}^{(t+1)}) \geq & L(\mathrm{f}^{(t)} + \gamma_t(\tilde{\mathrm{g}}^{(t)} - \mathrm{f}^{(t)})) \\ \geq & L(\mathrm{f}^{(t)}) + \gamma_t \langle L(\mathrm{f}^{(t)}), \tilde{\mathrm{g}}^{(t)} - \mathrm{f}^{(t)} \rangle - \frac{2\gamma_t^2}{\delta^2} \\ \geq & L(\mathrm{f}^{(t)}) + \gamma_t (\max_{\mathrm{g} \in \mathcal{P}_K} \langle \mathrm{g} - \mathrm{f}^{(t)}, \nabla L(\mathrm{f}^{(t)}) \rangle - e_t) - \frac{2\gamma_t^2}{\delta^2} \end{aligned}$$

where the last inequality uses Proposition 2.1.

For any $\mathrm{f} \in \mathrm{conv}(\mathcal{P}_K)$, we define the duality gap function

$$D(\mathrm{f}) := \max_{\mathrm{g} \in \mathcal{P}_K} \langle \mathrm{g} - \mathrm{f}, \nabla L(\mathrm{f}) \rangle.$$

We also denote the error function $\mathcal{E}(\mathrm{f}) = L(\hat{\mathrm{f}}) - L(\mathrm{f})$. Then (B.1) and (B.2) lead to that $E(\mathrm{f}) \leq D(\mathrm{f})$.

Under the simplified notation, we have

$$\begin{aligned} \mathcal{E}(\mathrm{f}^{(t+1)}) \leq & \mathcal{E}(\mathrm{f}^{(t)}) - \gamma_t D(\mathrm{f}^{(t)}) + \gamma_t e_t + \frac{2\gamma_t^2}{\delta^2} \\ \leq & \mathcal{E}(\mathrm{f}^{(t)}) - \gamma_t \mathcal{E}(\mathrm{f}^{(t)}) + \gamma_t e_t + \frac{2\gamma_t^2}{\delta^2} \\ = & (1 - \gamma_t) \mathcal{E}(\mathrm{f}^{(t)}) + \gamma_t (e_t + \frac{2\gamma_t}{\delta^2}). \end{aligned} \tag{B.3}$$

Next we shall use induction to prove that

$$\mathcal{E}(\mathrm{f}^{(t+1)}) \leq \frac{4}{t + 1 + 2} C,$$

where $C = e_t/\gamma_t + \frac{2}{\delta^2} = \frac{2}{\delta^2}(1+\epsilon)$.

The base case $t = 0$ holds with $\gamma_0 = 1$ by (B.3) applying to the start of the CGM algorithm. For $t \geq 1$, (B.3) gives

$$\begin{aligned}
\mathcal{E}(\mathrm{f}^{(t)}) &\leq (1-\gamma_t)\mathcal{E}(\mathrm{f}^{(t)}) + \gamma_t^2 C \\
&\leq \left(1 - \frac{2}{t+2}\right)\mathcal{E}(\mathrm{f}^{(t)}) + \left(\frac{2}{t+2}\right)^2 C \\
&\leq \left(1 - \frac{2}{t+2}\right)\frac{4C}{t+2} + \left(\frac{2}{t+2}\right)^2 C \\
&\leq \frac{4C}{t+3}.
\end{aligned}$$

□

Appendix C Hellinger accuracy

This section includes the proofs of Theorem 3.1 and Theorem 3.2. These results crucially use the metric entropy results proved in Appendix E.

Proof of Theorem 3.1. Let $S_0 := \{x_1, \ldots, x_n\}$ be the set of design points. Recall the definition of the class $\mathcal{M}_R$ from (3.5). Let $\|\cdot\|_\infty$ be the pseudometric on $\mathcal{M}_R$ given by

$$(f^G, f^{G'}) \mapsto \sup_{x \in S_0, y \in \mathbb{R}} \left| f_x^G(y) - f_x^{G'}(y) \right|.$$

Theorem E.1 gives an upper bound on the η-covering number $N(\eta, \mathcal{M}_R, \|\cdot\|_\infty)$ of $\mathcal{M}_R$ under the pseudometric $\|\cdot\|_\infty$ (the definition of covering numbers is recalled in Section E). This result will be crucially used in this proof.

Let $\{h^1, \ldots, h^N\} \subseteq \mathcal{M}_R$ be an η-covering set of $\mathcal{M}_R$ under $\|\cdot\|_\infty$ where $N = N(\eta, \mathcal{M}_R, \|\cdot\|_\infty)$. This ensures

$$\sup_{h \in \mathcal{M}_R} \inf_{1 \leq j \leq N} \|h - h^j\|_\infty \leq \eta. \tag{C.1}$$

For a fixed sequence $\{\gamma_n\}_{n \geq 1}$ and $t > 0$, let us now bound $\mathbb{P}\{\mathfrak{H}_n(f^{\hat{G}}, f^{G^*}) \geq t\gamma_n\}$ (the precise form for γ_n will be given later in the proof; it will equal a constant multiple of ϵ_n).

We define a set $J \subseteq \{1, \ldots, N\}$. Let J be composed of all index $j \in \{1, \ldots, N\}$ for which there exists $h^{0j} \in \mathcal{M}_R$ satisfying

$$\|h^{0j} - h^j\|_{\infty, S_0 \times \mathbb{R}} \leq \eta \quad \text{and} \quad \mathfrak{H}_n(h^{0j}, f^{G^*}) \geq t\gamma_n. \tag{C.2}$$

Let $j \in \{1, \ldots, N\}$ be such that $\|h^j - f^{\hat{G}}\|_\infty \leq \eta$ (such a j clearly exists because $h^1, \ldots, h^N$ form an η-covering set of $\mathcal{M}_R$). Now if $\mathfrak{H}_n(f^{\hat{G}}, f^{G^*}) \geq t\gamma_n$, then $j \in J$ and consequently $\|f^{\hat{G}} - h^{0j}\|_\infty \leq 2\eta$ which implies that

$$f_{x_i}^{\hat{G}}(y) \leq h_{x_i}^{0j}(y) + 2\eta \qquad \text{for all } i = 1, \ldots, n \text{ and } y \in \mathbb{R}.$$

Therefore

$$\prod_{i=1}^n f_{x_i}^{G^*}(Y_i) \leq \prod_{i=1}^n f_{x_i}^{\hat{G}}(Y_i) \leq \prod_{i=1}^n \{h_{x_i}^{0j}(Y_i) + 2\eta\} \leq \max_{j \in J} \prod_{i=1}^n \{h_{x_i}^{0j}(Y_i) + 2\eta\},$$

where the first inequality follows from the fact that $\hat{G}$ maximizes the likelihood. We thus get

$$\begin{aligned}\mathbb{P}(\mathfrak{H}_n(f^{\hat{G}}, f^{G^*}) \geq t\gamma_n) &\leq \mathbb{P}\left\{\max_{j\in J}\prod_{i=1}^n \frac{h_{x_i}^{0j}(Y_i)+2\eta}{f_{x_i}^{G^*}(Y_i)} \geq 1\right\} \\ &\leq \sum_{j\in J}\mathbb{P}\left\{\prod_{i=1}^n \frac{h_{x_i}^{0j}(Y_i)+2\eta}{f_{x_i}^{G^*}(Y_i)} \geq 1\right\} \\ &\leq \sum_{j\in J}\mathbb{E}\prod_{i=1}^n \sqrt{\frac{h_{x_i}^{0j}(Y_i)+2\eta}{f_{x_i}^{G^*}(Y_i)}} = \sum_{j\in J}\prod_{i=1}^n \mathbb{E}\sqrt{\frac{h_{x_i}^{0j}(Y_i)+2\eta}{f_{x_i}^{G^*}(Y_i)}},\end{aligned}$$

where we used the union bound in the second line and Markov's inequality (followed by the independence of $Y_1, \ldots, Y_n$) in the third line. For each $j \in J$,

$$\begin{aligned}\prod_{i=1}^n \mathbb{E}\sqrt{\frac{h_{x_i}^{0j}(Y_i)+2\eta}{f_{x_i}^{G^*}(Y_i)}} &= \exp\left(\sum_{i=1}^n \log \mathbb{E}\sqrt{\frac{h_{x_i}^{0j}(Y_i)+2\eta}{f_{x_i}^{G^*}(Y_i)}}\right) \\ &\leq \exp\left(\sum_{i=1}^n \mathbb{E}\sqrt{\frac{h_{x_i}^{0j}(Y_i)+2\eta}{f_{x_i}^{G^*}(Y_i)}} - n\right) \\ &= \exp\left(\sum_{i=1}^n \int \sqrt{(h_{x_i}^{0j}+2\eta) f_{x_i}^{G^*}} - n\right)\end{aligned}$$

where we used the inequality $\log a \leq a - 1$ in the second line, and the last equality follows from the fact that Y_i has density $f_{x_i}^{G^*}$. The simple inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ now gives, for each $1 \leq i \leq n$,

$$\begin{aligned}\int \sqrt{(h_{x_i}^{0j}+2\eta) f_{x_i}^{G^*}} &\leq \int \sqrt{h_{x_i}^{0j} f_{x_i}^{G^*}} + \sqrt{2\eta}\int \sqrt{f_{x_i}^{G^*}} \\ &\leq 1 - \frac{1}{2}\mathfrak{H}^2(h_{x_i}^{0j}, f_{x_i}^{G^*}) + \sqrt{2\eta}\sqrt{\int f_{x_i}^{G^*}} = 1 - \frac{1}{2}\mathfrak{H}^2(h_{x_i}^{0j}, f_{x_i}^{G^*}) + \sqrt{2\eta}.\end{aligned}$$

As a result, we deduce

$$\sum_{i=1}^n \int \sqrt{(h_{x_i}^{0j}+2\eta) f_{x_i}^{G^*}} \leq n - \frac{1}{2}\sum_{i=1}^n \mathfrak{H}^2(h_{x_i}^{0j}, f_{x_i}^{G^*}) + n\sqrt{2\eta}.$$

As we have assumed that for every $j \in J$,

$$\sum_{i=1}^n \mathfrak{H}^2(h_{x_i}^{0j}, f_{x_i}^{G^*}) = n\mathfrak{H}_n^2(h^{0j}, f^{G^*}) \geq nt^2\gamma_n^2,$$

we obtain

$$\sum_{i=1}^n \int \sqrt{(h_{x_i}^{0j}+2v_i) f_{x_i}^{G^*}} \leq n - \frac{n}{2}t^2\gamma_n^2 + n\sqrt{2\eta}.$$

We have thus proved

$$\prod_{i=1}^n \mathbb{E}\sqrt{\frac{h_{x_i}^{0j}(Y_i)+2\eta}{f_{x_i}^{G^*}(Y_i)}} \leq \exp\left(\sum_{i=1}^n \int \sqrt{(h_{x_i}^{0j}+2\eta) f_{x_i}^{G^*}} - n\right) \leq \exp(-\frac{n}{2}t^2\gamma_n^2 + n\sqrt{2\eta})$$

which gives (note that $|J| \leq N$)

$$\begin{aligned}\mathbb{P}\left\{\mathfrak{H}_n(f^{\hat{G}}, f^{G^*}) \geq t\gamma_n\right\} &\leq |J| \cdot \exp\left(-\frac{n}{2}t^2\gamma_n^2 + n\sqrt{2\eta}\right) \\ &\leq \exp\left(\log N - \frac{n}{2}t^2\gamma_n^2 + n\sqrt{2\eta}\right). \end{aligned} \tag{C.3}$$

We now use the metric entropy result in Theorem E.1 to bound $\log N$. Setting $S_0 = \{x_i\}_{i=1}^n$ and $K = \{\beta \in \mathbb{R}^p : \|\beta\| \leq R\}$ in Theorem E.1, we get

$$\log N(\eta, \mathcal{M}_R, \|\cdot\|_\infty) \leq C_p \zeta^p N(\{2\log(3\sigma^{-1}\eta^{-1})\}^{1/2}\sigma/\mathfrak{L}, \{\beta : \|\beta\| \leq R\})\{\log(\sigma^{-1}\eta^{-1})\}^{p+1}$$

where $\mathfrak{L} = \sup_{x \in S_0} \mathfrak{L}(x) = \sup_{1 \leq i \leq n} \mathfrak{L}(x_i)$ and $\mathfrak{L}(x_i)$ is defined in (E.2). It is clear that for the linear model, $\zeta = 1$ and $\mathfrak{L}(x_i) \leq \|x_i\| \leq B$ (note that we have made the assumption $\max_{1 \leq i \leq n} \|x_i\| \leq B$). The Euclidean covering number $N(\{2\log(3\sigma^{-1}\eta^{-1})\}^{1/2}\sigma/\mathfrak{L}, \{\beta : \|\beta\| \leq R\})$ is bounded in the following way. It is well-known that

$$N\left(\epsilon, \{\beta \in \mathbb{R}^p : \|\beta\| \leq R\}\right) \leq \left(1 + \frac{2R}{\epsilon}\right)^p \qquad \text{for all } \epsilon > 0$$

and consequently

$$N(\{2\log(3\sigma^{-1}\eta^{-1})\}^{1/2}\sigma/\mathfrak{L}, \{\beta : \|\beta\| \leq R\}) \leq \left(1 + \frac{2R\mathfrak{L}}{\{2\log(3\sigma^{-1}\eta^{-1})\}^{1/2}\sigma}\right)^p.$$

This and the fact that $\mathfrak{L} \leq B$ lead to

$$\begin{aligned}\log N = \log N(\eta, \mathcal{M}_R, \|\cdot\|_\infty) \leq & C_p \left(1 + \frac{2RB}{\{2\log(3\sigma^{-1}\eta^{-1})\}^{1/2}\sigma}\right)^p \{\log(\sigma^{-1}\eta^{-1})\}^{p+1} \\ &\leq C_p\{\log(\sigma^{-1}\eta^{-1})\}^{p+1} + C_p\left(\frac{RB}{\sigma}\right)^p \{\log(3\sigma^{-1}\eta^{-1})\}^{p/2+1},\end{aligned}$$

where we also used that $\mathfrak{L} \leq B$ and that C_p absorbs a coeffient 2^p. Using the above in (C.3), we obtain

$$\mathbb{P}\left\{\mathfrak{H}_n(f^{\hat{G}}, f^{G^*}) \geq t\gamma_n\right\} \leq \exp\left(C_p\{\log(\sigma^{-1}\eta^{-1})\}^{p+1} + C_p\left(\frac{RB}{\sigma}\right)^p \{\log(3\sigma^{-1}\eta^{-1})\}^{p/2+1} - \frac{n}{2}t^2\gamma_n^2 + n\sqrt{2\eta}\right).$$

We shall now take γ_n and η so that

$$n\gamma_n^2 \geq 12\max\left(C_p\{\log(\sigma^{-1}\eta^{-1})\}^{p+1}, C_p\left(\frac{RB}{\sigma}\right)^p \{\log(3\sigma^{-1}\eta^{-1})\}^{p/2+1}, n\sqrt{2\eta}\right). \tag{C.4}$$

This will ensure that, for $t \geq 1$,

$$\mathbb{P}\left\{\mathfrak{H}_n(f^{\hat{G}}, f^{G^*}) \geq t\gamma_n\right\} \leq \exp\left(\frac{n\gamma_n^2}{4}(1 - 2t^2)\right) \leq \exp\left(-\frac{nt^2\gamma_n^2}{4}\right). \tag{C.5}$$

To satisfy (C.4), we first take $\eta := \gamma_n^4/288$ (so that $12n\sqrt{2\eta} = n\gamma_n^2$). The quantity γ_n will then have to satisfy the two inequalities:

$$n\gamma_n^2 \geq 12C_p\left(\log\frac{288}{\sigma\gamma_n^4}\right)^{p+1} \tag{C.6}$$

and

$$n\gamma_n^2 \geq 12C_p\left(\frac{RB}{\sigma}\right)^p \left(\log\frac{864}{\sigma\gamma_n^4}\right)^{p/2+1}. \tag{C.7}$$

It is now elementary to check that (C.6) is satisfied whenever

$$\gamma_n \geq \sqrt{\frac{12C_p}{n}} \left(\text{Log } \frac{2n^2}{\sigma C_p^2} \right)^{(p+1)/2}$$

and (C.7) is satisfied whenever

$$\gamma_n \geq \sqrt{\frac{12C_p}{n}} \left(\frac{RB}{\sigma} \right)^{p/2} \left(\text{Log } \frac{6n^2\sigma^{2p}}{\sigma C_p^2 (RB)^{2p}} \right)^{(p/4)+(1/2)},$$

where we used the notation $\text{Log } x := \max(1, \log x)$.

We may now assume $C_p \geq \sqrt{6}$. It is then easy to see that both the above inequalities and consequently both (C.6) and (C.7) are satisfied whenever

$$\gamma_n \geq \sqrt{\frac{12C_p}{n}} \max \left(\left(\text{Log } \frac{n^2}{\sigma} \right)^{\frac{p+1}{2}}, \left(\frac{RB}{\sigma} \right)^{\frac{p}{2}} \left(\text{Log } \frac{n^2\sigma^{2p}}{\sigma (RB)^{2p}} \right)^{\frac{p}{4}+\frac{1}{2}} \right)$$

Using $\text{Log } x^2 \leq 2\text{Log } x$ and absorbing all the p-dependent constants in C_p, we deduce that inequality (C.5) holds for $\gamma_n = \sqrt{C_p}\epsilon_n$ where ϵ_n is defined in (3.2). This completes the proof of (3.3) (note that $\exp(-nt^2 C_p \epsilon_n^2/4)$ can be bounded by $\exp(-nt^2\epsilon_n^2)$ by taking C_p larger than 4).

To prove (3.4), we multiply both sides of (3.3) by t and integrate from $t = 1$ to $t = \infty$ to obtain

$$\mathbb{E} \left(\frac{\mathfrak{H}_n^2(f^{\hat{G}}, f^{G^*})}{C_p \epsilon_n^2} - 1 \right)_+ \leq \frac{1}{n\epsilon_n^2}$$

where $x_+ := \max(x, 0)$ which implies

$$\mathbb{E}\mathfrak{H}_n^2(f^{\hat{G}}, f^{G^*}) \leq C_p \epsilon_n^2 + \frac{C_p}{n}.$$

This proves (3.4) (after changing C_p to $2C_p$) as $\epsilon_n^2 \geq n^{-1}$. □

Proof of Theorem 3.2. The proof of Theorem 3.2 can be obtained without difficulty by carefully replacing all the $x^\top \beta$ in the proof of Theorem 3.1 by $r(x, \beta)$. Let us give a few remarks to validate why this the generalization can be done. Firstly, our metric entropy result proved in Section E is already generalized that also holds for polynomial regressions. Secondly, the definition of $\mathfrak{L}$ is readily generalized to be the Lipschitz constant, which makes all the inequalities involved with $\mathfrak{L}$ still hold in the polynomial regression case. In the remainder of the proof, we will emphasize on the differences and abbreviate common arguments that have been elaborated in the proof of Theorem 3.1.

Same as the proof of Theorem 3.1, (C.3) also applies, i.e.,

$$\mathbb{P} \left\{ \mathfrak{H}_n(f^{\hat{G}}, f^{G^*}) \geq t\gamma_n \right\} \leq \exp \left(\log N - \frac{n}{2} t^2 \gamma_n^2 + n\sqrt{2\eta} \right)$$

where $\log N = \log N(\eta, \mathcal{M}_R, \|\cdot\|_\infty)$.

Setting $S_0 = \{x_i\}_{i=1}^n$ and $K = \{\beta \in \mathbb{R}^p : \|\beta\| \leq R\}$ in Theorem E.1, we get

$$\log N(\eta, \mathcal{M}_R, \|\cdot\|_\infty) \leq C_p \zeta^p N(\{2\log(3\sigma^{-1}\eta^{-1})\}^{1/2}\sigma/\mathfrak{L}, \{\beta : \|\beta\| \leq R\}) \{\log(\sigma^{-1}\eta^{-1})\}^{p+1}$$

where $\mathfrak{L} = \sup_{x \in S_0} \mathfrak{L}(x) = \sup_{1 \leq i \leq n} \mathfrak{L}(x_i)$, and $\mathfrak{L}(x_i)$ is defined in (E.2) (same as in (3.6)). The proof of Theorem 3.1 is more special by seeting $\zeta = 1$ and using $\mathfrak{L} \leq B$. In contrast, we need to

keep both ζ and $\mathfrak{L}$ here. Again, we use the well-known bound on the Euclidean covering number that $N\left(\epsilon, \{\beta \in \mathbb{R}^p : \|\beta\| \leq R\}\right) \leq \left(1 + \frac{2R}{\epsilon}\right)^p, \forall \epsilon > 0$, which leads to

$$\log N = \log N(\eta, \mathcal{M}_R, \|\cdot\|_\infty) \leq \leq C_p \zeta^p \{\log(\sigma^{-1}\eta^{-1})\}^{p+1} + C_p \zeta^p \left(\frac{R\mathfrak{L}}{\sigma}\right)^p \{\log(3\sigma^{-1}\eta^{-1})\}^{p/2+1}$$

The next crucial step is to take γ_n and η so that

$$n\gamma_n^2 \geq 12 \max\left(C_p \zeta^p \{\log(\sigma^{-1}\eta^{-1})\}^{p+1}, C_p \zeta^p \left(\frac{R\mathfrak{L}}{\sigma}\right)^p \{\log(3\sigma^{-1}\eta^{-1})\}^{p/2+1}, n\sqrt{2\eta}\right). \tag{C.8}$$

Again, η is taken as $\eta := \gamma_n^4/288$. It is elementary to check that (C.8) holds when γ_n satisfy both

$$\gamma_n \geq \sqrt{\frac{12 C_p \zeta^p}{n}} \left(\text{Log } \frac{2n^2}{\sigma C_p^2 \zeta^{2p}}\right)^{(p+1)/2}$$

and

$$\gamma_n \geq \sqrt{\frac{12 C_p \zeta^p}{n}} \left(\frac{R\mathfrak{L}}{\sigma}\right)^{p/2} \left(\text{Log } \frac{6n^2\sigma^{2p}}{\sigma \zeta^{2p} C_p^2 (R\mathfrak{L})^{2p}}\right)^{(p/4)+(1/2)}.$$

We further simplify the conditions for γ_n and absorb constants into C_p, the crucial condition for γ_n becomes (assuming $C_p \geq \sqrt{6}$ and noting that $\zeta \geq 1$)

$$\gamma_n \geq \sqrt{\frac{C_p \zeta^p}{n}} \max\left(\left(\text{Log } \frac{n^2}{\sigma\zeta^{2p}}\right)^{\frac{p+1}{2}}, \left(\frac{R\mathfrak{L}}{\sigma\zeta^{2p}}\right)^{\frac{p}{2}} \left(\text{Log } \frac{n^2\sigma^{2p}}{\sigma(R\mathfrak{L}\zeta)^{2p}}\right)^{\frac{p}{4}+\frac{1}{2}}\right).$$

We then take ϵ_n as a upper bound of $\gamma_n/\sqrt{C_p}$ as below

$$\epsilon_n^2 = n^{-1}\left(\zeta^p \left(\text{Log } \frac{n}{\sqrt{\sigma}\zeta^p}\right)^{p+1}, \left(\frac{\zeta R\mathfrak{L}}{\sigma}\right)^p \left(\text{Log } \left\{\frac{n}{\sqrt{\sigma}} \left(\frac{\sigma}{\zeta R\mathfrak{L}}\right)^p\right\}\right)^{\frac{p}{2}+1}\right).$$

The remainder follows the same argument as in Theorem 3.1. □

Appendix D Random design

In this section, we first prove the bound on prediction error under random design in Theorem 3.3 , and then we prove weak consistency of the nonparametric maximum likelihood estimator in Theorem 3.4.

In the main body of the proof of Theorem 3.3, we bound the discrepancy between prediction error under fixed design and prediction error under random design by a small quantity. The main technique involves using a maximal inequality for expected supremum. Combining the discrepancy bound with the previous prediction error bound under fixed design, we complete the proof of the Theorem 3.3. To make the maximal inequality meaningful, we prove a bracketing number bound and show that the corresponding bracketing integral is finite.

Given functions l and u, the bracket $[l, u]$ refers to a class of functions $\{f | l \leq f \leq u\}$. Given measure P and $r > 0$, the size of a bracket $[l, u]$ under metric $L_r(P)$ is defined as $(\int P(u-l)^r)^{1/r}$. The bracketing number $N_{[]}(\epsilon, \mathcal{F}, L_r(P))$ is defined as the minimum number of brackets of size ϵ in order to cover the function class $\mathcal{F}$. The bracketing entropy is the logarithm of the bracketing number, and the bracketing integral is an integral defined as

$$J_{[]}(\delta, \mathcal{F}, L_2(P)) = \int_0^\delta \sqrt{1 + \log N_{[]}(\epsilon, \mathcal{F}, L_r(P))} \mathrm{d}\epsilon.$$

Proof of Theorem 3.3. Under the random design, we assume the regressors x_i, $i = 1, \ldots, n$ are independently generated from the same probability measure μ. The probability measure μ is supported on a set $S_0 \subseteq \mathbb{R}^p$. Given the true mixing probability measure G^*, we introduce function $D^G : S_0 \to \mathbb{R}_{\geq 0}$ defined as follows.

$$D^G(x) = \frac{1}{\sqrt{2}}\mathfrak{H}(f_x^G(y), f_x^{G^*}(y)). \tag{D.1}$$

Note that $D^G(x)$ is a function of x and it is indexed by G. By definition, $D^G(x)$ is simply a rescaling of the Hellinger distance between $f_x^G(y)$ and the density function$f_x^{G^*}(y)$. Here, we consider the Hellinger distance rather than squared Hellinger distance due to a technical reason. As shown later in this proof, there is a $n^{-1/2}$ term when we employ concentration inequality and we need to take a square of the prediction error, which is nearly parametric n^{-1}, in order to match that $n^{-1/2}$ term. For any G and x, $0 \leq D^G(x) \leq 1$ because of the boundness of Hellinger distance. We denote the collection of functions D^G by $\mathcal{H}_R$, i.e.,

$$\mathcal{H}_R = \{D^G(x) \mid G \text{ supported on } \mathrm{B}_p(0, R)\}.$$

Our main proof outline is based on the following decomposition

$$\begin{aligned}\int D^{\tilde{G}}(x)\mathrm{d}\mu(x) \leq & \frac{1}{n}\sum_{i=1}^n D^{\tilde{G}}(x_i) + \left|\frac{1}{n}\sum_{i=1}^n D^{\tilde{G}}(x_i) - \int D^{\tilde{G}}(x)\mathrm{d}\mu(x)\right| \\ \leq & \frac{1}{n}\sum_{i=1}^n D^{\tilde{G}}(x_i) + \sup_{D^G \in \mathcal{H}_R}\left|\frac{1}{n}\sum_{i=1}^n D^G(x_i) - \int D^G(x)\mathrm{d}\mu(x)\right|.\end{aligned} \tag{D.2}$$

The first term $\frac{1}{n}\sum_{i=1}^n D^{\tilde{G}}(x_i)$ is easily bounded using Theorem 3.1 and Cauchy-Schwarz inequality. We then focus on bounding the second term $\sup_{D^G \in \mathcal{H}_R}\left|\frac{1}{n}\sum_{i=1}^n D^G(x_i) - \int D^G(x)\mathrm{d}\mu(x)\right|$.

Since $\mathcal{H}_R$ is uniformly bounded by 1, by bounded differences concentration inequalities (see, e.g., Boucheron et al. (2013, Theorem 6.2)), we have

$$\begin{aligned}&\sup_{D^G \in \mathcal{H}_R}\left|\frac{1}{n}\sum_{i=1}^n D^G(x_i) - \int D^G(x)\mathrm{d}\mu(x)\right| \\ \leq & \mathbb{E}\left(\sup_{D^G \in \mathcal{H}_R}\left|\frac{1}{n}\sum_{i=1}^n D^G(x_i) - \int D^G(x)\mathrm{d}\mu(x)\right|\right) + t\sqrt{\frac{2}{n}\log n}\end{aligned} \tag{D.3}$$

with probability at least $1 - n^{-t^2}$.

We then bound the first term on the right hand side of (D.3). Note that this term is an expectation of a supremum across functions in $\mathcal{H}_R$. Then it is natural to bound it with a maximal inequality about empirical processes. The empirical process that we consider here is the mapping $\mathcal{H}_R \to \mathbb{R}$ given by

$$D^G \mapsto \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(D^G(x_i) - \int D^G(x)\mathrm{d}\mu(x)\right).$$

Since the function class $\mathcal{H}_R$ has an envelope function 1, by the maximal inequality (Van der Vaart, 2000, Lemma 19.34 and Corollary 19.35),

$$\mathbb{E}\sup_{D^G \in \mathcal{H}_R}\left|\frac{1}{n}\sum_{i=1}^n D^G(x_i) - \int D^G(x)\mathrm{d}\mu(x)\right| \leq \frac{C}{\sqrt{n}}J_{[]}(1, \mathcal{H}_R) \tag{D.4}$$

where C is a universal constant and $J_{[]}(1, \mathcal{H}_R, L^2(\mu))$ is the bracketing integral defined as

$$J_{[]}(1, \mathcal{H}_R, L^2(\mu)) = \int_0^1 \sqrt{1 + \log N_{[]}(\epsilon, \mathcal{H}_R, L^2(\mu))} \mathrm{d}\epsilon. \tag{D.5}$$

Here $N_{[]}(\epsilon, \mathcal{H}_R, L^2(\mu))$ refers to the bracketing number.Therefore (D.4) controls the discrepancy between the prediction error under random design and the prediction error under fixed design. By (D.3) and (D.4), we have

$$\sup_{D^G \in \mathcal{H}_R} \left| \frac{1}{n} \sum_{i=1}^n D^G(x_i) - \int D^G(x) \mathrm{d}\mu(x) \right| \leq \frac{C}{\sqrt{n}} J_{[]}(1, \mathcal{H}_R) + t\sqrt{\frac{2 \log n}{n}} \tag{D.6}$$

with probability at least $1 - n^{-t^2}$.

It remains to prove that $J_{[]}(1, \mathcal{H}_R)$ is finite so that the right hand side of (D.4) is at the scale of $O(n^{-1/2})$. This is the main goal in the remainder of this proof. Essentially we need to bound the bracketing number $N_{[]}(\epsilon, \mathcal{H}_R, L^2(\mu))$ in the bracketing integral (D.5).

Recall that $L^2(\mu)$ refers to metric induced by the L^2 norm under measure μ. The distance between D^{G_1} and D^{G_2} under $L^2(\mu)$ is $\left(\int [D^{G_1}(x) - D^{G_2}(x)]^2 \mathrm{d}\mu(x) \right)^{1/2}$. We consider another metric $\| \cdot \|_{\infty, S_0}$, and the distance between D^{G_1} and D^{G_2} under the metric $\| \cdot \|_{\infty, S_0}$ is $\sup_{x \in S_0} |D^{G_1}(x) - D^{G_2}(x)|$. It is clear that the metric $L^2(\mu)$ is no larger than the supremum metric $\| \cdot \|_{\infty, S_0}$. Therefore $N_{[]}(\epsilon, \mathcal{H}_R, L^2(\mu)) \leq N_{[]}(\epsilon, \mathcal{H}_R, \| \cdot \|_{\infty, S_0})$. We will bound $N_{[]}(\epsilon, \mathcal{H}_R, \| \cdot \|_{\infty, S_0})$ instead because the bracketing number under $\| \cdot \|_{\infty, S_0}$ is closely related to our metric entropy result in Theorem E.1.

For any two probability measures G_1 and G_2 on $\mathbb{R}^p$ and any $x \in \mathbb{R}^p$,

$$\begin{aligned}
|D^{G_1}(x) - D^{G_2}(x)| &= \frac{1}{\sqrt{2}} \left| \sqrt{1 - 2 \int \sqrt{f_x^{G_1}(y) f_x^{G^*}(y)} \mathrm{d}y} - \sqrt{1 - 2 \int \sqrt{f_x^{G_2}(y) f_x^{G^*}(y)} \mathrm{d}y} \right| \\
&\leq \sqrt{\int \sqrt{f_x^{G^*}(y)} \left| \sqrt{f_x^{G_1}(y)} - \sqrt{f_x^{G_2}(y)} \right| \mathrm{d}y} \\
&\leq \sqrt{\int \sqrt{f_x^{G^*}(y)} \sqrt{|f_x^{G_1}(y) - f_x^{G_2}(y)|} \mathrm{d}y},
\end{aligned} \tag{D.7}$$

where the last two inequalities use the basic inequality $\sqrt{a} - \sqrt{b} \leq \sqrt{|a - b|}$.

Further, by Jensen's inequality, we have

$$\begin{aligned}
\sqrt{\int \sqrt{f_x^{G^*}(y)} \sqrt{|f_x^{G_1}(y) - f_x^{G_2}(y)|} \mathrm{d}y} &= \sqrt{\int \sqrt{f_x^{G^*}(y) |f_x^{G_1}(y) - f_x^{G_2}(y)|} \mathrm{d}y} \\
&\leq \left(\int f_x^{G^*}(y) |f_x^{G_1}(y) - f_x^{G_2}(y)| \mathrm{d}y \right)^{1/4} \\
&\leq \Big(\sup_{(x,y) \in S_0 \times \mathbb{R}} |f_x^{G_1}(y) - f_x^{G_2}(y)| \int f_x^{G^*}(y) \mathrm{d}y \Big)^{1/4} \\
&= \sup_{(x,y) \in S_0 \times \mathbb{R}} |f_x^{G_1}(y) - f_x^{G_2}(y)|^{1/4}
\end{aligned} \tag{D.8}$$

where we use the fact that $\int f_x^{G^*}(y) = 1$ in the last line.

To conclude (D.7) and (D.8), we have

$$|D^{G_1}(x) - D^{G_2}(x)| \leq \sup_{(x,y)\in S_0\times\mathbb{R}} |f_x^{G_1}(y) - f_x^{G_2}(y)|^{1/2}, \forall x \in S_0. \tag{D.9}$$

Recall that $\mathcal{M}_R$ is a class of conditional density functions according to the definition in (3.5). Let $f^{G_1}, \ldots, f^{G_N}$ be a $(\epsilon/2)^4$-net of $\mathcal{M}_R$ under the $\|\cdot\|_{\infty,S_0\times\mathbb{R}}$ metric, we claim that $[D^{G_i}-\epsilon/2, D^{G_i}+\epsilon/2]$, $i = 1, \ldots, N$ form an ϵ-bracketing cover of $\mathcal{H}_R$. Indeed, for any $D^G \in \mathcal{H}_R$, there exists $G_j \in \{G_1, \ldots, G_N\}$ such that

$$\|f^G - f^{G_j}\|_{\infty,S_0\times\mathbb{R}} \leq \left(\frac{\epsilon}{2}\right)^4.$$

By (D.9),

$$\|D^G - D^{G_j}\|_{\infty,S_0} \leqslant \frac{\epsilon}{2},$$

so for all $x \in S_0$, $D^{G_j}(x) - \epsilon/2 \leqslant D^G(x) \leqslant D^{G_j}(x) + \epsilon/2$ and D^G belongs to the bracket $[D^{G_j} - \epsilon/2, D^{G_j} + \epsilon/2]$.

Therefore,

$$N_{[]}(\epsilon, \mathcal{H}_R, \|\cdot\|_{\infty,S_0}) \leq N((\epsilon/2)^4, \mathcal{M}_R, \|\cdot\|_{\infty,S_0\times\mathbb{R}}),$$

and $\log N_{[]}(\epsilon, \mathcal{H}_R, \|\cdot\|_{\infty,S_0}) \leq \log N((\epsilon/2)^4, \mathcal{M}_R, \|\cdot\|_{\infty,S_0\times\mathbb{R}})$.

We invoke metric entropy bound in Theorem E.1 by taking η in Theorem E.1 as $(\epsilon/2)^4$ and S_0 as the support set of μ. As defined in Theorem E.1, $\mathfrak{L}_{S_0} = \sup_{x\in S_0} \mathfrak{L}(x) \leq \sup_{x\in S_0} \|x\| \leq B$. Therefore,

$$\begin{aligned}
&\log N_{[]}(\epsilon, \mathcal{H}_R, \|\cdot\|_{\infty,S_0}) \\
\leq &\log N((\epsilon/2)^4, \mathcal{M}_R, \|\cdot\|_{\infty,S_0\times\mathbb{R}}) \\
\leq &C_p\zeta^p N(\frac{\sigma}{\mathfrak{L}_{S_0}}\sqrt{2\log\frac{48}{\sigma\epsilon^4}}, \mathrm{B}_p(0,R)) \left(\log\frac{16}{\sigma\epsilon^4}\right)^{p+1}
\end{aligned} \tag{D.10}$$

when $(\epsilon/2)^4 < e^{-1}\sigma^{-1}$, i.e., $\epsilon < 2(e\sigma)^{-1/4}$.

By existing results on bounding Euclidean covering numbers, we have

$$\begin{aligned}
&N(\frac{\sigma}{\mathfrak{L}_{S_0}}\sqrt{2\log\frac{48}{\sigma\epsilon^4}}, \mathrm{B}_p(0,R)) \\
\leq &\left(1 + \frac{2R\mathfrak{L}_{S_0}}{\{2\log(48\sigma^{-1}\epsilon^{-4})\}^{1/2}\sigma}\right)^p
\end{aligned}$$

Plugging the result back to the metric entropy bound and bracketing covering number, we have

$$\log N_{[]}(\epsilon, \mathcal{H}_R, \|\cdot\|_{\infty,S_0}) \leq C_p\zeta^p \left(1 + \frac{2R\mathfrak{L}_{S_0}}{\{2\log(48\sigma^{-1}\epsilon^{-4})\}^{1/2}\sigma}\right)^p \left(\log\frac{16}{\sigma\epsilon^4}\right)^{p+1}.$$

On the other hand, since the $L^2(\mu)$ norm is bounded by the uniform norm $\|\cdot\|_{\infty,S_0}$, it follows that

$$\begin{aligned}
\log N_{[]}(\epsilon, \mathcal{H}_R, L^2(\mu)) &\leq \log N_{[]}(\epsilon, \mathcal{H}_R, \|\cdot\|_{\infty,S_0}) \\
&\leq C_p'\zeta^p \left(1 + \frac{2R\mathfrak{L}_{S_0}}{\{2\log(48\sigma^{-1}\epsilon^{-4})\}^{1/2}\sigma}\right)^p \left(\log\frac{16}{\sigma\epsilon^4}\right)^{p+1}.
\end{aligned} \tag{D.11}$$

Note that the bracketing entropy is non-increasing in ϵ, we can then bound $J_{[]}(1, \mathcal{H}_R, L^2(\mu))$ by quantity $\rho_p(\mathfrak{L}_{S_0}, R, \sigma)$ defined as

$$\begin{aligned}
&\rho_p(\mathfrak{L}_{S_0}, R, \sigma) \\
= &\max((e\sigma)^{-1/4})/2, 1) \int_0^{\min(1,2(e\sigma)^{-1/4})} \sqrt{1 + C_p'\zeta^p \left(1 + \frac{2R\mathfrak{L}_{S_0}}{\{2\log(48\sigma^{-1}\epsilon^{-4})\}^{1/2}\sigma}\right)^p \left(\log\frac{16}{\sigma\epsilon^4}\right)^{p+1}} \mathrm{d}\epsilon.
\end{aligned}$$

where the term $2(e\sigma)^{-1/4}$ appearing is solely because the metric entropy bound in (D.10) is only proved for covering that are smaller enough. It is not hard to verify that this integral in $\rho_p(\mathfrak{L}_{S_0}, R, \sigma)$ is indeed finite through of change of variables $\delta = 1/\epsilon$ and using the fact that $\int_1^\infty \frac{(\log \delta)^r}{\delta^2} \mathrm{d}\delta < \infty$ for any $r > 0$.

By Theorem 3.1 and Cauchy-Schwarz inequality,

$$\frac{1}{n}\sum_{i=1}^n D^{\tilde{G}}(x_i) \le \sqrt{\frac{1}{n}\sum_{i=1}^n (D^{\tilde{G}}(x_i))^2} \le tC_p\epsilon_n \tag{D.12}$$

with probability at least $1 - n^{-t^2}$, where $epsilon_n$ is defined the same as in Theorem 3.1.

Now we can bound $\int D^{\tilde{G}}(x)\mathrm{d}\mu(x)$ as outlined in (D.2),

$$\begin{aligned}
\int D^{\tilde{G}}(x)\mathrm{d}\mu(x) \le & \frac{1}{n}\sum_{i=1}^n D^{\tilde{G}}(x_i) + \left|\frac{1}{n}\sum_{i=1}^n D^{\tilde{G}}(x_i) - \int D^{\tilde{G}}(x)\mathrm{d}\mu(x)\right| \\
\le & \frac{1}{n}\sum_{i=1}^n D^{\tilde{G}}(x_i) + \sup_{D^G \in \mathcal{H}_R}\left|\frac{1}{n}\sum_{i=1}^n D^{G}(x_i) - \int D^{G}(x)\mathrm{d}\mu(x)\right| \\
\le & tC_p\epsilon_n + \frac{C}{\sqrt{n}}\rho_p(\mathfrak{L}_{S_0}, R, \sigma) + t\sqrt{\frac{2\log n}{n}}
\end{aligned}$$

with probability at least $1 - 2n^{-t^2}$. Here we use the union bound of probabilities. Lastly, we can set $t = 1$, and we further relax $\rho_p(\mathfrak{L}_{S_0}, R, \sigma)$ to $\rho_p(B, R, \sigma)$ since $\mathfrak{L}_{S_0} \le B$. Specifically, $\zeta = 1$ corresponds to the linear regressions. □

Our second topic of this section is weak consistency of $\hat{G}$, summarized in Theorem 3.4. This result is built upon the prediction error bound in Theorem 3.3. We utilize characteristic functions in the proof of Theorem 3.4, and this part is greatly inspired by Beran and Millar (1994).

Recall that under the random design setting, the mixture of linear regressions model is

$$Y_i = X_i^\top \beta_i + \sigma Z_i, \beta \sim G^*, X_i \sim \mu, Z_i \sim N(0, \sigma^2). \tag{D.13}$$

We review and introduce the following notation throughout the proof of weak consistency. We use K to refer to a compact set in $\mathbb{R}^p$, and more specifically $\mathrm{B}_p(0, R)$ to be consistent with the notation in previous sections. $\mathcal{G}_K$ is the class of all probability measures supported on K. $G^* \in \mathcal{G}_K$ is some probability measure of β. $\hat{G}_n \in \mathcal{G}_K$ is an NPMLE from n sample points $\{(x_i, y_i)\}_{i=1}^n$. μ is the probability measure of X_i. $P(G^*, \mu)$ is the joint probability measure of (X_i, Y_i) under model (D.13). $\hat{P}_n$ is the empirical measure of $\{(x_i, y_i)\}_{i=1}^n$. $\hat{F}_{x,n}$ is the empirical measure of $\{x_i\}_{i=1}^n$. d is any metric that metrizes the weak convergence of probability measures on Euclidean spaces.

Before proving Theorem 3.4, we state and prove two lemmas, Lemma D.1 and Lemma D.2. Lemma D.1 is regarding the continuity of probability measure consistency and Lemma D.2 is regarding the identifiability of the mixture of linear regression models. Both lemmas provide key ingredients to our proof of Theorem 3.4.

Lemma D.1. *Suppose G_n is some probability measure supported on K, $n = 1, 2, \ldots$. If*

$$d(G_n, G^*) \to 0, d(F_{x,n}, \mu) \to 0,$$

then

$$d[P(G_n, F_{x,n}), P(G^*, \mu)] \to 0.$$

Proof of Lemma D.1. The following proof of this lemma goes similarly as the proof of Proposition 2.1 of Beran and Millar (1994). It suffices to show that the characteristic function of $P(G_n, F_{x,n})$ converges to the characteristic function of $P(G^*, \mu)$.

The characteristic function of $P(G_n, F_{x,n})$ is $\mathbb{E}_{(X,Y)\sim P(G_n,F_{x,n})} e^{iu^\top X + itY}$. To calculate this expectation, we first take expectation conditioning on x, so

$$\begin{aligned}\mathbb{E}_{(X,Y)\sim P(G_n,F_{x,n})} e^{iu^\top X+itY} &= \mathbb{E}_{X\sim F_{x,n}}\left[e^{iu^\top X}\mathbb{E}_{Z\sim N(0,\sigma^2),\beta\sim G_n}(e^{itZ}e^{itX^\top\beta}|X)\right] \\ &= \int e^{iu^\top x}\mathbb{E}e^{itZ}\mathbb{E}_{G_n}e^{itx^\top\beta}\mathrm{d}F_{x,n}(x).\end{aligned}$$

Similarly, the characteristic function of $P(G^*, \mu)$ is $\mathbb{E}_{(X,Y)\sim P(G^*,\mu)} e^{iu^\top X+itY}$, and it can be calculated as

$$\int e^{iu^\top x}\mathbb{E}e^{itZ}\mathbb{E}_{G^*}e^{itx^\top\beta}\mathrm{d}\mu(x).$$

Therefore, to show the characteristic function of $P(G_n, F_{x,n})$ converges to the characteristic function of $P(G^*, \mu)$, it is equivalent to show that

$$\int e^{iu^\top x}\mathbb{E}e^{itZ}\mathbb{E}_{G_n}e^{itx^\top\beta}\mathrm{d}F_{x,n}(x)$$

converges to

$$\int e^{iu^\top x}\mathbb{E}e^{itZ}\mathbb{E}_{G^*}e^{itx^\top\beta}\mathrm{d}\mu(x)$$

for all $u \in \mathbb{R}^p, t \in \mathbb{R}$.

We introduce shorthand notation $p_n(x) = e^{iu^\top x}\mathbb{E}e^{itZ}\mathbb{E}_{G_n}e^{itx^\top\beta}$ and $p(x) = e^{iu^\top x}\mathbb{E}e^{itZ}\mathbb{E}_{G^*}e^{itx^\top\beta}$, then we need to show that

$$\int p_n(x)\mathrm{d}F_{x,n}(x) \to \int p(x)\mathrm{d}\mu(x).$$

By the weak convergence of G_n to G^*, $p_n(x)$ converges to $p(x)$ for all $x \in \mathbb{R}^p$. $\{p_n(x)\}_{n\geq 1}$ and $p(x)$ are equicontinuous and bounded by 1, thus $p_n(x)$ converges uniformly to $p(x)$ on any compact set of x in $\mathbb{R}^p$.

$F_{x,n}$ converges weakly to μ, thus there exists a compact set K_δ such that $\mu(K_\delta) < \delta$ and $F_{x,n}(K_\delta) < \delta$ for n sufficiently large.

$$\left|\int p_n\mathrm{d}F_{n,x}(x) - \int p(x)\mathrm{d}\mu(x)\right| \leq \int_{K_\delta}|p_n - p|\mathrm{d}F_{n,x}(x) + \left|\int_{K_\delta} p\mathrm{d}(F_{n,x}-\mu)(x)\right| + 2\delta. \tag{D.14}$$

The first term in (D.14) is less than δ for sufficiently large n by uniform convergence of x, and the second term in (D.14) is less than δ for sufficiently large n by the weak convergence of $F_{x,n}$. Since the above argument holds for arbitrarily small $\delta > 0$, we complete the proof by letting δ go to 0.

□

Lemma D.2. *Let $\{G_n\}$ denote a series of probability measures in $\mathcal{G}_K$, and let $\{F_{x,n}\}$ denote a series of probability measures on $\mathbb{R}^p$. Assuming* $\mathrm{supp}(\mu)$ *contains an open set in $\mathbb{R}^p$. If*

$$d[P(G_n, F_{x,n}), P(G^*, F_x)] \to 0,$$

then

$$d(G_n, G^*) \to 0.$$

Proof of Lemma D.2. This proof of this lemma goes similarly as the proof of Proposition 2.2 of Beran and Millar (1994). By definition of P and $d[P(G_n, F_{x,n}), P(G^*, F_x)] \to 0$, it follows that $d(F_{x,n}, \mu) \to 0$. Since $\{G_n\}$ is supported on a compact set K, $\{G_n\}$ is tight. $\{G_n\}$ has a subsequence converging weakly (see, e.g., Theorem 3.10.3 in Durrett (2019)). Let $\tilde{G}$ denote the limiting probability measure of the weakly convergent subsequence. The convergence of the characteristic functions gives

$$\lim_{n\to\infty} \int e^{iu^\top x} \mathbb{E} e^{itZ} \mathbb{E}_{G^*} e^{itx^\top \beta} \mathrm{d}F_{x,n}(x) = \int e^{iu^\top x} \mathbb{E} e^{itZ} \mathbb{E}_{\tilde{G}} e^{itx^\top \beta} \mathrm{d}\mu(x).$$

By the convergence of $F_{x,n}$ to F_x, it follows that

$$\int e^{iu^\top x} \mathbb{E} e^{itZ} \mathbb{E}_{\tilde{G}} e^{itx^\top \beta} \mathrm{d}F_x(x) = \int e^{iu^\top x} \mathbb{E} e^{itZ} \mathbb{E}_{G^*} e^{itx^\top \beta} \mathrm{d}\mu(x)$$

for all $u \in \mathbb{R}^p, t \in \mathbb{R}$.

Thus for all x in the support of μ,

$$\mathbb{E}_{\beta\sim\tilde{G}} e^{itx^\top \beta} = \mathbb{E}_{\beta\sim G^*} e^{itx^\top \beta}. \tag{D.15}$$

Both sides of the equation (D.15) are bounded and thus analytic as function of x. Since μ contains a open set, (D.15) holds for all $x \in \mathbb{R}^p$. Because (D.15) also holds for all t, $\tilde{G}$ and G^* have the same characteristic functions. Thus $\tilde{G}$ and G^* are identical. Every weakly convergent subsequence of $\{G_n\}$ has the same limit, which is equivalent to that $\{G_n\}$ converges weakly to G^*.

□

Now we are ready to show the proof of Theorem 3.4.

Proof of Theorem 3.4. When the support S_0 of μ is compact, then Hellinger distance bound holds with $\mathfrak{L}_{S_0} = \sup_{x\in S_0} \|x\|$ for any sequence $\{x_i\}_{i=1}^n$. More specifically, we have proved in Theorem 3.3 that

$$\mathfrak{H}_n^2(\hat{f}, f^*) = \mathfrak{H}^2[P(\hat{G}_n, \hat{F}_{x,n}), P(G^*, \hat{F}_{x,n})] = O_p\left(\frac{(\log n)^{p+1}}{n}\right).$$

Since convergence under the Hellinger distance implies weak convergence, we have

$$d[P(\hat{G}_n, \hat{F}_{x,n}), P(G^*, \hat{F}_{x,n})] \to 0 \text{ in probability.} \tag{D.16}$$

By weak convergence of empirical measures,

$$d(\hat{F}_{x,n}, \mu) \to 0 \text{ in probability.}$$

By Lemma D.1,

$$d[P(G^*, \hat{F}_{x,n}), P(G^*, \mu)] \to 0 \text{ in probability.} \tag{D.17}$$

Applying triangle inequality with (D.16) and (D.17), we have

$$d[P(\hat{G}_n, \hat{F}_{x,n}), P(G^*, F_x)] \to 0 \text{ in probability.}$$

By Lemma D.2,

$$d(\hat{G}_n, G^*) \to 0 \text{ in probability.}$$

Lemma D.1 and Lemma D.2 are directly applicable when almost surely convergence holds rather than merely convergence in probability, but we can get around this by using the subsequences argument (see, e.g., Theorem 2.3.2 in Durrett (2019)).

□

Appendix E Metric entropy results

In this section, we prove our metric entropy results, and these results provide key ingredients for the proof of Theorem 3.1. Let us first formally define the notion of metric entropy. Let T be a subset of a metric space with metric $\mathfrak{d}$. For $\eta > 0$, we say that a set S is an η-covering of T if $\sup_{t\in T} \inf_{s\in S} \mathfrak{d}(s,t) \leq \eta$. The smallest possible cardinality of an η-covering of T is known as the η-covering number of T under the metric $\mathfrak{d}$ and this is denoted by $N(\eta, T, \mathfrak{d})$. The logarithm of $N(\eta, T, \mathfrak{d})$ is called the η-metric entropy of T under $\mathfrak{d}$. When T is a subset of $\mathbb{R}^p$ and the metric $\mathfrak{d}$ is the usual Euclidean metric on $\mathbb{R}^p$, we shall denote $N(\eta, T, \mathfrak{d})$ by simply $N(\eta, T)$.

The main theorem of this section is Theorem E.1. We work here under a more general setting than linear regression functions. Specifically, we use the function $r(x, \beta)$ to represent the mean of the response y given x and β so that the conditional density function of y given x is

$$f_x^G(y) := \int \frac{1}{\sigma}\phi\left(\frac{y - r(x,\beta)}{\sigma}\right) \mathrm{d}G(\beta).$$

Let K denote an arbitrary compact set in $\mathbb{R}^p$ and

$$\mathcal{M}_K := \{f_x^G(y) : G \text{ is a probability measure supported on } K\}. \tag{E.1}$$

The goal of this section is to prove an upper bound on the covering number $N(\eta, \mathcal{M}_K, \|\cdot\|_{\infty, S_0\times\mathbb{R}})$ of $\mathcal{M}_K$ under the metric $\|\cdot\|_{\infty,S_0\times\mathbb{R}}$:

$$\sup_{x\in S_0, y\in\mathbb{R}} \left|f_x^G(y) - f_x^{G'}(y)\right|.$$

for an arbitrary set S_0 of x-values. For each x, let $\mathfrak{L}(x)$ be defined as

$$\mathfrak{L}(x) := \sup_{\beta_1,\beta_2\in K:\beta_1\neq\beta_2} \frac{|r(x,\beta_1) - r(x,\beta_2)|}{\|\beta_1 - \beta_2\|} \tag{E.2}$$

so that

$$|r(x,\beta_1) - r(x,\beta_2)| \leq \mathfrak{L}(x)\|\beta_1 - \beta_2\| \qquad \text{for all } \beta_1, \beta_2 \in K.$$

The following is the main theorem of this section.

Theorem E.1. *Suppose that, for every x, the function $\beta \mapsto r(x,\beta)$ is a polynomial function of degree at most ζ. Then there exists a constant C_p depending only on p such that for every $0 < \eta < e^{-1}\sigma^{-1}$, we have*

$$\log N(\eta, \mathcal{M}_K, \|\cdot\|_{\infty,S_0\times\mathbb{R}}) \leq C_p\zeta^p N\left(\frac{\sigma}{\mathfrak{L}_{S_0}}\sqrt{2\log\frac{3}{\sigma\eta}}, K\right)\left(\log\frac{1}{\sigma\eta}\right)^{p+1}, \tag{E.3}$$

where $\mathfrak{L}_{S_0} = \sup_{x\in S_0}\mathfrak{L}(x)$.

We prove Theorem E.1 by modifying appropriately the proof of the metric entropy results for Gaussian location mixtures in Zhang (2009) (see also Ghosal and Van Der Vaart (2007) and Saha and Guntuboyina (2020)). Actually Theorem E.1 can be seen as a generalization of metric entropy results for Gaussian location mixtures. Indeed, in the special case when $p = 1$, $\sigma = 1$, $S_0 = \{0\}$, $r(x,\beta) = \beta$ and $K = [-M, M]$ (for some $M > 0$), the class $\mathcal{M}_K$ becomes

$$\mathcal{H}_M := \left\{y \mapsto \int \phi(y-\beta)dG(\beta) : G[-M, M] = 1\right\}$$

and inequality (E.3) gives that the η-metric entropy of $\mathcal{H}_M$ under the L_∞ metric on $\mathbb{R}$ is bounded by

$$CN\left(\sqrt{2\log\frac{3}{\eta}}, [-M, M]\right)\left(\log\frac{1}{\eta}\right)^2 \leq C\left(1 + \frac{2M}{\sqrt{2\log(3/\eta)}}\right)\left(\log\frac{1}{\eta}\right)^2$$

for all $0 < \eta < e^{-1}$. This is essentially Zhang (2009, inequality (5.8)).

The proof of Theorem E.1 crucially relies on Lemma E.1 (moment matching accuracy) and Lemma E.2 (approximation by discrete mixtures) which are given next. Lemma E.1 follows almost directly from the corresponding result for Gaussian location mixtures (see Jiang and Zhang (2009, Lemma 1) or Saha and Guntuboyina (2020, Lemma D.2)) but Lemma E.2 requires additional arguments.

Lemma E.1. *Fix a pair (x, y) and let A be a subset of $\mathbb{R}^p$ such that*

$$\mathring{O}((x, y), a) \subseteq A \subseteq O((x, y), ca)$$

for some $a > 1$ and $c \geq 1$ where

$$O((x, y), a) = \{\beta \in K : |y - r(x, \beta)|/\sigma \leq a\}.$$

and

$$\mathring{O}((x, y), a) = \{\beta \in K : |y - r(x, \beta)|/\sigma < a\}.$$

Let G and G' be two probability measures on $\mathbb{R}^p$ such that for some $m \geq 1$ and all integers $0 \leq k \leq 2m$, we have

$$\int_A \{r(x, \beta)\}^k \mathrm{d}G(\beta) = \int_A \{r(x, \beta)\}^k \mathrm{d}G'(\beta). \tag{E.4}$$

Then

$$|f_x^G(y) - f_x^{G'}(y)| \leq \frac{1}{2\pi\sigma}\left(\frac{c^2a^2e}{2(m+1)}\right)^{m+1} + \frac{2e^{-a^2/2}}{(2\pi)^{1/2}\sigma}. \tag{E.5}$$

Proof of Lemma E.1. This result follows from the moment matching lemma for the univariate Gaussian location mixtures in Jiang and Zhang (2009, Lemma 1) or Saha and Guntuboyina (2020, Lemma D.2). These results are stated for the $\sigma = 1$ case but the extension to arbitrary σ is straightforward. □

Lemma E.2. *Let G be a probability measure supported on K. For every $a \geq 1$, there exists a discrete probability measure G' supported on at most*

$$(2\lfloor 13.5a^2 \rfloor \zeta + 1)^p N(a\sigma/\mathfrak{L}_{S_0}, K) + 1, \tag{E.6}$$

points in K such that

$$\sup_{(x,y)\in S_0\times\mathbb{R}} |f_x^G(y) - f_x^{G'}(y)| \leq \left(1 + \frac{1}{\sqrt{2\pi}}\right)\frac{e^{-a^2/2}}{(2\pi)^{1/2}\sigma}. \tag{E.7}$$

Proof of Lemma E.2. Let us introduce a pseudometric $d_{S_0,r}$ on K as

$$d_{S_0,r}(\beta_1, \beta_2) = \sup_{x\in S_0} |r(x, \beta_1) - r(x, \beta_2)|/\sigma. \tag{E.8}$$

Fix $a \geq 1$ and let $L := N(a, K, d)$ denote the a-covering number of K under the pseudometric $d_{S_0,r}$. Let $E_1, \ldots, E_L$ denote balls of radius a (with respect to $d_{S_0,r}$) within K whose union is equal to K, and

let $\bar{\beta}_1, \ldots, \bar{\beta}_L$ denote the centers of these balls. Then we define $B_1 = E_1$ and $B_i = E_i \cap (\cup_{j=1}^{i-1} B_j)^c$ for $i = 2, \ldots, L$. Let $m = \lfloor 13.5a^2 \rfloor$ and consider the following collection of $(2m\zeta+1)^p L$-dimensional vectors:

$$T_{int} := \left\{ \left(\int \beta_1^{k_1} \ldots \beta_p^{k_p} \mathbb{I}\{\beta \in B_i\} \mathrm{d}G(\beta) \right)_{0 \le k_1, \ldots, k_p \le 2m\zeta, 1 \le i \le L} : G \text{ is any probability measure over } K \right\}.$$

By standard results, it follows that T_{int} is the convex hull of

$$T := \left\{ \left(\beta_1^{k_1} \ldots \beta_p^{k_p} \mathbb{I}\{\beta \in B_i\} \right)_{0 \le k_1, \ldots, k_p \le 2m\zeta, 1 \le i \le L} : \beta \in K \right\}.$$

This follows, for example, by Parthasarathy (2005, Theorem 6.3) and the fact that T is closed. Notice that both T_{int} and T lie in the Euclidean space of dimension $(2m\zeta+1)^p L$. By Carathéodory's theorem, any vector in T_{int} can be written as a convex combination of at most $\{(2m\zeta+1)^p L + 1\}$ elements in T. This implies that for every probability measure G on K, there exists a discrete measure G' which is supported on a discrete subset of K of cardinality at most $\{(2m\zeta+1)^p L + 1\}$ such that

$$\int_{B_i} \beta_1^{k_1} \ldots \beta_p^{k_p} \mathrm{d}G(\beta) = \int_{B_i} \beta_1^{k_1} \ldots \beta_p^{k_p} \mathrm{d}G'(\beta) \text{ for } 0 \le k_1, \ldots, k_p \le 2m\zeta, \text{ and } 1 \le i \le L. \tag{E.9}$$

Fix $x \in S_0$ and $y \in \mathbb{R}$. We shall prove the bound (E.7) for $|f_x^G(y) - f_x^{G'}(y)|$ by using Lemma E.1. First note that since $\mathring{O}((x,y),a)$ is contained in K, the sets $B_1, \ldots, B_L$ cover $\mathring{O}((x,y),a)$. Let $F := \{1 \le i \le L : B_i \bigcap \mathring{O}((x,y),a) \ne \emptyset\}$ so that

$$\mathring{O}((x,y),a) \subseteq \bigcup_{i \in F} B_i.$$

We shall prove below that

$$\bigcup_{i \in F} B_i \subseteq O((x,y),3a) \tag{E.10}$$

which will enable us to apply Lemma E.1 with $A = \bigcup_{i \in F} B_i$. To see (E.10), note that for each fixed $i \in F$, there exists $\beta_0 \in B_i$ such that $\beta_0 \in \mathring{O}((x,y),a)$, i.e. $|y - r(x,\beta_0)|/\sigma \le a$. As the diameter of B_i (under the metric $d_{S_0,r}$) is atmost $2a$, it follows that $d_{S_0,r}(\beta,\beta_0) \le 2a$ for every $\beta \in B_i$. Consequently

$$|y - r(x,\beta)|/\sigma \le |y - r(x,\beta_0)|/\sigma + |r(x,\beta) - r(x,\beta_0)|/\sigma \le a + d_{S_0,r}(\beta,\beta_0) \le 3a.$$

This proves (E.10). In order to apply Lemma E.1, we need to next check that inequalty (E.4) holds. This basically follows from (E.9) and the fact that $r(x,\beta)$ is assumed to be a polynomial function of the components of β with degree ζ (this will ensure that the terms being integrated on both sides of (E.4) are polynomials of components of β with degree up to $2m\zeta$). Lemma E.1 can thus be applied (with $A = \bigcup_{i \in F} B_i$ and $c = 3$) which gives

$$|f_x^G(y) - f_x^{G'}(y)| \le \frac{1}{2\pi\sigma} \left(\frac{9a^2 e}{2(m+1)} \right)^{m+1} + \frac{e^{-a^2/2}}{(2\pi)^{1/2}\sigma}.$$

Because $m = \lfloor 13.5a^2 \rfloor$, we have $m + 1 \ge 13.5a^2$ and

$$\left(\frac{9a^2 e}{2(m+1)} \right)^{m+1} \le \left(\frac{e}{3} \right)^{m+1} \le \exp(-\frac{m+1}{12}) \le \exp\left(-\frac{27a^2}{24} \right) \le e^{-a^2/2},$$

where we used the simple fact that $e/3 \le e^{-1/12}$. This proves (E.7). It remains to prove that the cardinality of the support of G' is at most (E.6). As we have already seen that the cardinality of the

support of G' is at most $\{(2m\zeta+1)^p L+1\}$, we only need to show that $L = N(a, K, d)$ is at most the Euclidean covering number $N(a\sigma/\mathfrak{L}_{S_0}, K)$. For this, note that by definition of $\mathfrak{L}_{S_0}$, we have

$$d_{S_0,r}(\beta_1,\beta_2) = \sup_{x\in S_0} |r(x,\beta_1) - r(x,\beta_2)|/\sigma \le \mathfrak{L}_{S_0}\sigma^{-1}\|\beta_1-\beta_2\|,$$

for every β_1, β_2. This gives

$$N(a, K, d_{S_0,r}) \le N(a\sigma/\mathfrak{L}_{S_0}, K) \tag{E.11}$$

which completes the proof of Lemma E.2. □

Proof of Theorem E.1. Fix a probability measure G that is supported on K. By Lemma E.2, for each fixed $a \ge 1$, there exists a probability measure G' supported on K such that

$$\sup_{(x,y)\in S_0\times\mathbb{R}} |f_x^G(y) - f_x^{G'}(y)| \le \left(1+\frac{1}{\sqrt{2\pi}}\right)\frac{e^{-a^2/2}}{(2\pi)^{1/2}\sigma},$$

and such that the cardinality of the support of G' is at most ℓ where ℓ is given by (E.6).

Now let $\alpha = \nu = e^{-a^2/2}$. Let $s_1, \ldots, s_{N_1}$ be an α-covering of K under the $d_{S_0,r}$ pseudometric (defined in (E.8)) where (via (E.11))

$$N_1 := N(\alpha, K, d_{S_0,r}) \le N(\alpha\sigma/\mathfrak{L}_{S_0}, K) \tag{E.12}$$

Also let $t_1, \ldots, t_{N_2}$ be a ν-covering of the probability simplex $\Delta_\ell := \{(p_1,\ldots,p_\ell) : p_j \ge 0, \sum_j p_j = 1\}$ under the L^1-metric $(p,q) \mapsto \sum_j |p_j - q_j|$ where $N_2 := N(\nu, \Delta_\ell, L_1)$. We can write $G' = \sum_{i=1}^\ell w_i\delta_{a_i}$ for some $(w_1,\ldots,w_\ell) \in \Delta_\ell$ and $a_1,\ldots,a_\ell \in K$. Since $s_1,\ldots,s_{N_1}$ form an α-covering of K, we can find ℓ (not necessarily distinct) elements $s_{G'1},\ldots,s_{G'\ell}$ from $\{s_1,\ldots,s_{N_1}\}$ such that $d_{S_0,r}(a_i, s_{G'i}) \le \alpha, i = 1,\ldots,\ell$. Letting $G'' = \sum_{i=1}^\ell w_i\delta_{s_{G'i}}$, we have

$$\begin{aligned}
|f_x^{G'}(y) - f_x^{G''}(y)| &= \frac{1}{\sigma}\left|\sum_{i=1}^\ell w_i\phi\left(\frac{y-r(x,a_i)}{\sigma}\right) - \sum_{i=1}^\ell w_i\phi\left(\frac{y-r(x,s_{G'i})}{\sigma}\right)\right| \\
&\le \frac{1}{\sigma}\sum_{i=1}^\ell w_i\cdot\left|\phi\left(\frac{y-r(x,a_i)}{\sigma}\right) - \phi\left(\frac{y-r(x,s_{G'i})}{\sigma}\right)\right| \\
&\le \frac{1}{\sigma}\sum_{i=1}^\ell w_i\cdot\sup_z|\phi'(z)|\cdot d_{S_0,r}(a_i,s_{G'i}) \le \alpha\frac{e^{-1/2}}{(2\pi)^{1/2}\sigma}
\end{aligned}$$

for every $x \in S_0$ and $y \in \mathbb{R}$. Also since $t_1,\ldots,t_{N_2}$ is a ν-covering of Δ_ℓ under the L^1 metric, there exist $t_{G'1},\ldots,t_{G'\ell}$ from $\{t_1,\ldots,t_{N_2}\}$ such that $\sum_{i=1}^\ell |t_{G'i} - w_i| \le \nu$. Denote $G''' = \sum_{i=1}^\ell t_{G'i}\delta_{s_{G'i}}$, then for every $x \in S_0$ and any $y \in \mathbb{R}$, we have

$$\begin{aligned}
|f_x^{G''}(y) - f_x^{G'''}(y)| &= \frac{1}{\sigma}\left|\sum_{i=1}^\ell w_i\phi\left(\frac{y-r(x,s_{G'i})}{\sigma}\right) - \sum_{i=1}^\ell t_{G'i}\phi\left(\frac{y-r(x,s_{G'i})}{\sigma}\right)\right| \\
&\le \frac{1}{\sigma}\sum_{i=1}^\ell |w_i - t_{G'i}|\cdot\phi\left(\frac{y-r(x,s_{G'i})}{\sigma}\right) \le \frac{\nu}{\sigma}\cdot\sup_z|\phi(z)| \le \frac{\nu}{\sigma}\frac{1}{(2\pi)^{1/2}}.
\end{aligned}$$

Combining three inequalities together, we have

$$\begin{aligned}
|f_x^G(y) - f_x^{G'''}(y)| &\le |f_x^G(y) - f_x^{G'}(y)| + |f_x^{G'}(y) - f_x^{G''}(y)| + |f_x^{G''}(y) - f_x^{G'''}(y)| \\
&\le \left(1+(2\pi)^{-1/2}\right)\frac{e^{-a^2/2}}{(2\pi)^{1/2}\sigma} + \alpha\frac{e^{-1/2}}{(2\pi)^{1/2}\sigma} + \nu\frac{1}{(2\pi)^{1/2}\sigma},
\end{aligned} \tag{E.13}$$

for all $x \in S_0$ and $y \in \mathbb{R}$. We now take $\alpha = \nu = \eta\sigma/3$ so that $a = \{2\log(\alpha^{-1})\}^{1/2} = \{2\log(3\sigma^{-1}\eta^{-1})\}^{1/2}$. The the right hand side of (E.13) is bounded by

$$\alpha/\sigma\left(2(2\pi)^{-1/2} + (2\pi)^{-1} + (2\pi)^{-1/2}e^{-1/2}\right) \leq \eta.$$

Therefore, as G''' varies, the collection of functions $(x, y) \mapsto f_x^{G'''}(y)$ forms an η-covering of $\mathcal{M}_K$ under the metric $\|\cdot\|_{\infty,S_0\times\mathbb{R}}$. It remains to bound the cardinality of this collection which equals $\binom{N_1}{\ell}N_2$. Thus

$$\log N(\eta, \mathcal{M}_K, \|\cdot\|_{\infty,S_0\times\mathbb{R}}) \leq \log\binom{N_1}{\ell} + \log N_2$$

By Stirling's formula,

$$\binom{N_1}{\ell} \leq \frac{N_1^\ell}{\ell!} \leq \left(\frac{N_1 e}{\ell}\right)^\ell.$$

By (E.12) and (E.6), we have $N_1 \leq \ell$ so that $\log\binom{N_1}{\ell} \leq \ell$. Also N_2 is the ν-covering number of Δ_ℓ under the L^1-metric which implies, by a well known result, that $\log N_2 \leq \ell\log(1 + 2/v)$. We thus get

$$\log N(\eta, \mathcal{M}_K, \|\cdot\|_{\infty,S_0\times\mathbb{R}}) \leq \ell(\log(1 + 2/\nu) + 1).$$

By $1/\nu = 3\sigma^{-1}\eta^{-1}$ and $\eta < e^{-1}\sigma^{-1}$, we get

$$\log N(\eta, \mathcal{M}_K, \|\cdot\|_{\infty,S_0\times\mathbb{R}}) \leq \ell(\log(1 + 2/\nu) + 1) \leq C\ell\log(\sigma^{-1}\eta^{-1}) \tag{E.14}$$

for a universal constant C. It also follows from (E.6) that

$$\ell = (2\lfloor 13.5a^2\rfloor\zeta + 1)^p N(a\sigma/\mathfrak{L}_{S_0}, K) \leq C_p\{\log(\sigma^{-1}\eta^{-1})\}^p\zeta^p N\left(\frac{\sigma}{\mathfrak{L}_{S_0}}\sqrt{2\log\frac{3}{\sigma\eta}}, K\right).$$

This, combined with (E.14), completes the proof of Theorem E.1. □