The Annals of Statistics
2010, Vol. 38, No. 3, 1733–1766
DOI: 10.1214/09-AOS765

APPROXIMATION OF CONDITIONAL DENSITIES BY SMOOTH MIXTURES OF REGRESSIONS

By Andriy Norets

Princeton University

This paper shows that large nonparametric classes of conditional multivariate densities can be approximated in the Kullback–Leibler distance by different specifications of finite mixtures of normal regressions in which normal means and variances and mixing probabilities can depend on variables in the conditioning set (covariates). These models are a special case of models known as "mixtures of experts" in statistics and computer science literature. Flexible specifications include models in which only mixing probabilities, modeled by multinomial logit, depend on the covariates and, in the univariate case, models in which only means of the mixed normals depend flexibly on the covariates. Modeling the variance of the mixed normals by flexible functions of the covariates can weaken restrictions on the class of the approximable densities. Obtained results can be generalized to mixtures of general location scale densities. Rates of convergence and easy to interpret bounds are also obtained for different model specifications. These approximation results can be useful for proving consistency of Bayesian and maximum likelihood density estimators based on these models. The results also have interesting implications for applied researchers.

1. Introduction. This paper explores approximation properties of finite smooth mixtures of normal regressions as flexible models for conditional densities. These models are a special case of mixtures of experts (ME) introduced by Jacobs et al. (1991). ME have become increasingly popular is statistical literature since they are very flexible, easy to interpret and reasonably easy to estimate. See, for example, papers by Jordan and Jacobs (1994) and Jordan and Xu (1995) who employ the expectation maximization (EM) estimation algorithm or papers by Peng, Jacobs and Tanner (1996), Wood, Jiang and Tanner (2002), Geweke and Keane (2007) and

Received August 2009; revised November 2009.

AMS 2000 subject classifications. Primary 62G07; secondary 41A30.

Key words and phrases. Finite mixtures of normal distributions, smoothly mixing regressions, mixtures of experts, Bayesian conditional density estimation.

Villani, Kohn and Giordani (2009) who use Markov chain Monte Carlo methods for estimation of ME in the Bayesian framework. This paper contributes to the literature that provides a theoretical explanation of the success of ME models in applications. In particular, I show that large classes of conditional densities can be approximated in the Kullback–Leibler (KL) distance by finite smooth mixtures of normal regressions. Approximation results are obtained in the KL distance for the following reason. If a data generating density is in the KL closure of a class of models then this density can be consistently estimated from data by these models under weak regularity conditions [see, e.g., Ghosh and Ramamoorthi (2003) for a textbook treatment of Schwarz's theorem on posterior consistency and Roeder and Wasserman (1997) for posterior consistency results for finite mixture of normals].

Consider a joint probability distribution F on a product space $Y \times X$, $Y \subset R^d$ and $X \subset R^{d_x}$. Assume the conditional distribution $F(y|x)$ has a density $f(y|x)$ with respect to the Lebesgue measure. The marginal density of x with respect to some generic measure is denoted by $f(x)$. A model $\mathcal{M}$ for the conditional density $f(y|x)$ is described by $p(y|x, \mathcal{M})$. The KL distance between $f(y|x)f(x)$ and $p(y|x, \mathcal{M})f(x)$ is defined by

$$d_{\mathrm{KL}}(F, \mathcal{M}) = \int \log \frac{f(y|x)}{p(y|x, \mathcal{M})} F(dy, dx).$$

This distance can also be interpreted as the expected KL distance between the conditional distributions. Either way, this is the distance useful for obtaining estimation consistency results. Also, convergence in the KL distance implies convergence in the total variation distance. Below, I consider several different specifications of mixture of normal regressions models, $p(y|x, \mathcal{M})$, and provide conditions on F under which $d_{\mathrm{KL}}(F, \mathcal{M})$ can be made arbitrarily small. I also derive rates of convergence and easy to interpret bounds for $d_{\mathrm{KL}}(F, \mathcal{M})$.

In general, a finite mixture of normal regressions model can be written as

$$p(y|x, \mathcal{M}) = \sum_{j=1}^{m} \alpha_j^m(x) \phi(y, \mu_j^m(x), \sigma_j^m(x)),$$

where mixing probabilities satisfy $\alpha_j^m(x) \in [0, 1]$ and $\sum_j \alpha_j^m(x) = 1$, and $\phi(y, \mu, \sigma)$ is a normal density with mean μ and standard deviation σ evaluated at y (if y is multidimensional then the variance–covariance matrix is diagonal $\sigma^2 I$). Most of the results obtained in the paper can be easily extended to models in which general location scale densities $\sigma^{-d}K((y-\mu)/\sigma)$ are mixed instead of the normal densities $\phi(y, \mu, \sigma)$. Models, in which the mixing weights depend on x, are referred in this paper as smooth mixtures. In practice, $\alpha_j^m(x)$'s are often modeled by a multinomial choice model, for example, multinomial logit [Peng, Jacobs and Tanner (1996)] or probit

[Geweke and Keane (2007)], or it might not depend on x. The mean $\mu_j^m(x)$ can be constant, linear or flexible, for example, polynomial, in x. An exponentiated polynomial or spline in x can be used for modeling the standard deviation $\sigma_j^m(x)$ [Villani, Kohn and Giordani (2009)].

To the best of my knowledge, previous literature on smooth mixtures of regressions (or experts) does not provide a theory on what specifications for α_j^m, μ_j^m and σ_j^m deliver a model that can approximate and consistently estimate large nonparametric classes of densities F. There are theoretical results on approximation of smooth functions and estimation of conditional expectations by ME [see Zeevi, Meir and Maiorov (1998) and Maiorov and Meir (1998)]. The only paper on approximation of conditional densities by ME seems to be Jiang and Tanner (1999) who develop approximation and estimation results for target densities from a single parameter exponential family, in which the parameter is a smooth function of covariates. A detailed comparison with results in Jiang and Tanner (1999) is presented in Section 6. In this paper, I do not restrict the functional form of $f(y|x)$ and use weak regularity conditions to describe a class of F that can be approximated. Conditions on approximable classes of $f(y|x)$ and $f(x)$ that are common for different model specifications include bounded support for $f(x)$, continuity of $f(y|x)$ in (y,x), finite expectation of a change of $\log f(y|x)$ in a neighborhood of y and existence of the second moments of y. The latter restriction can be weakened by adding densities with fat tails to the mixtures in addition to normal densities.

In Section 4, I show that considerable flexibility is already attained when α_j^m's are modeled by multinomial logit with linear indices in x, and (μ_j^m, σ_j^m) are independent of x. Results in Sections 3 and 4 suggest that using polynomials in the logit specification reduces the number of mixture components m required to achieve a specified approximation precision. As shown in Section 5, models for univariate response y in which the mixing probabilities and the variances of the mixed normals are independent of x, and the means are flexible, for example, polynomial in x, can approximate large classes of $f(y|x)$. Differences in quantiles of $f(y|x)$ from these classes have to be bounded above and below uniformly in x. These restrictions on $f(y|x)$ can be weakened if the variances of the mixed normals are modeled by flexible functions of x. Section 7 summarizes the findings.

2. Infeasible model.

In this section, I explicitly construct a smooth mixture of normals model that converges to a given F in the KL distance as m increases. This model is not feasible in the sense that it is not based on components employed in practice, for example, logit/probit mixing probabilities. However, the results for feasible models presented in the following sections follow from this one or are similar.

Let A_j^m, $j = 0, 1, \ldots, m$, be a partition of Y consisting of adjacent half-open half-closed hypercubes $A_1^m, \ldots, A_m^m$ with side length h_m and the rest of the space A_0^m. As m increases the fine part of the partition becomes finer, $h_m \to 0$. Also, it covers larger and larger part of Y: for any $y \in Y$ there exists M_0 such that

$$\forall m \geq M_0 \qquad C_{\delta_m}(y) \cap A_0^m = \varnothing, \tag{2.1}$$

where $C_{\delta_m}(y)$ is a hypercube with center y and side length $\delta_m \to 0$. It is always possible to construct such a partition. For example, if $Y = [0, \infty)$ let $A_0^m = [\log m, \infty)$, $A_j^m = [(j-1)\log m/m, j \log m/m)$ for $j \neq 0$, and $h_m = \log m/m$.

A candidate model $\mathcal{M}_0$ for approximating $f(y|x)$ is

$$p(y|x, \mathcal{M}_0) = \sum_{j=1}^{m} F(A_j^m|x)\phi(y, \mu_j^m, \sigma_m) + F(A_0^m|x)\phi(y, 0, \sigma_0), \tag{2.2}$$

where σ_0 is fixed, σ_m converges to zero as m increases and μ_j^m is the center of A_j^m. One can always construct a model $\mathcal{M}_0$ and a partition A_j^m so that

$$\delta_m \to 0, \qquad \sigma_m/\delta_m \to 0, \qquad \delta_m^{d-1} h_m/\sigma_m^d \to 0, \tag{2.3}$$

for example, in the example for $Y = [0, \infty)$ from the previous paragraph let $\sigma_m = h_m^{0.5}$ and $\delta_m = h_m^{0.25}$.

For a partition satisfying (2.1) and (2.3), let us introduce the following restrictions on F.

Assumption 2.1. 1. $f(y|x)$ is continuous in y a.s. F.
2. The second moments of y are finite.
3. For any (y, x) there exists a hypercube $C(r, y, x)$ with side length $r > 0$ and $y \in C(r, y, x)$ such that (i)

$$\int \log \frac{f(y|x)}{\inf_{z \in C(r,y,x)} f(z|x)} F(dy, dx) < \infty \tag{2.4}$$

and (ii) exists M_3 such that for any $m \geq M_3$, if $y \in A_0^m$ then $C(r, y, x) \cap A_0^m$ contains a hypercube $C_0(r, y, x)$ with side length $r/2$ and a vertex at y and if $y \in Y \setminus A_0^m$, then $C(r, y, x) \cap (Y \setminus A_0^m)$ contains a hypercube $C_1(r, y, x)$ with side length $r/2$ and a vertex at y.

Parameter σ_0 can always be chosen so that

$$1 > 2^{-(d+1)} > \phi(y, 0, \sigma_0)\lambda(C_0(r, y, x)), \tag{2.5}$$

where λ is the Lebesgue measure.

PROPOSITION 2.1. *If the model $p(y|x,\mathcal{M}_0)$ and the partition A_j^m are constructed so that (2.1), (2.2), (2.3) and (2.5) hold, and F satisfies Assumption 2.1, then $d_{\mathrm{KL}}(F,\mathcal{M}_0)\to 0$ as $m\to\infty$.*

The proposition is rigorously proved in the Appendix. Here, I briefly describe the intuition behind the argument and the role of the assumptions. Convergence in the KL distance is proved by the dominated convergence theorem (DCT). First, I establish point-wise convergence of the integrand, $\log f(y|x)/p(y|x,\mathcal{M}_0)$, to zero, and then I derive an integrable upper bound on the integrand for the DCT applicability. Nonnegativity of the KL distance is fruitfully exploited in the proof as it allows working only with upper bounds and ignoring the lower ones in convergence arguments.

The first term on the right-hand side of (2.2) (the sum from 1 to m) approximates the integral

$$\int \phi(y,\mu,\sigma_m)f(\mu|x)\,d\mu = \int f(y-\sigma_m z|x)\phi(z,0,1)\,dz, \tag{2.6}$$

when h_m is much smaller than σ_m, and the fine part of the partition is large. The integral on the right-hand side of (2.6) is obtained by the change of variables. For a small δ_m and z satisfying $\|\sigma_m z\|\le\delta_m$, $f(y-\sigma_m z|x)$ is close to $f(y|x)$ as $f(y|x)$ is assumed to be continuous in y. Therefore, when σ_m is much smaller than δ_m the right-hand side of (2.6) should be close to $f(y|x)$. Thus, this intuitive argument explains the role of conditions (2.3) and continuity of $f(y|x)$.

The second term on the right-hand side of (2.2) converges to zero. This term is not needed for point-wise convergence. It can be omitted when the support of $f(y|x)$ is bounded uniformly in x as in this case we can set $A_0^m=\varnothing$ and use the same variance σ_m^2 in all mixture components (there is no need to define σ_0). This term together with part 2 of Assumption 2.1 prevents tails of $p(y|x,\mathcal{M}_0)$ from becoming too thin relative to $f(y|x)$ in the unbounded support case (in the absence of this term the tails would be too thin as $\sigma_m\to 0$).

Parts 2 and 3 of Assumption 2.1 together guarantee existence of an integrable upper bound for the DCT applicability. An upper bound on $\log f(y|x)/p(y|x,\,\mathcal{M}_0)$ involves a lower bound on $p(y|x,\mathcal{M}_0)$. Both terms on the right-hand side in the definition of $p(y|x,\mathcal{M}_0)$ in (2.2) can be bounded below by an expression proportional to $\inf_{z\in C(r,y,x)} f(z|x)$. That is how condition (2.4) is deduced. The lower bound for the second term in (2.2) also includes $\phi(y,0,\sigma_0)$ and that is why finiteness of the second moments of y is assumed.

One interpretation of condition (2.4) [part 3(i) of Assumption 2.1] is that local relative changes in $f(y|x)$ due to changes in y should not be infinitely large on average. It seems difficult to think of an unconditional density, which is well behaved and positive everywhere, that would violate (2.4). This part of the assumption though can be violated by reasonable conditional densities as Example 2.1 below illustrates.

When $f(y|x)$ is positive everywhere, part 3(ii) of Assumption 2.1 is not needed. It always holds if $C(r,y,x)$ is a hypercube with center at y. Part 3(ii) becomes important when $f(y|x)$ can be equal to zero. In particular, the sets $C_0(r,y,x)$ and $C_1(r,y,x)$ in part 3(ii) of Assumption 2.1 are introduced to specify that $C(r,y,x)$ needs to be defined differently near the boundary of the support and in the tails if one wants to use condition (2.4) in its present form. This is illustrated in Figure 1.

The support of $f(\cdot|x)$ should include $C(r,y,x)$ a.s. F; otherwise, part 3(i) of Assumption 2.1 is not satisfied. Therefore, for $f(y|x)$ in Figure 1, it has to be the case that $C(r,y,x)=[y,y+r]$ at the boundary of the support (the intersection of the axes). Setting $C(r,y,x)=[y,y+r]$ near the boundary of the support makes the ratio $f(y|x)/\inf_{z\in C(r,y,x)} f(z|x)$ smallest possible (equal to one) and thus helps with condition (2.4). Parts of Y near the boundary of the support are covered by the fine part of the partition $A_1^m,\dots,A_m^m$ for all sufficiently large m, and part 3(ii) of Assumption 2.1 holds for $C_1(r,y,x)=[y,y+r/2]$. Using $C(r,y,x)=[y,y+r]$ for all y would not work. Since for any m one can find $y\in A_m^m$ such that $C(r,y,x)\cap Y\setminus A_0^m$ is arbitrary small, and part 3(ii) of Assumption 2.1 fails. Thus, for y that are arbitrary far from the boundary of the support, one has to use $C(r,y,x)=[y-r/2,y+r/2]$ eventually. Then, part 3(ii) of the assumption clearly holds for $C_1(r,y,x)=[y-r/2,y]$, $C_0(r,y,x)=[y,y+r/2]$ and any m.

Results in this section and similar results in the following sections can be generalized in several different ways. First, the derivation of the integrable upper bound in the proof of Proposition 2.1 suggests that the requirement of finite second moments of y can be weakened by adding a density with thicker than normal tails to the mixture of normals; for example, substitute $\phi(y,0,\sigma_0)$ in (2.2) with a Student t-density. Second, more

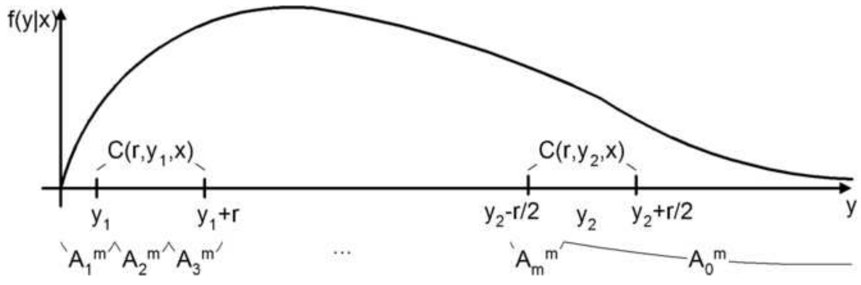

FIG. 1. *Construction of $C(r,y,x)$.*

general shapes of the support of F can be accommodated if instead of hypercubes $C(r,y,x)$, $C_0(r,y,x)$, and $C_1(r,y,x)$ in Assumption 2.1 different sets with positive Lebesgue measure are used. For example, if the support of $f(\cdot|x)$ is a triangle in R^2 then small triangles can be used instead of the squares $C(r,y,x)$, $C_0(r,y,x)$ and $C_1(r,y,x)$. Third, general location scale densities $\sigma^{-d}K((y-\mu)/\sigma)$ can be used in mixtures instead of normal densities. As long as analogs of Lemmas A.1, A.2 and A.3 (see the Appendix) are available for a particular type of densities, results in this and the following sections will hold for mixtures of these densities. Lemmas A.1 and A.3 hold for $\sigma^{-d}K((y-\mu)/\sigma)$ if $K(z)$ is bounded and nonincreasing in $|z|$ (proofs of the lemmas use only these facts about the normal distributions). The derivation of bounds in Lemma A.2 exploits normality; however, the qualitative results of the lemma hold as long as $\int_R K(z)\,dz = 1$ and $K(z)$ is positive in a neighborhood of zero. Thus, all the results in this paper that establish $d_{\mathrm{KL}}(F,\mathcal{M}) \to 0$ do not depend on the normality assumption; however, bounds and convergence rates for $d_{\mathrm{KL}}(F,\mathcal{M})$ derived below are specific to mixtures of normal densities, and they might be different for mixtures of other densities. All these generalizations seem to be straight forward and I do not pursue them in this paper to keep the arguments short and simple.

Examples below demonstrate that Assumption 2.1 is satisfied for a large class of densities. They also describe some situations in which the assumption fails.

Example 2.1. Exponential distribution, $f(y|x) = \gamma(x)\exp\{-\gamma(x)y\}$, $\gamma(x) > 0$. The density is continuous in y (part 1 of Assumption 2.1). Let $\int \gamma^{-2}\,dF < \infty$ so that the second moment of y is finite (part 2 of Assumption 2.1). Define the partition A_j^m and $C(r,y,x)$, $C_0(r,y,x)$ and $C_1(r,y,x)$ as shown in Figure 1, for example, for some $r > 0$ let $C(r,y,x) = [y, y+r]$ for $y \in [0,r]$ and $C(r,y,x) = [y - r/2, y + r/2]$ for $y \in (r,\infty)$. Thus, from the discussion of Figure 1 above it follows that part 3(ii) of Assumption 2.1 is satisfied. Because $\log f(y|x)/\inf_{z \in C(r,y,x)} f(z|x) \le r\gamma(x)$, part 3(i) of Assumption 2.1 holds as long as $\gamma(x)$ is integrable with respect to $f(x)$. If $\gamma(x)$ is not integrable, then part 3(i) of the assumption fails.

Example 2.2. A Student t-distribution, in which scale and location parameters are functions of x, $f(y|x) \propto [\nu + ((y-b(x))/c(x))^2]^{-(\nu+1)/2}$, $\nu > 2$ and $b(x)^2$, $c(x)^{-2}$ and $c(x)^2$ are integrable w.r.t. $f(x)$. The second moment of y is finite since

$$\int y^2\,dF = \int \left(c(x)^2 \left[\frac{y - b(x)}{c(x)} \right]^2 + 2b(x)y - b(x)^2 \right) dF$$

$$= \int \left(c(x)^2 \frac{\nu}{\nu - 2} + 2b(x)^2 - b(x)^2 \right) dF < \infty.$$

As I discuss above, for densities positive everywhere part 3(ii) of Assumption 2.1 always holds with $C(r,y,x) = [y - r/2, y + r/2]$. Part 3(i) of Assumption 2.1 is also satisfied because

$$\begin{aligned}
&\int \log \frac{f(y|x)}{\inf_{z \in C(r,y,x)} f(z|x)} F(dy, dx) \\
&\qquad = 2 \int_X \int_{b(x)}^{\infty} -\frac{\nu + 1}{2} \log \frac{\nu + ((y - b(x))/c(x))^2}{\nu + ((y + r - b(x))/c(x))^2} f(y|x)\, dy F(dx) \\
&\qquad \leq (\nu + 1) 2 \int_X \int_{b(x)}^{\infty} [\nu + ((y + r - b(x))/c(x))^2] f(y|x)\, dy F(dx) < \infty,
\end{aligned}$$

where the last inequality follows by the integrability of $((y - b(x))/c(x)$, its square and $c(x)^{-2}$.

Example 2.3. Suppose that conditional density $f(y|x)$ is continuous in y and bounded above and away from zero, $\infty > \overline{f} \geq f(y|x) \geq \underline{f} > 0$ for any $y \in Y = [a, b]$ and $x \in X$. Then we can set $A_0^m = \varnothing$. For $r \in (0, (b - a)/4)$, let $C(r, y, x) = [y, y + r]$ and $C_1(r, y, x) = [y, y + r/2]$ for $y \in [a, (a + b)/2]$ and $C(r, y, x) = [y - r, y]$ and $C_1(r, y, x) = [y - r/2, y]$ for $y \in ((a + b)/2, b]$. Clearly, part 3(ii) of Assumption 2.1 is satisfied. Because $f(y|x)/\inf_{z \in C(r,y,x)} f(z|x) \leq \overline{f}/\underline{f}$ part 3(i) of Assumption 2.1 also holds. The second moment of y is finite and thus all parts of Assumption 2.1 hold.

The boundedness away from zero condition can be replaced by a monotonicity condition at the boundary of the support. For example, let $f(y|x)$ be nondecreasing on $[a, a + 2r]$, nonincreasing on $[b - 2r, b]$ and bounded below by $\underline{f} > 0$ on $[a + r, b - r]$. In this case $f(y|x)/\inf_{z \in C(r,y,x)} f(z|x) \leq \max\{1, \overline{f}/\underline{f}\}$ for any $y \in [a, b]$. Thus, part 3(i) of Assumption 2.1 holds. The other parts of the assumption are not affected by this change.

Example 2.4. Consider a uniform distribution $f(y|x) = x^{-1} 1_{[0,x]}(y)$ and $f(x) > 0$ for any $x \in [1, \infty)$. A natural choice of the partition would be $A_0^m = [m h_m, \infty)$ and $A_j^m = [(j - 1) h_m, j h_m)$ for $j \in \{1, \ldots, m\}$. When $y = x$, the only reasonable choice of $C(r, y, x)$ is $C(r, y, x) = [y - r, y]$. For an arbitrary m and $y = x = m h_m + r/4$, $C(r, y, x)$ violates part 3(ii) of Assumption 2.1 since the only possible $C_0(r, y, x) = [y - r/2, y]$ is not included in A_0^m. For $f(x)$ with bounded support, this example would satisfy Assumption 2.1 since in this case we could set $A_0^m = \varnothing$.

This example illustrates that Assumption 2.1 rules out some cases in which the support of $f(\cdot|x)$ is increasing in x without a bound. In Section 5, I consider model specifications in which means and variances of the mixed normals can be flexible functions of x. Those specifications seem to be more promising for modeling densities $f(\cdot|x)$ with support increasing in x without a bound (see Example 5.2).

2.1. *Approximation error bounds.* The proof techniques of this section can also be used to derive explicit bounds on the approximation error. The bounds for positive everywhere and especially differentiable $f(y|x)$ are particularly informative. It is also easy to deduce an approximation rate from them. Thus, I present below the bounds and approximation rate for these special albeit important cases. Convergence rates and bounds for other special classes can be obtained in a similar way, for example, for densities bounded away from zero. However, rates and bounds for the general case seem to be difficult to calculate.

COROLLARY 2.1. *Part* (i). *Suppose the model $p(y|x,\mathcal{M}_0)$ and the partition A_j^m are constructed so that (2.1), (2.2), (2.3) and (2.5) hold. Suppose $f(y|x)$ is positive and continuous in y on $Y = R^d$ for all x, second moments of y are finite and (2.4) holds with $C(r,y,x) = C_r(y)$ taken to be a hypercube with center at y and radius r. Then, for all sufficiently large m,*

$$d_{\mathrm{KL}}(F,\mathcal{M}_0) \le \int \log \frac{f(y|x)}{\inf_{z\in C_{\delta_m}(y)} f(z|x)} F(dy,dx) \tag{2.7}$$

$$+\, 2\frac{3d^{3/2}\delta_m^{d-1}h_m}{(2\pi)^{d/2}\sigma_m^d} + 2\exp\left\{-\frac{(\delta_m/\sigma_m)^2}{8}\right\} \tag{2.8}$$

$$+ \int_{B_{\delta_m}(A_0^m)} \log \frac{f(y|x)}{\inf_{z\in C_r(y)} f(z|x)} F(dy,dx) \tag{2.9}$$

$$+ \int_{B_{\delta_m}(A_0^m)} \left[\frac{y'y}{2\sigma_0^2} - \log \frac{(r/2)^d}{(2\pi\sigma_0^2)^{d/2}}\right] F(dy,dx), \tag{2.10}$$

where $B_{\delta_m}(A_0^m) = \{(y,x) : C_{\delta_m}(y) \cap A_0^m \ne \varnothing\}$ and bounds in (2.7)–(2.10) converge to zero as $m \to \infty$.

Part (ii). *If $f(y|x)$ is continuously differentiable in y for all x and instead of (2.4) the following condition holds:*

$$\int \sup_{z\in C_r(y)} \left\| \frac{d \log f(z|x)}{dz} \right\| F(dy,dx) < \infty, \tag{2.11}$$

then for all sufficiently large m,

$$d_{\mathrm{KL}}(F,\mathcal{M}_0) \le \delta_m \cdot \frac{d^{1/2}}{2} \int \sup_{z\in C_{\delta_m}(y)} \left\| \frac{d \log f(z|x)}{dz} \right\| F(dy,dx) \tag{2.12}$$

$$+\, 2\frac{3d^{3/2}\delta_m^{d-1}h_m}{(2\pi)^{d/2}\sigma_m^d} + 2\exp\left\{-\frac{(\delta_m/\sigma_m)^2}{8}\right\} \tag{2.13}$$

$$+ \frac{rd^{1/2}}{2} \int_{B_{\delta_m}(A_0^m)} \sup_{z\in C_r(y)} \left\| \frac{d \log f(z|x)}{dz} \right\| F(dy,dx) \tag{2.14}$$

$$+\int_{B_{\delta_m}(A_0^m)}\left[\frac{y'y}{2\sigma_0^2}-\log\frac{(r/2)^d}{(2\pi\sigma_0^2)^{d/2}}\right]F(dy,dx), \tag{2.15}$$

and bounds in (2.12)–(2.15) converge to zero as $m\to\infty$.

Part (iii). *If, in addition to assumptions from part* (ii), *for some $q>2$ and some $i_1\in\{1,\dots,d\}$*

$$\int|y_i|^q F(dy)<\infty,\qquad i\in\{1,\dots,d\}, \tag{2.16}$$

and

$$\int|y_{i_1}|^{q-2}\sup_{z\in C_r(y)}\left\|\frac{d\log f(z|x)}{dz}\right\|F(dy,dx)<\infty, \tag{2.17}$$

then the approximation error bound can be written as

$$d_{\mathrm{KL}}(F,\mathcal{M}_0)\le c\cdot\left(\frac{1}{m}\right)^{1/(d\cdot[2+1/(q-2)+\varepsilon])}, \tag{2.18}$$

where $\varepsilon>0$ can be arbitrarily close to zero and c does not depend on m.

The corollary is proved in the Appendix. The bounds in part (i) of the corollary follow from the proof of Proposition 2.1. The bounds in part (ii) are derived from the bounds in part (i), and they are especially easy to interpret. The larger the "average" derivative of $\log f(y|x)$ is the smaller δ_m has to be to achieve a prespecified level for the right-hand side of (2.12). Constant h_m has to be much smaller than σ_m, and σ_m has to be much smaller than δ_m [condition (2.3)] so that (2.13) becomes sufficiently small. Size of (2.14) and (2.15) depends on how fast and by how much tails of $f(y|x)f(x)$ dominate $d\log f(y|x)/dy$, y^2, and a constant.

The approximation rate in part (iii) is derived from the bounds in part (ii). Expressions in (2.12) and (2.13) can be immediately converted in expressions in terms of m. To convert (2.14) and (2.15) in expressions in terms of m one seems to need slightly more than integrability of $\sup_{z\in C_r(y)}\|d\log f(z|x)/dz\|$ [condition (2.17)] and slightly more than finiteness of the second moments of y [condition (2.16)]. Under these conditions, (2.14) and (2.15) are bounded by $(h_m m^{1/d})^{-(q-2)}$ times a constant (see the corollary proof). An upper bound on $(h_m m^{1/d})^{-(q-2)}$, (2.12) and (2.13) gives the rate in (2.18). This upper bound has to be strictly larger than (2.18) with $\varepsilon=0$ as I show in the corollary proof. For distributions with exponentially declining tails, (2.14) and (2.15) can be decreasing exponentially in $h_m m^{1/d}$. In this case, one can set $q=\infty$ in (2.18) (see Example 5.3 below).

The dimension of y enters the approximation bounds exponentially. The dimension of x does not affect the bound and the approximation rate for the "infeasible" model because this model is constructed with the use of

$F(A_j^m|x)$'s, which are unknown functions of x. The following sections shed some light on the role of the dimension of x in approximating $f(y|x)$ by feasible models.

3. Flexible multinomial choice models for mixing probabilities.

This section gives conditions under which approximation results for "infeasible" model $\mathcal{M}_0$ also hold for a model with logit mixing probabilities that include polynomial terms in x. It also shows how to extend these results to multinomial probit and other models for mixing probabilities.

Assumption 3.1. X is compact and for partitions A_j^m, $j = 0, 1, \ldots, m$ satisfying (2.1), $F(A_j^m|x)$ is a continuous function of x on X and $F(A_j^m|x) > 0$ [the support of $f(\cdot|x)$ does not depend on x].

Under this assumption (by the Stone–Weierstrass theorem) for any sequence of $\varepsilon_m \to 0$, $\varepsilon_m > 0$ there exist finite order polynomials in x, $P_j^m(x)$ such that

$$|\log F(A_j^m|x) - P_j^m(x)| < \varepsilon_m \qquad \forall x \in X, j = 1, \ldots, m. \tag{3.1}$$

Let $p(y|x, \mathcal{M}_1)$ denote a model with σ_j^m and μ_j^m independent of x and logit mixing probabilities,

$$\begin{aligned}\alpha_j^m(x, \mathcal{M}_1) &= \frac{\exp\{P_j^m(x)\}}{\sum_{k=1}^m \exp\{P_k^m(x)\}}\\ &= \frac{F(A_j^m|x)\exp\{P_j^m(x) - \log F(A_j^m|x)\}}{\sum_{k=1}^m F(A_k^m|x)\exp\{P_k^m(x) - \log F(A_k^m|x)\}}.\end{aligned}$$

Condition (3.1) implies $\alpha_j^m(x, \mathcal{M}_1) \in (F(A_j^m|x)\exp\{-2\varepsilon_m\}, F(A_j^m|x)\exp\{2\varepsilon_m\})$. The following corollary immediately follows.

Corollary 3.1. *If Assumption 3.1 and the conditions of Proposition 2.1 hold then $d_{\mathrm{KL}}(F, \mathcal{M}_1)$ is bounded above and below by $d_{\mathrm{KL}}(F, \mathcal{M}_0) \pm 2\varepsilon_m$ and thus converges to zero.*

It seems possible to extend this corollary to other models for mixing probabilities, in particular, to a class of multinomial choice models in which mixing probabilities have the following representation:

$$\alpha_j^m(x) = \Pr[(e_0, \ldots, e_m) : v_j(x) + e_j \geq v_k(x) + e_k, k \in \{0, \ldots, m\}],$$

where $v_j(x)$ are flexible functions of x and e_k's are i.i.d. Multinomial logit and probit models fall into this category with polynomial $v_j(x)$ and extreme value and normal distributions for e_k's. The proof of Proposition 1 in

Hotz and Miller (1993) implies that if e_k are i.i.d. and have a density with respect to the Lebesgue measure, which is positive on R, then

$$(v_0(x),\dots,v_{m-1}(x)) = Q(\alpha_0^m(x),\dots,\alpha_{m-1}^m(x)),$$

where $v_m(x)$ is normalized to 0 and Q and Q^{-1} are differentiable mappings defined correspondingly on R^m, and the interior of the m-dimensional simplex. Flexible functional forms for $(v_0(x),\dots,v_{m-1}(x))$ can be used to approximate $Q(F(A_0^m|x),\dots,F(A_{m-1}^m|x))$. Then $(\alpha_0^m(x),\dots,\alpha_{m-1}^m(x)) = Q^{-1}(v_0(x),\dots, v_{m-1}(x))$ will approximate $(F(A_0^m|x),\dots,F(A_0^{m-1}|x)$. To get an analog of Corollary 3.1 one only needs to show that Q^{-1} transfers small additive approximation errors in $v_j(x)$ into multiplicative approximation errors for $\alpha_j^m(x)$, that are close to one. Since the mapping Q^{-1} is continuous this is the case as long as $F(A_j^m|x)$ are positive. Thus, it seems one does not need more than Assumption 3.1 to extend Corollary 3.1 to other models for mixing probabilities.

Of course, Corollary 3.1 can be formulated for any other method for approximating continuous functions in the sup norm on compacts, for example, for splines instead of the polynomials in the logit mixing probabilities.

The corollary implies that for F satisfying conditions of Corollary 2.1, bounds on the approximation error for model $\mathcal{M}_1$ are given by the bounds in the corollary for $\mathcal{M}_0$ plus ε_m. Results from the function approximation theory [see, e.g., Section 3.3 in Rust (1996) for a survey] suggest that to achieve a worst case approximation bound ε_m, computable approximations to Lipschitz continuous functions must involve the number of parameters proportional to $\varepsilon_m^{-d_x}$ ($\varepsilon_m^{-d_x/n}$ if the function has bounded derivatives up to order $n+1$). Thus, the number of parameters in the polynomials (or splines) $P_j^m(x)$ depends at best exponentially on the dimension of x.

It might be very difficult to estimate a model with high order polynomials in the logit mixing probabilities. The following section shows that it is not necessary to use high order polynomials in logit specification to attain flexibility. However, as I discuss at the end of the following section, polynomial terms might reduce the number of mixture components required to achieve a specified approximation precision.

4. Linear indices in logit.

In this section I explore an alternative approximation to $F(A_j^m|x)$ based on logit mixing probabilities that use only linear indices in x. The following assumption is a slightly stricter analog of Assumption 2.1.

ASSUMPTION 4.1. 1. $X = [0,1]^{d_x}$ (the arguments would go through for a bounded X).
2. $f(y|x)$ is continuous in (y,x) a.s. F.

3. The second moments of y are finite.
4. For any (y,x) there exists a hypercube $C(r,y,x)$ with side length $r>0$ and $y\in C(r,y,x)$ such that (i)

$$\int \log \frac{f(y|x)}{\inf_{z\in C(r,y,x),\|t-x\|\le r} f(z|t)} F(dy,dx) < \infty \tag{4.1}$$

and (ii) exists M such that for any $m\ge M$, if $y\in A_0^m$ then $C(r,y,x)\cap A_0^m$ contains a hypercube $C_0(r,y,x)$ with side length $r/2$ and a vertex at y and if $y\in Y\setminus A_0^m$, then $C(r,y,x)\cap(Y\setminus A_0^m)$ contains a hypercube $C_1(r,y,x)$ with side $r/2$ and a vertex at y.

Let B_i^m, $i=1,\dots,N(m)$ be equal size half-open half-closed hypercubes forming a partition of $X=[0,1]^{d_x}$. The partition becomes finer as m increases, $\lambda(B_i^m)=N(m)^{-1}\to 0$. Let x_i^m denote the center of B_i^m. Before looking at logit let us consider an "infeasible" model $\mathcal{M}_2$,

$$p(y|x,\mathcal{M}_2)=\sum_{i=1}^{N(m)}\left[\sum_{j=1}^{m}\alpha_{ij}^m(x,\mathcal{M}_2)\phi(y,\mu_j^m,\sigma_m)+\alpha_{i0}^m(x,\mathcal{M}_2)\phi(y,0,\sigma_0)\right],$$

where the mixing probabilities $\alpha_{ij}^m(x,\mathcal{M}_2)=1_{B_i^m}(x)F(A_j^m|x_i^m)$. As the partition of X becomes finer, model $\mathcal{M}_2$ approximates $\mathcal{M}_0$ because $F(A_j^m|x)\approx\sum_{i=1}^{N(m)}1_{B_i^m}(x)F(A_j^m|x_i^m)$ under continuity of $f(y|x)$ in x (part 2 of Assumption 4.1). Since, $\mathcal{M}_2$ is not interesting on its own I do not make this argument precise here. Instead I employ this idea to get approximation results for model $\mathcal{M}_3$ constructed similarly to $\mathcal{M}_2$ but with logit mixing probabilities,

$$\begin{aligned}\alpha_{ij}^m(x,\mathcal{M}_3)&=\frac{\exp\{\log F(A_j^m|x_i^m)-R_m(x_i^{m\prime}x_i^m-2x_i^{m\prime}x)\}}{\sum_{k,l}\exp\{\log F(A_k^m|x_l^m)-R_m(x_l^{m\prime}x_l^m-2x_l^{m\prime}x)\}}\\&=F(A_j^m|x_i^m)\frac{\exp\{-R_m(x_i^{m\prime}x_i^m-2x_i^{m\prime}x)\}}{\sum_l\exp\{-R_m(x_l^{m\prime}x_l^m-2x_l^{m\prime}x)\}}.\end{aligned}\tag{4.2}$$

In this expression, R_m is a positive diverging to infinity sequence that satisfies the following condition:

$$\exp\{-R_m s_m\}/s_m^{d_x/2}\to 0 \qquad \text{where } s_m=d_x\lambda(B_i^m)^{2/d_x}\to 0, \tag{4.3}$$

is the squared diagonal of B_i^m. This condition specifies that R_m should increase fast relative to how fine the partition of X becomes. It is always possible to define sequence R_m satisfying (4.3), for example, $R_m=s_m^{-2}$.

PROPOSITION 4.1. *If condition (4.3), Assumption 4.1, and conditions of Proposition 2.1 hold then* $d_{\mathrm{KL}}(F,\mathcal{M}_3)\to 0$ *as* $m\to\infty$.

The proposition is proved in the Appendix. The proof shows that the expression in (4.2) multiplying $F(A_j^m|x_i^m)$ behaves like $1_{B_i^m}(x)$ when R_m becomes large and then uses the same arguments as in the proof of Proposition 2.1. Attempts to develop similar results for mixing probabilities modeled by multinomial probit [see, e.g., Geweke and Keane (2007) for applications] were not successful. It would not be hard to make multinomial probit mixing probabilities behave like indicator functions. However, making them behave like an indicator times $F(A_j^m|x_i^m)$ as in (4.2) seems to be more difficult.

The bounds on the approximation error for $\mathcal{M}_3$ and $f(y|x)$ positive everywhere are similar to bounds for $\mathcal{M}_0$ obtained in Corollary 2.1. This is formalized in the following corollary.

Corollary 4.1. *Part* (i). *Suppose conditions of Proposition 4.1 hold, $f(y|x)$ is positive for any $y \in Y = R^d$ and any $x \in X$, $f(y|x)$ is continuously differentiable in (y,x), and instead of (4.1) the following condition holds:*

$$\int \sup_{y \in C_r(y), \|x-t\| \le r} \left\| \frac{d \log f(z|t)}{d(z,t)} \right\| F(dy,dx) < \infty; \tag{4.4}$$

then, for all sufficiently large m,

$$\begin{align} d_{\mathrm{KL}}(F, \mathcal{M}_3) \le \left(\delta_m \frac{d^{1/2}}{2} + s_m^{1/2} \right) & \tag{4.5} \\ \times \int \sup_{z \in C_{\delta_m}(y), \|x-t\| \le s_m^{1/2}} \left\| \frac{d \log f(z|t)}{d(z,t)} \right\| F(dy,dx) & \nonumber \\ + 2 \frac{3 d^{3/2} \delta_m^{d-1} h_m}{(2\pi)^{d/2} \sigma_m^d} + 2 \exp\left\{ -\frac{(\delta_m/\sigma_m)^2}{8} \right\} & \tag{4.6} \\ + \frac{r d^{1/2}}{2} \int_{B_{\delta_m}(A_0^m)} \sup_{z \in C_r(y), \|x-t\| \le r} \left\| \frac{d \log f(z|t)}{d(z,t)} \right\| F(dy,dx) & \tag{4.7} \\ + \int_{B_{\delta_m}(A_0^m)} \left[\frac{y'y}{2\sigma_0^2} - \log \frac{(r/2)^d}{(2\pi\sigma_0^2)^{d/2}} \right] F(dy,dx) & \tag{4.8} \\ + \log[1 - d_x^{d_x/2} \exp\{-R_m s_m\}/s_m^{d_x/2}], & \tag{4.9} \end{align}$$

and bounds in (4.5)–(4.9) converge to zero as $m \to \infty$.

Part (ii). *If, in addition to assumptions from part* (i), *for some $q > 2$ and some $i_1 \in \{1, \dots, d\}$,*

$$\int |y_i|^q F(dy) < \infty, \qquad i \in \{1, \dots, d\}, \tag{4.10}$$

and

$$\int |y_{i_1}|^{q-2} \sup_{z \in C_r(y), \|x-t\| \le r} \left\| \frac{d \log f(z|t)}{d(z,t)} \right\| F(dy, dx) < \infty, \tag{4.11}$$

then the approximation error bound can be written as

$$d_{\mathrm{KL}}(F, \mathcal{M}_3) \le \mathit{constant} \cdot [mN(m)]^{-1/(d_x + d \cdot [2 + 1/(q-2) + \varepsilon])}, \tag{4.12}$$

where $mN(m)+1$ is the number of mixture components in $\mathcal{M}_3$ and $\varepsilon > 0$ can be arbitrarily close to zero.

From the definition of models $\mathcal{M}_2$ and $\mathcal{M}_3$ and from the comparison of the convergence rates in (2.18) and (4.12), it is clear that using only linear indices in x in the mixing probabilities does not come without a cost. The number of mixing components in model $\mathcal{M}_3$ that approximates an infeasible model $\mathcal{M}_0$ is equal to $mN(m)+1$ while for model with polynomial terms in logit, $\mathcal{M}_1$, this number is $m+1$ (Corollary 3.1). The proof of Corollary 4.1 implies that the number of hypercubes in the partition of X, $N(m)$, increases exponentially with the dimensionality of X. Thus, the number of parameters in model $\mathcal{M}_3$ grows exponentially in the dimension of x (the exponential growth of the number of parameters in $\mathcal{M}_1$ is discussed at the end of the previous section). Overall, approximation results for $\mathcal{M}_1$ and $\mathcal{M}_3$ do not seem to suggest which model might perform better in practice; however, they seem to identify a tradeoff between the number of components in the mixture and the flexibility of models for the mixing probabilities.

5. Flexible means and variances. In this section, I show that a finite mixture of normal regressions models, in which mixing probabilities do not depend on x, can be quite flexible. However, the results also suggest that specifications in which mixing probabilities are flexible functions of x might perform better.

There is a large literature on finite mixture of regressions models. In early work, mixtures of two normal regressions were considered [see, e.g., Quandt and Ramsey (1978) and Kiefer (1978)]. Jones and McLachlan (1992) applied the EM algorithm for estimation of finite mixtures of normal regressions. Fitting of more general finite mixtures of generalized linear models has been considered in Jansen (1993) and Wedel and DeSarbo (1995) among others. Many more references can be found in a comprehensive book on finite mixture models by McLachlan and Peel (2000).

To the best of my knowledge, the literature on finite mixtures of regressions does not contain any approximation results for conditional densities. The closest analogs of the results I obtain can be found in the literature on finite mixtures of unconditional densities [see, e.g., Zeevi and Meir (1997) and references therein and Li and Barron (1999)]. Even for mixtures of unconditional densities approximation results for the KL distance, which is useful for establishing consistency of Bayesian or classical maximum likelihood estimators, seem to be scarce. Approximation results in the KL distance for convex combinations of densities in Zeevi and Meir (1997) and Li and Barron (1999) seem to apply to mixtures of truncated normals and to target densities that are compactly supported. Some of these results are very strong. For example, for target densities that are general mixtures of the densities mixed in the model, approximation error bounds obtained by Li and Barron (1999) are proportional to m^{-1}. If there are no covariates x, then the infeasible model from Section 2 is simply a finite mixture of multivariate normals. For an elaboration on this idea in the context of joint and conditional density estimation and for consistency results for a Bayesian estimator based on this model see Norets and Pelenis (2009). The convergence rates obtained for this model in Section 2.1 are slower than m^{-1}. However, the convergence rates are not directly comparable as the target densities in Li and Barron (1999) are different from those considered here.

Model $\mathcal{M}_4$ constructed in this section is very similar to model $\mathcal{M}_0$ except for one important difference. In $\mathcal{M}_4$, fine equal probability partitions of Y are used instead of fine equal length partitions in $\mathcal{M}_0$. As will be clear below, $\mathcal{M}_4$ defined in this way allows mixing probabilities to be independent of x. However, it requires the means of the mixed normals to be flexible functions of x. In this section, I assume that the response variable is univariate: $Y \subset R$ or $d = 1$ (all the results from previous sections were obtained for arbitrary d). If fine equal probability partitions can be well defined for distributions of multivariate random variables and if these partitions depend smoothly on covariates, then it might be possible to extend the results of this section to multivariate responses. I do not pursue this conjecture here.

Define model $\mathcal{M}_4$ as follows:

$$p(y|x, \mathcal{M}_4) = \sum_{j=1}^{m} \alpha_j^m \phi(y, \mu_j^m(x), \sigma_j^m(x)).$$

For a given x let $A_j^m(x)$, $j = 0, 1, \ldots, m$, be a partition of Y such that $\bigcup_{j=1}^m A_j^m(x)$ is a nondecreasing interval and

$$\begin{aligned} &F(A_j^m(x)|x) = p_m, \qquad j > 0, \\ &F(A_0^m(x)|x) = 1 - mp_m \quad \text{and} \quad mp_m \to 1, \end{aligned} \tag{5.1}$$

for some $p_m \in (0, m^{-1}]$ that does not depend on x. Define an upper bound on the length of an element of the fine part of the partition $h_m(x) \geq \max_{j>0} \lambda(A_j^m(x))$. The candidate mixing probabilities are given by $\alpha_j^m = F(A_j^m(x)|x)$ and $\mu_j^m(x) \in A_j^m(x)$. The standard deviations $\sigma_j^m(x) = \sigma_m(x)$

for $j>0$ and $\sigma_0^m(x)=\sigma_0(x)$ are treated as functions of x which is not essential but it weakens the restrictions on F (Corollaries 5.1 and 5.2 and Examples 5.1 and 5.2 below illustrate this point). Note that $\mathcal{M}_4$ is an infeasible model; in Corollary 5.2 below, I consider a feasible model $\mathcal{M}_5$ in which $\mu_j^m(x)$ are approximated by polynomials (see also Examples 5.1 and 5.2).

Suppose sequences $\delta_m(x)$, $\sigma_m(x)$, and $h_m(x)$ satisfy

$$\delta_m(x)\to 0, \qquad \frac{\sigma_m(x)}{\delta_m(x)}\to 0, \qquad \frac{h_m(x)}{\sigma_m(x)}\to 0. \tag{5.2}$$

Next, let us introduce the following restrictions on F.

Assumption 5.1. 1. Partitions $A_j^m(x)$ used in construction of $p(y|x,\mathcal{M}_4)$ satisfy (5.1), and (5.2) holds.
2. $f(y|x)$ is continuous in y a.s. F.
3. For any (y,x) there exists interval $C(r(x),y,x)$ with length $r(x)>0$ and $y\in C(r(x),y,x)$ such that (i)

$$\int \log\frac{f(y|x)}{\inf_{z\in C(r(x),y,x)} f(z|x)} F(dy,dx)<\infty \tag{5.3}$$

and (ii) exists M such that for any $m\geq M$, if $y\in A_0^m(x)$, then $C(r(x),y,x)\cap A_0^m(x)$ contains an interval $C_0(r(x),y,x)$ with an end at y and length $r(x)/2$, and if $y\in Y\setminus A_0^m(x)$, then $C(r(x),y,x)\cap(Y\setminus A_0^m(x))$ contains an interval $C_1(r(x),y,x)$ with an end at y and length $r(x)/2$.
4. $h_m(x)$, $\sigma_m(x)$, and $r(x)$ satisfy

$$\sup_x \frac{\sigma_m(x)}{r(x)}\to 0, \qquad \sup_x \frac{h_m(x)}{\sigma_m(x)}\to 0. \tag{5.4}$$

5. $\sigma_0(x)$ and $r(x)$ satisfy

$$1>1/4\geq \phi(y,0,\sigma_0(x))r(x)/2, \tag{5.5}$$

which holds, for example, when $\sigma_0(x)\geq 2(2\pi)^{-1/2}\cdot r(x)$.
6. $|\int \log[\phi(y,0,\sigma_0(x))r(x)/2]F(dy,dx)|<\infty$.

Proposition 5.1. *If Assumption 5.1 holds then $d_{\mathrm{KL}}(F,\mathcal{M}_4)\to 0$ as $m\to\infty$.*

The proposition is proved in the Appendix. The assumptions of the proposition and their role in the proof are similar to those discussed in detail in Section 2 for $\mathcal{M}_0$. The assumptions are satisfied by a large class of densities as illustrated by the following corollaries and examples. Approximation error bounds for $\mathcal{M}_4$ are presented below in Corollary 5.3.

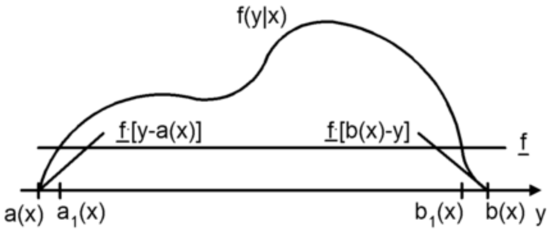

FIG. 2. *Approximation of densities with bounded support by $\mathcal{M}_4$.*

COROLLARY 5.1. *Assume:*

1. *$f(y|x)$ is continuous in y in the interior of the support of $f(y|x)$ for all $x \in X$.*
2. *There exists $\overline{f} < \infty$, such that $f(y|x) \le \overline{f}$ for all (y,x).*
3. *The support of $f(\cdot|x)$ is given by a finite interval $[a(x), b(x)]$, where $a(x)$ and $b(x)$ are square integrable. Also, for some $\underline{f} \in (0,1)$, a positive integer n, and $a(x) \le a_1(x) \le b_1(x) \le b(x), f(y|x) \ge \underline{f}$ on $[a_1(x), b_1(x)]$, $f(y|x) \ge \underline{f} \cdot [y - a(x)]^n$ on $(a(x), a_1(x))$, and $f(y|x) \ge \underline{f} \cdot [b(x) - y]^n$ on $(b_1(x), b(x))$. Figure 2 provides an illustration for $n = 1$.*
4. *There exists $r > 0$ such that $f(\cdot|x)$ is nondecreasing on $(a(x), a_1(x) + r/2)$ and nonincreasing on $(b_1(x) - r/2, b(x))$ for all $x \in X$.*

Then for $\mathcal{M}_4$ constructed so that $p_m = 1/m$, $A_0^m = \varnothing$, $\mu_j^m(x) \in A_j^m(x)$ and $\sigma_m(x) = p_m^{1/[4(n+1)]}$ and $\sigma_0(x) = 2(2\pi)^{-1/2} \cdot r$ are independent of x, $d_{\mathrm{KL}}(F, \mathcal{M}_4) \to 0$.

COROLLARY 5.2. *Assume conditions from Corollary 5.1, $F^{-1}(p|x)$ is continuous in x for all $p \in [0,1]$, X is compact. Then there exists a sequence of polynomials $P_j^m(x)$ such that $d_{\mathrm{KL}}(F, \mathcal{M}_5) \to 0$ where*

$$p(y|x, \mathcal{M}_5) = \sum_{j=1}^{m} p_m \phi(y, P_j^m(x), p_m^{1/8}).$$

PROOF. Let $\mu_j^m(x) = F^{-1}((j - 1/2)p_m|x)$. Note that $\mu_j^m(x) \in A_j^m(x) = [F^{-1}((j-1)p_m|x), F^{-1}(jp_m|x)]$ and

$$p_m/2 = \int_{\mu_j^m(x)}^{F^{-1}(jp_m|x)} f(y|x)\,dy \le (F^{-1}(jp_m|x) - \mu_j^m(x))\overline{f}.$$

Similarly, $p_m/2 \le (\mu_j^m(x) - F^{-1}((j-1)p_m|x))\overline{f}$. Thus, for $\varepsilon_m = p_m/(2\overline{f})$, $(\mu_j^m(x) - \varepsilon_m, \mu_j^m(x) + \varepsilon_m) \subset A_j^m(x)$. By the Stone–Weierstrass theorem there exist finite order polynomials in x, $P_j^m(x)$ such that $|P_j^m(x) - \mu_j^m(x)| < \varepsilon_m$.

Therefore, $P_j^m(x) \in A_j^m(x)$, which was the only requirement on the means of the mixture components in Corollary 5.1. $\square$

EXAMPLE 5.1. Exponential distribution, $f(y|x) = \gamma(x)\exp\{-\gamma(x)y\}$, $\gamma(x) \geq \underline{\gamma} > 0$, $\gamma(x)$ is continuous, $\int \gamma\, dF < \infty$ and the second moment of y is finite ($\int \gamma^{-2}\, dF < \infty$). The quantile function is given by $F^{-1}(p|x) = -\gamma(x)^{-1}\log(1-p)$. Let the partition be such that $A_0^m = [F^{-1}(mp_m|x), \infty)$. Since the exponential density is decreasing the largest interval in the fine part of the partition is given by $A_m^m = [F^{-1}((m-1)p_m|x), F^{-1}(mp_m|x))$. Therefore, $h_m(x) = h_m = \underline{\gamma}^{-1}\log(1 + p_m/(1 - p_m m))$. Choosing $p_m = (m - m^{0.5})/m^2$ guarantees that $h_m \to 0$. For $\sigma_m = h_m^{1/4}$, and $\delta_m(x) = h_m^{1/8}$, and $r(x) = 1$ conditions (5.1), (5.2) and (5.4) hold.

Next, let $C(1, y, x) = [y, y+1]$ if $y \in [0, 1/2]$, $C(1, y, x) = [y - 1/2, y + 1/2]$ if $y \in [1/2, \infty)$. Since

$$\inf_{z \in C(1,y,x)} f(z|x) \geq \gamma(x)\exp\{-\gamma(x)(y+1)\},$$

we have

$$1 \leq f(y|x) / \inf_{z \in C(1,y,x)} f(z|x) \leq \exp\{\gamma(x)\}.$$

Inequality (5.3) is satisfied since $\gamma(x)$ is assumed to be integrable. Finally, let $\sigma_0(x) = 2(2\pi)^{-1/2}$ so that equation (5.5) in Assumption 5.1 holds. Then,

$$\left| \int \log[\phi(y, 0, \sigma_0(x)) r(x)/2] F(dy, dx) \right| = \left| \int \left[-\log(4) - \frac{y^2\pi}{4} \right] F(dy, dx) \right| < \infty$$

since the second moment of y is assumed to be finite. Thus, condition 6 of Assumption 5.1 holds.

If X is compact the same argument as in the proof of Corollary 5.2 can be used to show that $\mu_j^m(x)$ can be polynomial in x [for fixed m there exists $\varepsilon_m > 0$ such that $\lambda(A_j^m(x)) > \varepsilon_m$ for all x and j].

It is possible to give sufficient conditions for approximation results when $\gamma(x)$ is not bounded away from zero, for example, let $r(x) = \gamma(x)^{-1}$, $h_m(x) = \gamma(x)^{-1}\log(1 + p_m/(1 - p_m m))$, etc. However, then σ_m and σ_0 would have to be functions of x [not necessarily flexible functions of x but functions that would have the same order as $\gamma(x)$]. Also, $\gamma(x)^{-1}$ is not continuous and the argument I use for justifying the use of polynomial $\mu_j^m(x)$ breaks down in this case.

EXAMPLE 5.2. Uniform distribution, $f(y|x) = b(x)^{-1} 1_{[0,b(x)]}(y)$, $b(x) > 0$ is continuous, $\int \log b\, dF < \infty$ and the second moment of y is finite ($\int b^2\, dF < \infty$). This example demonstrates that the support of $f(y|x)$ does not have to be (un)bounded uniformly in x as long as normal variances are modeled as flexible functions of x.

Let the partition be such that $A_0^m = \varnothing$ and $p_m = F(A_j^m|x) = m^{-1}$, $j > 0$. Note that $h_m(x) = b(x)/m$. For $\sigma_m(x) = b(x)p_m^{1/4}$, and $\delta_m(x) = b(x)p_m^{1/8}$, and $r(x) = b(x)$ conditions (5.1), (5.2) and (5.4) hold.

Next, let $C(r(x), y, x) = [0, b(x)]$. Note that $f(y|x)/\inf_{z\in C(r(x),y,x)} f(z|x) = 1$, and inequality (5.3) is satisfied. Finally, let $\sigma_0(x) = 2(2\pi)^{-1/2}b(x)$ so that inequality (5.5) in Assumption 5.1 holds. Then,

$$\left|\int \log[\phi(y, 0, \sigma_0(x))r(x)/2]F(dy, dx)\right| = |-\log(4) - \pi/(3\cdot 4)| < \infty$$

and condition 6 of Assumption 5.1 holds.

If X is compact and $b(x)$ is bounded away from zero then the same argument, as in the proof of Corollary 5.2, can be used to show that $\mu_j^m(x)$ can be polynomial in x [for fixed m there exists $\varepsilon_m > 0$ such that $\lambda(A_j^m(x)) > \varepsilon_m$ for all x and j].

Corollary 5.3. *Suppose conditions of Proposition 5.1 are satisfied for $h_m(x) = h_m$, $\sigma_m(x) = \sigma_m$, $\delta_m(x) = \delta_m$ and $r(x) = r$ that do not depend on x. Also, suppose conditions from parts* (i) *and* (ii) *of Corollary 2.1 hold. Then for all sufficiently large m,*

$$d_{\mathrm{KL}}(F, \mathcal{M}_4) \le \delta_m \cdot \frac{d^{1/2}}{2}\int \sup_{z\in C_{\delta_m}(y)}\left\|\frac{d\log f(z|x)}{dz}\right\| F(dy, dx) \tag{5.6}$$

$$+ 2\frac{3h_m}{(2\pi)^{1/2}\sigma_m} + 2\exp\left\{-\frac{(\delta_m/\sigma_m)^2}{8}\right\} \tag{5.7}$$

$$+ \frac{r}{2}\int_{B_{\delta_m}(A_0^m(x))} \sup_{z\in C_r(y)}\left\|\frac{d\log f(z|x)}{dz}\right\| F(dy, dx) \tag{5.8}$$

$$+ \int_{B_{\delta_m}(A_0^m(x))}\left[\frac{y'y}{2\sigma_0^2} - \log\frac{(r/2)}{(2\pi\sigma_0^2)^{1/2}}\right]F(dy, dx), \tag{5.9}$$

where $B_{\delta_m}(A_0^m(x)) = \{(y, x,) : C_{\delta_m}(y) \cap A_0^m(x) \neq \varnothing\}$and bounds in (5.6)–(5.9) converge to zero as $m \to \infty$.

Proof. The proof is identical to the proof of Corollary 2.1. □

The bounds for $\mathcal{M}_4$, (5.6)–(5.9), are almost the same as the bounds for $\mathcal{M}_0$, (2.12)–(2.15), obtained in Corollary 2.1, except for a difference between $B_{\delta_m}(A_0^m(x))$ in $\mathcal{M}_4$ and $B_{\delta_m}(A_0^m)$ in $\mathcal{M}_0$. For the same value of h_m, the length of the complement of $A_0^m(x)$ in $\mathcal{M}_4$ is bounded above by mh_m [$h_m = \max_{j>0}\lambda(A_j^m(x))$] which is the length of the complement of A_0^m in $\mathcal{M}_0$. Thus the bounds obtained for $\mathcal{M}_4$ are likely to be larger than the bounds obtained for $\mathcal{M}_0$. Compact and interpretable conditions sufficient for deriving an explicit approximation rate for $\mathcal{M}_4$ from (5.6)–(5.9) seem to be difficult to find. Instead, I show in the following example that not only bounds for $\mathcal{M}_0$ can be smaller but also that convergence for $\mathcal{M}_0$ can be slightly faster than for $\mathcal{M}_4$.

EXAMPLE 5.3. Laplace distribution, $f(y|x) = 0.5\gamma(x)\exp\{-\gamma(x)|y|\}$, $\gamma(x) \geq \underline{\gamma} > 0$, $\gamma(x)$ is continuous, $\int \gamma\, dF < \infty$ and the second moment of y is finite ($\int \gamma^{-2}\, dF < \infty$). Note that nondifferentiability of $f(y|x)$ at zero does not affect any of the theoretical results above.

First consider $\mathcal{M}_4$. Let $A_j^m(x) = [F^{-1}((1-p_m m)/2+(j-1)p_m|x), F^{-1}((1-p_m m)/2+jp_m|x))$. Note that $F^{-1}(p|x) = \log(2p)/\gamma(x)$ for $p < 0.5$ and $F^{-1}(p|x) = -\log(2(1-p))/\gamma(x)$ for $p \geq 0.5$. Then,

$$\begin{aligned} h_m &\geq F^{-1}((1-p_m m)/2 + p_m|x) - F^{-1}((1-p_m m)/2|x) \\ &= \frac{1}{\gamma(x)} \log\left(1 + \frac{2p_m}{1-p_m m}\right). \end{aligned} \tag{5.10}$$

Since $h_m \to 0$ and $mp_m \to 1$ we can write

$$p_m = \frac{1}{m + g(m)},$$

where $g(m)$ satisfies $g(m)/m \to 0$ and $g(m) \to \infty$. Note that

$$B_{\delta_m}(A_0^m(x)) \subset \left(-\infty, \frac{\log(1-p_m m)(1-\varepsilon_0)}{\gamma(x)}\right) \cup \left(-\frac{\log(1-p_m m)(1-\varepsilon_0)}{\gamma(x)}, \infty\right)$$

for any $\varepsilon_0 \in (0,1)$ and all sufficiently large m. A direct calculation shows that integrals in (5.8) and (5.9) can be bounded by

$$\text{constant} \cdot (1-p_m m)^{1-\varepsilon} \leq \text{constant} \cdot (g(m)/m)^{1-\varepsilon}$$

for any $\varepsilon \in (\varepsilon_0, 1)$ and all sufficiently large m. From (5.10) and the mean value theorem,

$$h_m \geq \text{constant} \cdot \gamma(x)^{-1} \cdot g(m)^{-1}.$$

Since the approximation error bounds increase in h_m, we should choose the smallest possible value for $h_m = \text{constant} \cdot \underline{\gamma}^{-1} \cdot g(m)^{-1}$. One can verify that the smallest upper bound for δ_m, h_m/σ_m, $\exp\{-(\delta_m/\sigma_m)^2/8\}$ and $(g(m)/m)^{1-\varepsilon}$ is inside the interval $(m^{-1/3}, m^{-1/[3+\varepsilon_1]}]$ for any $\varepsilon_1 > 0$ and all sufficiently large m. Thus,

$$d_{\mathrm{KL}}(F, \mathcal{M}_4) \leq \text{constant} \cdot \left(\frac{1}{m}\right)^{1/[3+\varepsilon_1]}.$$

Next, consider $\mathcal{M}_0$. Expressions (2.14) and (2.15) are exponentially decreasing in $h_m m$. Setting h_m to a power of m, one can show that

$$d_{\mathrm{KL}}(F, \mathcal{M}_0) \leq \text{constant} \cdot \left(\frac{1}{m}\right)^{1/[2+\varepsilon_2]},$$

for any $\varepsilon_2 > 0$ and all sufficiently large m. These results suggest that $\mathcal{M}_0$ converges to the target density faster than $\mathcal{M}_4$.

It might be unfair to compare approximation errors for $\mathcal{M}_0$ and $\mathcal{M}_4$. Although both models are "infeasible" and include m functions that need to be approximated by polynomials (or splines), the error from approximation by the polynomials enters the total approximation error in different ways. Nevertheless, the results obtained in this section do seem to suggest that models in which mixing probabilities depend on covariates might perform better in practice.

6. Comparison with Jiang and Tanner (1999).

Jiang and Tanner (1999) is the only work on approximation of conditional densities by ME that I am aware of. Jiang and Tanner (1999) develop approximation and estimation results for target densities of the form

$$\pi(y|x; h(\cdot)) = \exp(a(h(x))y + b(h(x)) + c(y)). \tag{6.1}$$

Functions a, b and c are assumed to be known, a and b are assumed to have nonzero derivatives and $h(x)$ is assumed to have uniformly bounded continuous second order derivatives. It seems that their results could still hold if a, b and c are known only up to some parameters (see their Remark 4). Jiang and Tanner (1999) show that $\pi(y|x; h(\cdot))$ can be approximated in the KL distance by ME of the form

$$\sum_{j=1}^{m} \alpha_j^m(x) \pi(y|x; h_j(\cdot)), \tag{6.2}$$

where $\pi(\cdot|\cdot; \cdot)$ is defined in (6.1), $h_j(x)$ is a linear function of x and the mixing probabilities $\alpha_j^m(x)$ can be modeled by logit (more general specifications for mixing weights are also allowed). The idea of their argument is to divide X into a fine partition B_j^m, approximate $1_{B_j^m}(x)$ by $\alpha_j^m(x)$ and approximate $h(x)$ by linear function $h_j(x)$ on B_j^m. Jiang and Tanner (1999) prove that for their target class of densities a bound on the approximation error is proportional to m^{-4/d_x}.

There are several important differences between the present work and Jiang and Tanner (1999). First, I consider multivariate responses, y, while Jiang and Tanner (1999) consider univariate responses. Most importantly, I do not assume that functional form of $f(y|x)$ is known, for example, known π, a, b and c. The components of the model I employ, for example, normal densities and logit mixing probabilities, are generally not related to the true density. As Examples 2.2 and 2.3 and Corollary 5.1 illustrate, many densities that are not from (6.1) are shown to be approximable by ME models. Examples 2.1 and 5.1 also show that some of the densities from class (6.1) satisfy sufficient conditions for approximation results I obtain. However, there might exist densities from (6.1) that violate these sufficient conditions. This would not be surprising since the "correct" functional forms are mixed in (6.2). For the same reason it is not surprising that the approximation rate obtained by Jiang and Tanner (1999), m^{-4/d_x}, differs from the ones obtained here, for example, $m^{-1/[d_x+2+1/(q-2)+\varepsilon]}$ for model $\mathcal{M}_3$ in Corollary 4.1.

Finally, responses in Jiang and Tanner (1999) class (6.1) can be discrete, for example, Poisson. To accommodate discrete responses in the framework of the present paper one could map the discrete values of response y into a partition of R and introduce a corresponding latent variable $y^* \sim p(y^*|x, \mathcal{M})$. For example, for binary $y \in \{0, 1\}$ let $y^* \in (-\infty, 0)$ if $y = 0$ and $y^* \in [0, \infty)$ if $y = 1$. Any discrete distribution can be represented by a continuously distributed latent variable in this fashion. This continuous distribution can be flexibly modeled by $p(y^*|x, \mathcal{M})$. Models with latent variables are easy to estimate in the Bayesian framework using MCMC methods [see, e.g., Tanner and Wong (1987) and Albert and Chib (1993)].

7. Discussion.

This paper shows that large classes of conditional densities can be approximated in the Kullback–Leibler distance by different specifications of finite smooth mixtures of normal densities or regressions. The theory can be generalized to smooth mixtures of location scale densities. These results have interesting implications for applied researchers.

First of all, smooth mixtures of densities or experts can be used as flexible models for estimation of multivariate conditional densities. It seems this issue has not been explored in the literature and it would be interesting to see how specifications studied in the paper work in these settings.

Second, smooth mixtures of simple components, for example, models in which mixing probabilities are modeled by multinomial logit linear in covariates and the means and variances do not depend on covariates, can be quite flexible. A simulation study in Villani, Kohn and Giordani (2009) suggests though that models with more complex components perform better in practice. This issue should be further explored in simulation studies.

Third, results in Section 4 suggest that making mixing probabilities more flexible, for example, by using polynomials in logit, might reduce the number of necessary mixture components. However, these models are more difficult to estimate.

Fourth, models in which mixing probabilities do not depend on covariates can be very flexible at least for univariate response variables. However, they seem to require a lot of mixture components and very flexible models for the means of the mixed normals. Also, approximation error bounds and convergences rates (Example 5.3) obtained in Section 5 suggest that models with flexible mixing probabilities might perform better in practice than models with flexible means of the mixed normals and constant mixing probabilities. Nevertheless, it would be interesting to see how these specifications perform in actual applications and simulation studies.

On the basis of a simulation study, Villani, Kohn and Giordani (2009) generally recommend using heteroscedastic experts (mixture components with variances that depend on covariates). The theory obtained here suggests that heteroscedastic experts might be necessary when differences in quantiles of $f(\cdot|x)$ are not uniformly bounded in x and, especially, when the support bounds of $f(\cdot|x)$ are increasing without a bound in x (see Examples 2.4 and 5.2). This suggestion is likely to remain useful when the differences in quantiles and/or support of $f(\cdot|x)$, although bounded, still change considerably with covariates.

Practical implications of the theoretical results obtained in the paper and summarized in this section are deduced under the assumption of no estimation and parameter uncertainty. Exploring the behavior of the estimation error in addition to the approximation error would result in a more complete understanding of the ME models. This issue is left for future work.

Overall, the paper provides a number of encouraging approximation results for (smooth) mixtures of densities or experts which might stimulate more theoretical and applied work in this area of research.

APPENDIX

Proof of Proposition 2.1. Since d_{KL} is always nonnegative,

$$0 \le \int \log \frac{f(y|x)}{p(y|x,\mathcal{M}_0)} F(dy,dx) \le \int \log \max\left\{1, \frac{f(y|x)}{p(y|x,\mathcal{M}_0)}\right\} F(dy,dx).$$

Thus, it suffices to show that the last integral in the inequality above converges to zero as m increases. The dominated convergence theorem (DCT) is used for that. First, I establish conditions for point-wise convergence of the integrand to zero a.s. F. Then, I present conditions for existence of an integrable upper bound on the integrand required by the DCT.

For fixed (y,x),

$$\begin{aligned} p(y|x,\mathcal{M}_0) &= \sum_{j=1}^{m} F(A_j^m|x)\phi(y,\mu_j^m,\sigma_m) + F(A_0^m|x)\phi(y,0,\sigma_0) \\ &\ge \inf_{z\in C_{\delta_m}(y)} f(z|x) \sum_{j\,:\,A_j^m\subset C_{\delta_m}(y)} \lambda(A_j^m)\phi(y,\mu_j^m,\sigma_m), \end{aligned} \tag{A.1}$$

where λ is the Lebesgue measure.

In Lemmas A.1 and A.2, I derive the following bounds for the Riemann sum in (A.1) (the Riemann sum is not far from the corresponding normal integral, and the integral is not far from 1):

$$\begin{aligned}
&\sum_{j:\, A_j^m \subset C_{\delta_m}(y)} \lambda(A_j^m)\phi(y,\mu_j^m,\sigma_m)\\
&\geq 1 - \frac{3d^{3/2}\delta_m^{d-1}h_m}{(2\pi)^{d/2}\sigma_m^d} - \frac{8(\sigma_m/\delta_m)}{(2\pi)^{1/2}}\exp\left\{-\frac{(\delta_m/\sigma_m)^2}{8}\right\} \qquad \text{(A.2)}\\
&\geq 1 - \frac{3d^{3/2}\delta_m^{d-1}h_m}{(2\pi)^{d/2}\sigma_m^d} - \exp\left\{-\frac{(\delta_m/\sigma_m)^2}{8}\right\},
\end{aligned}$$

where the last inequality holds for all sufficiently large m ($\delta_m/\sigma_m \to \infty$). Given $\varepsilon > 0$ there exists M_1 such that for $m \geq M_1$, expressions in (A.2) are bounded below by $(1-\varepsilon)$.

If $f(y|x)$ is continuous in y at (y,x) and $f(y|x) > 0$ there exists M_2 such that for $m \geq M_2$, $[f(y|x)/\inf_{z\in C_{\delta_m}(y)} f(z|x)] \leq (1+\varepsilon)$ since $\delta_m \to 0$. For any $m \geq \max\{M_0, M_1, M_2\}$,

$$\begin{aligned}
1 &\leq \max\left\{1, \frac{f(y|x)}{p(y|x,\mathcal{M}_0)}\right\}\\
&\leq \max\left\{1, \frac{f(y|x)}{\inf_{z\in C_{\delta_m}(y)} f(z|x)(1-\varepsilon)}\right\} \leq \frac{1+\varepsilon}{1-\varepsilon}.
\end{aligned}$$

Thus, $\log\max\{1, f(y|x)/p(y|x,\mathcal{M}_0)\} \to 0$ a.s. F as long as $f(y|x)$ is continuous in y a.s. F [$f(y|x)$ is always positive a.s. F].

Parts 2 and 3 of Assumption 2.1 are used for establishing an integrable upper bound for the DCT

$$\begin{aligned}
p(y|x,\mathcal{M}_0) &= \sum_{j=1}^m F(A_j^m|x)\phi(y,\mu_j^m,\sigma_m) + F(A_0^m|x)\phi(y,0,\sigma_0)\\
&\geq [1 - 1_{A_0^m}(y)] \qquad \text{(A.3)}\\
&\quad \times \inf_{z\in C_1(r,y,x)} f(z|x)\cdot \sum_{j:\, A_j^m\subset C_1(r,y,x)} \lambda(A_j^m)\phi(y,\mu_j^m,\sigma_m)\\
&\quad + 1_{A_0^m}(y)\cdot \inf_{z\in C_0(r,y,x)} f(z|x)\cdot\lambda(C_0(r,y,x))\phi(y,0,\sigma_0).
\end{aligned}$$

Lemmas A.1 and A.2 imply that the Riemann sum in (A.3) is bounded below by $2^{-d} - 2^{-(d+1)} = 2^{-(d+1)}$ for any m larger then some M_4. Inequalities (A.3) and (2.5) imply

$$\begin{aligned}
&\log \max\left\{1, \frac{f(y|x)}{p(y|x,\mathcal{M}_0)}\right\} \\
&\qquad \le \log \max\left\{1, \frac{f(y|x)}{\inf_{z\in C(r,y,x)} f(z|x)\cdot \phi(y,0,\sigma_0)\cdot (r/2)^d}\right\} \\
&\qquad = \log \frac{1}{\phi(y,0,\sigma_0)(r/2)^d} \max\left\{\phi(y,0,\sigma_0)(r/2)^d, \frac{f(y|x)}{\inf_{z\in C(r,y,x)} f(z|x)}\right\} \\
&\qquad \le -\log(\phi(y,0,\sigma_0)(r/2)^d) + \log \frac{f(y|x)}{\inf_{z\in C(r,y,x)} f(z|x)},
\end{aligned} \tag{A.4}$$

where inequality (A.4) follows by the first inequality in (2.5). The first expression in (A.4) is integrable by Assumption 2.1, part 2. The second expression in (A.4) is integrable by Assumption 2.1, part 3. Thus the proposition is proved. □

Proof of Corollary 2.1. The proof of the first part of the proposition is a simple implication of the argument in the proof of Proposition 2.1. Note that

$$\begin{aligned}
d_{\mathrm{KL}}(F,\mathcal{M}_0) &= \int_{Y\times X\setminus B_{\delta_m}(A_0^m)} \log \frac{f(y|x)}{p(y|x,\mathcal{M}_0)} F(dy,dx) \\
&\quad + \int_{B_{\delta_m}(A_0^m)} \log \frac{f(y|x)}{p(y|x,\mathcal{M}_0)} F(dy,dx).
\end{aligned} \tag{A.5}$$

For $(y,x)\in Y\times X\setminus B_{\delta_m}(A_0^m)$, inequalities (A.1) and (A.2) apply. Thus, the first integral in (A.5) is bounded by the sum of (2.7) and (2.8), where the bound in (2.8) is obtained by the mean value theorem for $-\log(1-x)$ and a small positive x,

$$\begin{aligned}
&-\log\left(1 - \frac{3d^{3/2}\delta_m^{d-1}h_m}{(2\pi)^{d/2}\sigma_m^d} - \exp\left\{-\frac{(\delta_m/\sigma_m)^2}{8}\right\}\right) \\
&\qquad \le 2\left(\frac{3d^{3/2}\delta_m^{d-1}h_m}{(2\pi)^{d/2}\sigma_m^d} + \exp\left\{-\frac{(\delta_m/\sigma_m)^2}{8}\right\}\right).
\end{aligned} \tag{A.6}$$

By inequality (A.3), the second integral in (A.5) is bounded by the sum of (2.9) and (2.10).

Expression (2.7) converges to zero by the DCT. The point-wise convergence follows by the assumed continuity and positivity of $f(y|x)$. An integrable upper bound is given by (2.4). Expression (2.7) converges to zero by (2.3). Expressions (2.9) and (2.10) converge to zero because $Y \times X \setminus$

$B_{\delta_m}(A_0^m) \nearrow Y \times X$ and the integrands are integrable by (2.4) and by the assumed finiteness of the second moment of y. Thus, the first part of the proposition is proved.

The second part of the proposition [bounds for differentiable $f(y|x)$] follows from the first part since

$$\left|\log \frac{f(y|x)}{\inf_{z\in C_r(y)} f(z|x)}\right| \le \sup_{z\in C_r(y)} \left\|\frac{d\log f(z|x)}{dz}\right\| \frac{d^{1/2}r}{2},$$

which is implied by the multivariate mean value theorem: for any (z_1, z_2)

$$|\log f(z_1|x) - \log f(z_2|x)| \le \|f'(cz_1 + (1-c)z_2)\| \|z_1 - z_2\|$$

for some $c \in [0,1]$. Convergence of the bounds to zero is obtained in the same way as in the first part of the proposition.

To obtain the third part let us suppose that the fine part of the partition $\{A_j^m, 1 \le j \le m\}$ is centered at 0. If $(y,x) \in B_{\delta_m}(A_0^m)$, then $|y_i| \ge h_m m^{1/d}/2 - \delta_m > h_m m^{1/d}/3$ for $i \in \{1,\ldots,d\}$ and all sufficiently large m and

$$\begin{aligned}
&\int_{B_{\delta_m}(A_0^m)} y_i^2 F(dy,dx) \\
&\qquad \le \int_{\{(y,x)\colon |y_i|>h_m m^{1/d}/3, \forall i\}} y_i^2 F(dy,dx) \\
&\qquad \le (h_m m^{1/d}/3)^{-(q-2)} \\
&\qquad\quad \times \int_{\{(y,x)\colon |y_i|>h_m m^{1/d}/3, \forall i\}} (h_m m^{1/d}/3)^{q-2} y_i^2 F(dy,dx) \\
&\qquad \le (h_m m^{1/d}/3)^{-(q-2)} \int_{Y\times X} y_i^q F(dy,dx).
\end{aligned} \tag{A.7}$$

Similarly,

$$\begin{aligned}
&\int_{B_{\delta_m}(A_0^m)} \sup_{z\in C_r(y)} \left\|\frac{d\log f(z|x)}{dz}\right\| F(dy,dx) \\
&\qquad \le \int_{\{(y,x)\colon |y_i|>h_m m^{1/d}/3, \forall i\}} \sup_{z\in C_r(y)} \left\|\frac{d\log f(z|x)}{dz}\right\| F(dy,dx) \\
&\qquad \le \biggl(\int_{\{(y,x)\colon |y_i|>h_m m^{1/d}/3, \forall i\}} (h_m m^{1/d}/3)^{q-2} \\
&\qquad\qquad\qquad \times \sup_{z\in C_r(y)} \left\|\frac{d\log f(z|x)}{dz}\right\| F(dy,dx)\biggr)
\end{aligned} \tag{A.8}$$

$$\begin{aligned}&\times\left((h_m m^{1/d}/3)^{q-2}\right)^{-1}\\ &\le (h_m m^{1/d}/3)^{-(q-2)}\int_{Y\times X} y_{i_1}^{q-2}\sup_{z\in C_r(y)}\left\|\frac{d\log f(z|x)}{dz}\right\|F(dy,dx).\end{aligned}$$

Since integrals in (A.7) and (A.8) are finite by assumption, (2.14) and (2.15) can be bounded above by an expression proportional to $(h_m m^{1/d})^{-(q-2)}$. Thus, the sum of (2.12)–(2.15) is bounded by

$$\begin{aligned}&c_1\cdot\delta_m + c_2\cdot\exp\{-(\delta_m/\sigma_m)^2/8\} + c_3\cdot\delta_m^{d-1}h_m/\sigma_m^d\\ &\quad + c_4\cdot 1/(h_m m^{1/d})^{q-2},\end{aligned}\tag{A.9}$$

where constants c_1, c_2, c_3 and c_4 do not depend on m. Let b_m be the smallest number satisfying $b_m\ge\delta_m$, $b_m\ge\delta_m^{d-1}h_m/\sigma_m^d$, $b_m\ge 1/(h_m m^{1/d})^{q-2}$ and $b_m\ge\exp\{-(\delta_m/\sigma_m)^2/8\}$. The first three of these inequalities imply

$$b_m\ge\{[(\delta_m/\sigma_m)^d]/m^{1/d}\}^{1/[2+1/(q-2)]}.$$

It implies that for all sequences δ_m, σ_m and h_m allowed by the corollary,

$$b_m > \left(\frac{1}{m}\right)^{1/(d\cdot[2+1/(q-2)])}.$$

One can verify that

$$b_m\le\left(\frac{(4\log m/d)^{d/2}}{m^{1/d}}\right)^{1/[2+1/(q-2)]}\le\left(\frac{1}{m}\right)^{1/(d\cdot[2+1/(q-2)+\varepsilon])},\tag{A.10}$$

when δ_m equal to the first bound in (A.10), $(\delta_m/\sigma_m)^2 = 4\log m/d$ and $h_m = \delta_m^2/(\delta_m/\sigma_m)^d$. □

Proof of Proposition 4.1. Define $I_1^m(x,s_m)=\{i\colon\|x_i^m - x\|^2 < s_m\}$ and $I_2^m(x,s_m)=\{i\colon\|x_i^m - x\|^2 > 2s_m\}$. For $i\in I_1^m(x,s_m)$,

$$[-R_m(x_i^{m\prime}x_i^m - 2x_i^{m\prime}x)] > [-R_m(s_m - x'x)]\tag{A.11}$$

and for $i\in I_2^m(x,s_m)$,

$$[-R_m(x_i^{m\prime}x_i^m - 2x_i^{m\prime}x)] < [-R_m(2s_m - x'x)].\tag{A.12}$$

Note that

$$\begin{aligned}&\frac{\sum_{i\in I_1^m(x,s_m)}\exp\{-R_m(x_i^{m\prime}x_i^m - 2x_i^{m\prime}x)\}}{\sum_l\exp\{-R_m(x_l^{m\prime}x_l^m - 2x_l^{m\prime}x)\}}\\ &\quad\ge 1-\frac{\sum_{i\in I_2^m(x,s_m)}\exp\{-R_m(x_i^{m\prime}x_i^m - 2x_i^{m\prime}x)\}}{\sum_{i\in I_1^m(x,s_m)}\exp\{-R_m(x_i^{m\prime}x_i^m - 2x_i^{m\prime}x)\}}\\ &\quad\ge 1-\frac{\operatorname{card}(I_2^m(x,s_m))}{\operatorname{card}(I_1^m(x,s_m))}\exp\{-R_m s_m\}\ge 1 - d_x^{d_x/2}\frac{\exp\{-R_m s_m\}}{s_m^{d_x/2}},\end{aligned}\tag{A.13}$$

where the second inequality follows from (A.11) and (A.12). The last inequality follows from the following bounds on the number of elements in $I_1^m(x,s_m)$ and $I_2^m(x,s_m)$: $\operatorname{card}(I_1^m(x,s_m)) \geq 1$ [s_m is chosen in (4.3) so that any ball in X with radius $s_m^{1/2}$ has to contain at least one x_i^m] and

$$\operatorname{card}(I_2^m(x,s_m)) \leq N(m) = d_x^{d_x/2} s_m^{-d_x/2}.$$

For $i \in I_1^m(x,s_m)$ and $A_j^m \subset C_{\delta_m}(y)$,

$$F(A_j^m | x_i^m) \geq \lambda(A_j^m) \inf_{z \in C_{\delta_m}(y), \|t-x\|^2 \leq s_m} f(z|t). \tag{A.14}$$

Inequalities (A.13), (A.14) and (A.2) imply that $p(y|x,\mathcal{M}_3)$ exceeds

$$\sum_{j\,:\, A_j^m \subset C_{\delta_m}(y)} \sum_{i \in I_1^m(x,s_m)} F(A_j^m|x_i^m) \frac{\exp\{-R_m(x_i^{m\prime}x_i^m - 2x_i^{m\prime}x)\}}{\sum_l \exp\{-R_m(x_l^{m\prime}x_l^m - 2x_l^{m\prime}x)\}} \phi(y,\mu_j^m,\sigma_m)$$

$$\begin{aligned}
&\geq \inf_{z \in C_{\delta_m}(y), \|t-x\|^2 \leq s_m} f(z|t) \cdot \Bigg[1 - \frac{3d^{3/2}\delta_m^{d-1}h_m}{(2\pi)^{d/2}\sigma_m^d} \\
&\qquad\qquad - \frac{8d\sigma_m}{(2\pi)^{1/2}\delta_m} \exp\left\{-\frac{(\delta_m/\sigma_m)^2}{8}\right\}\Bigg] \\
&\qquad\qquad \times \left[1 - d_x^{d_x/2} \frac{\exp\{-R_m s_m\}}{s_m^{d_x/2}}\right].
\end{aligned}$$

The expression on the last line of this inequality converges to 1 by (4.3). The rest of the proof is exactly the same as the proof of Proposition 2.1. □

Proof of Corollary 4.1. The proof of part (i) is identical to the proof of Corollary 2.1 part (ii).

The proof of part (ii) is also similar to the proof of Corollary 2.1 part (iii). Just set $s_m^{1/2} = \delta_m$ and note that (4.9) can be made arbitrarily smaller than the other parts of the bound by an appropriate choice of R_m. Thus, the bound is the same as in (2.18), we just need to express m in terms of the number of mixture components in $\mathcal{M}_3$, $mN(m)$. From the definition of $N(m)$ and s_m, $N(m) = \lambda(B_i^m)^{-1} = d_x^{d_x/2} s_m^{-d_x/2}$. Since we set $s_m^{1/2} = \delta_m$ and $\delta_m = m^{-1/(d\cdot[2+1/(q-2)])}$ in the proof of Corollary 2.1,

$$mN(m) = d_x^{d_x/2} m^{1+d_x/(d\cdot[2+1/(q-2)])}.$$

From this equation, one can express m as a function of $mN(m)$ and plug it in (2.18) to obtain (4.12). □

Proof of Proposition 5.1. First, consider point-wise convergence a.s. F. For fixed (y,x) and an interval $C_{\delta_m(x)}(y)$ with center y and length

$\delta_m(x) > 0$,

$$
\begin{aligned}
p(y|x,\mathcal{M}_4) &= \sum_{j=1}^{m} F(A_j^m(x)|x)\phi(y,\mu_j^m(x),\sigma_m(x)) \\
&\quad + F(A_0^m(x)|x)\phi(y,0,\sigma_0(x)) \\
&\geq \inf_{z\in C_{\delta_m(x)}(y)} f(z|x) \sum_{j=1}^{m} \lambda(A_j^m(x)\cap C_{\delta_m(x)}(y)) \\
&\qquad \times \phi(y,\mu_j^m(x),\sigma_m(x)) \\
&\geq \inf_{z\in C_{\delta_m(x)}(y)} f(z|x)\biggl(1 - \frac{6h_m(x)}{(2\pi)^{1/2}\sigma_m(x)} \\
&\qquad - \frac{16\sigma_m(x)}{(2\pi)^{1/2}\delta_m(x)} \exp\biggl\{-\frac{(\delta_m/\sigma_m)^2}{8}\biggr\}\biggr),
\end{aligned}
\tag{A.15}
$$

where the last inequality follows from Lemma A.3 [if $\delta_m(x)\to 0$ and $mp_m \to 1$ then for any (y,x) there exists M such that $\forall m \geq M$, $C_{\delta_m(x)}(y)\cap A_0^m(x) = \varnothing$ and the lemma applies]. Convergence of the bound in (A.15) to $f(y|x)$ a.s. F is implied by a.s. positivity and continuity in y of $f(y|x)$ and conditions in (5.2). The rest of the argument establishing point-wise convergence is the same as for $\mathcal{M}_0$ [details are below (2.3)].

Next, let us derive an integrable upper bound for the DCT,

$$
\begin{aligned}
p(y|x,\mathcal{M}_4) &= \sum_{j=1}^{m} F(A_j^m(x)|x)\phi(y,\mu_j^m(x),\sigma_m(x)) \\
&\quad + F(A_0^m(x)|x)\phi(y,0,\sigma_0(x)) \\
&\geq [1 - 1_{A_0^m(x)}(y)] \\
&\quad \times \inf_{z\in C_1(r(x),y,x)} f(z|x) \\
&\qquad \times \sum_{j:\, A_j^m(x)\subset C_1(r(x),y,x)} \lambda(A_j^m(x)) \\
&\qquad\qquad \times \phi(y,\mu_j^m(x),\sigma_m(x)) \\
&\quad + 1_{A_0^m(x)}(y)\cdot \inf_{z\in C_0(r(x),y,x)} f(z|x)\cdot \lambda(C_0(r(x),y,x)) \\
&\qquad \times \phi(y,0,\sigma_0(x)).
\end{aligned}
\tag{A.16}
$$

Lemma A.3 and condition (5.4) imply that the sum in (A.16) is bounded below by $1/2-1/4=1/4$ for all sufficiently large m. Equation (5.5) implies

$$
\begin{aligned}
&\log\max\left\{1,\frac{f(y|x)}{p(y|x,\mathcal{M}_4)}\right\}\\
&\quad\le\log\max\left\{1,\frac{f(y|x)\cdot(r(x)/2)^{-1}}{\inf_{z\in C(r(x),y,x)}f(z|x)\cdot\phi(y,0,\sigma_0(x))}\right\}\\
&\quad\le\log\frac{1}{\phi(y,0,\sigma_0(x))(r(x)/2)}\max\biggl\{\phi(y,0,\sigma_0(x))(r(x)/2),\\
&\qquad\qquad\qquad\qquad\qquad\frac{f(y|x)}{\inf_{z\in C(r(x),y,x)}f(z|x)}\biggr\} \qquad\qquad \text{(A.17)}\\
&\quad\le-\log[\phi(y,0,\sigma_0(x))r(x)/2]+\log\frac{f(y|x)}{\inf_{z\in C(r(x),y,x)}f(z|x)}.
\end{aligned}
$$

Inequality (A.17) follows by (5.5). The first expression in (A.17) is integrable by Assumption 5.1, part 6. The second expression in (A.17) is integrable by Assumption 5.1, part 3. This completes the proof of the proposition. □

Proof of Corollary 5.1. It suffices to show that Assumption 5.1 is satisfied. First, let us obtain a suitable h_m. Note that

$$
p_m\ge\int_{A_j^m(x)\cap[a_1(x),b_1(x)]}f(y|x)\,dy\ge\lambda(A_j^m(x)\cap[a_1(x),b_1(x)])\underline{f}. \tag{A.18}
$$

Also,

$$
\begin{aligned}
p_m&\ge\int_{A_j^m(x)\cap[a(x),a_1(x)]}f(y|x)\,dy\\
&\ge\int_{A_j^m(x)\cap[a(x),a_1(x)]}\underline{f}\cdot[y-a(x)]^n\,dy \qquad\qquad \text{(A.19)}\\
&\ge(n+1)^{-1}\lambda(A_j^m(x)\cap[a(x),a_1(x)])^{n+1}\underline{f}
\end{aligned}
$$

and similarly $p_m\ge(n+1)^{-1}\lambda(A_j^m(x)\cap[b_1(x),b(x)])^{n+1}\underline{f}$. Combining this inequality with (A.18) and (A.19) we get for all x and j,

$$
\begin{aligned}
\lambda(A_j^m(x))&\le\frac{p_m}{\underline{f}}+\frac{2\cdot(n+1)^{1/(n+1)}\cdot p_m^{1/(n+1)}}{\underline{f}^{1/(n+1)}}\\
&\le\frac{7p_m^{1/(n+1)}}{\underline{f}}=h_m.
\end{aligned}
$$

For $\sigma_m(x) = p_m^{1/4(n+1)}$ and $\delta_m(x) = p_m^{1/8(n+1)}$ conditions (5.1), (5.2) and (5.4) hold.

Next, let $C(r, y, x) = [y, y + r]$ if $y \in (a(x), a_1(x) + r/2)$, $C(r, y, x) = [y - r/2, y + r/2]$ if $y \in [a_1(x) + r/2, b_1(x) - r/2]$ and $C(r, y, x) = [y - r/2, y]$ if $y \in (b_1(x) - r/2, b(x))$. By condition 4 of the corollary $\inf_{z \in C(r(x),y,x)} f(z|x) = f(y|x)$ for $y \notin [a_1(x) + r/2, b_1(x) - r/2]$. For $y \in [a_1(x) + r/2, b_1(x) - r/2]$, $\inf_{z \in C(r(x),y,x)} f(z|x) \geq \underline{f}$ and

$$\int \log \frac{f(y|x)}{\inf_{z \in C(r(x),y,x)} f(z|x)} F(dy, dx) \leq \log(\overline{f}/\underline{f}) < \infty.$$

Condition 2 and (5.5) in Assumption 5.1 are assumed in the corollary. Since $a(x)$ and $b(x)$ are assumed to be square integrable, the second moment of y is finite, and condition 6 of Assumption 5.1 holds. □

Lemma A.1. *Define a hypercube $C_\delta(y) = \{\mu \in R^d : y_i \leq \mu_i \leq y_i + \delta, i = 1, \ldots, d\}$. Let $A_1, \ldots, A_m$ be adjacent hypercubes with centers μ_j and side length h such that $C_\delta(y) \subset \bigcup_{j=1}^m A_j$ and $\delta > 3d^{1/2}h$. Define $J = \{j : A_j \subset C_\delta(y)\}$. Then*

$$\sum_{j \in J} \lambda(A_j)\phi(y; \mu_j, \sigma) \geq \int_{C_\delta(y)} \phi(\mu; y; \sigma)\, d\mu - \frac{3d^{3/2}\delta^{d-1}h}{(2\pi)^{d/2}\sigma^d}.$$

By symmetry, this result holds for any hypercube with vertex at y and side length δ. This implies that for hypercube $D_\delta(y) = \{x : y_i - \delta/2 \leq x_i \leq y_i + \delta/2, i = 1, \ldots, d\}$,

$$\sum_{j:\, A_j \subset D_\delta(y)} \lambda(A_j)\phi(y; \mu_j, \sigma) \geq \int_{D_\delta(y)} \phi(\mu; y; \sigma)\, d\mu - 2^d \frac{3d^{3/2}(\delta/2)^{d-1}h}{(2\pi)^{d/2}\sigma^d}$$

as long as $D_\delta(y) \subset \bigcup_{j=1}^m A_j$ and $\delta > 6d^{1/2}h$.

Proof. For $j \in J$ let $B_j = \{x : \mu_{ji} \leq x_i \leq \mu_{ji} + h, i = 1, \ldots, d\}$ be a shifted and rotated version of A_j. Note that $\mu_j = \arg\max_{\mu \in B_j} \phi(\mu; y; \sigma)$, and therefore

$$\begin{aligned}
&\sum_{j \in J} \lambda(A_j)\phi(y; \mu_j, \sigma) \\
&\qquad = \sum_{j \in J} \lambda(B_j)\phi(y; \mu_j, \sigma) \geq \int_{\bigcup_{j \in J} B_j} \phi(\mu; y; \sigma)\, d\mu \\
&\qquad \geq \int_{C_\delta(y)} \phi(\mu; y; \sigma)\, d\mu - \int_{C_\delta(y) \setminus \bigcup_{j \in J} B_j} \phi(\mu; y; \sigma)\, d\mu.
\end{aligned}$$

Since $\{x : \min_J \mu_{ji} \le x_i \le \max_J \mu_{ji}, i = 1, \ldots, d\} \subset C_\delta(y) \cap [\bigcup_J B_j]$ and $\max_{j\in J} \mu_{ji} - \min_{j\in J} \mu_{ji} \ge \delta - 3d^{1/2}h$, we get $\lambda(C_\delta(y) \cap [\bigcup_J B_j]) \ge (\delta - 3d^{1/2}h)^d$ and

$$\begin{aligned}\lambda\biggl(C_\delta(y) \setminus \biggl[\bigcup_J B_j\biggr]\biggr) &= \lambda(C_\delta(y)) - \lambda\biggl(C_\delta(y) \cap \biggl[\bigcup_{j\in J} B_j\biggr]\biggr)\\ &\le \delta^d - (\delta - 3d^{1/2}h)^d \le 3d^{3/2}h\delta^{d-1},\end{aligned}$$

where the last inequality follows by induction. Thus,

$$\begin{aligned}\int_{C_\delta(y)\setminus\bigcup_J B_j} \phi(\mu; y; \sigma)\, d\mu &\le \lambda\biggl(C_\delta(y) \setminus \biggl[\bigcup_J B_j\biggr]\biggr) \frac{1}{(2\pi)^{d/2}\sigma^d}\\ &\le \frac{3d^{3/2}h\delta^{d-1}}{(2\pi)^{d/2}\sigma^d}. \qquad \square\end{aligned}$$

Lemma A.2. *Let $C_\delta(y)$ be a d-dimensional hypercube with center y and side length $\delta > 0$. Then*

$$\int_{C_\delta(y)} \phi(\mu; y; \sigma)\, d\mu > 1 - \frac{8d\sigma/\delta}{(2\pi)^{1/2}} \exp\biggl\{-\frac{(\delta/\sigma)^2}{8}\biggr\}.$$

Note that this inequality immediately implies that for any sub-hypercube of $C_\delta(y)$, $\tilde{C}$, with vertex at y and side length $\delta/2$, for example, $\tilde{C} = C_\delta(y) \cap [\mu \ge y]$,

$$\begin{aligned}\int_{\tilde{C}} \phi(\mu; y; \sigma)\, d\mu &= \frac{1}{2^d} \int_{C_\delta(y)} \phi(\mu; y; \sigma)\, d\mu\\ &> \frac{1}{2^d} - \frac{8d\sigma/\delta}{2^d(2\pi)^{1/2}} \exp\biggl\{-\frac{(\delta/\sigma)^2}{8}\biggr\}.\end{aligned}$$

Proof.

$$\begin{aligned}\int_{C_{\delta(y)}} \phi(\mu; y; \sigma)\, d\mu &= \int_{\bigcap_{i=1}^d [|\mu_i| \le \delta/2]} \phi(\mu; 0; \sigma)\, d\mu\\ &= 1 - \int_{\bigcup_{i=1}^d [|\mu_i| \ge \delta/2]} \phi(\mu; 0; \sigma)\, d\mu\\ &\ge 1 - \sum_{i=1}^d \int_{|\mu_i| \ge \delta/2} \phi(\mu_i; 0; \sigma)\, d\mu_i\\ &= 1 - 2d \int_{\delta/2}^\infty \phi(\mu_1; 0; \sigma)\, d\mu_1\\ &> 1 - \frac{2d}{(2\pi)^{1/2}\sigma} \int_{\delta/2}^\infty \exp\biggl\{-\frac{0.5(\delta/2)\mu_1}{\sigma^2}\biggr\}\, d\mu_1\end{aligned}$$

$$= 1 - \frac{2d}{(2\pi)^{1/2}\sigma}\frac{-\sigma^2}{0.5(\delta/2)}\exp\{-0.5(\delta/2)\mu_1/\sigma^2\}|_{\delta/2}^{\infty}$$

$$= 1 - \frac{8d(\sigma/\delta)}{(2\pi)^{1/2}}\exp\left\{-\frac{(\delta/\sigma)^2}{8}\right\}. \qquad \square$$

LEMMA A.3. *Let $A_1, \dots, A_m$ be a partition of an interval on R such that $\lambda(A_j) \le h$ and $\mu_j \in A_j$. Assume $C_\delta(y) = [y-\delta, y+\delta] \subset \cup A_j$ is an interval with center y and length δ. Then*

$$\sum_{j=1}^{m}\lambda(A_j \cap C_\delta(y))\phi(y,\mu_j,\sigma) \ge 1 - \frac{6h}{(2\pi)^{1/2}\sigma} - \frac{8(\sigma/\delta)}{(2\pi)^{1/2}}\exp\left\{-\frac{(\delta/\sigma)^2}{8}\right\}.$$

If $C_\delta(y) = [y-\delta, y]$ or $C_\delta(y) = [y, y+\delta]$ the lower bound in the above expression should be divided by 2.

PROOF. Let $J = \{j : A_j \cap C_\delta(y) \subset [y-\delta, y]\}$. For any $j \in J$ and $\mu \in A_j \cap C_\delta(y)$, $\mu - h \le \mu_j$ as $\lambda(A_j) < h$ and $\mu_j \in A_j$, which implies $\phi(y,\mu_j,\sigma) \ge \phi(y,\mu-h,\sigma)$. Therefore,

$$\sum_{j\in J}\lambda(A_j \cap C_\delta(y))\phi(y,\mu_j,\sigma) \ge \int_{\bigcup_{j\in J}[A_j\cap C_\delta(y)]}\phi(y,\mu-h,\sigma)\,d\mu. \tag{A.20}$$

Note next that

$$\begin{aligned}
&\int_{\bigcup_{j\in J}[A_j\cap C_\delta(y)]}\phi(y,\mu-h,\sigma)\,d\mu \\
&\qquad \ge \int_{y-\delta}^{y-h}\phi(y,\mu-h,\sigma)\,d\mu = \int_{y-\delta-h}^{y-2h}\phi(y,\mu,\sigma)\,d\mu \\
&\qquad = \int_{y-\delta}^{y}\phi(y,\mu,\sigma)\,d\mu \\
&\qquad\quad - \int_{y-\delta-h}^{y-\delta}\phi(y,\mu,\sigma)\,d\mu - \int_{y-2h}^{y}\phi(y,\mu,\sigma)\,d\mu \\
&\qquad \ge \int_{y-\delta}^{y}\phi(y,\mu,\sigma)\,d\mu - \frac{3h}{(2\pi)^{1/2}\sigma}.
\end{aligned}$$

By symmetry the same results can be obtained for $J = \{j : A_j \cap C_\delta(y) \subset [y, y+\delta]\}$. Thus

$$\sum_{j=1}^{m}\lambda(A_j \cap C_\delta(y))\phi(y,\mu_j,\sigma) \ge \int_{y-\delta}^{y+\delta}\phi(y,\mu,\sigma)\,d\mu - 2\frac{3h}{(2\pi)^{1/2}\sigma}.$$

The claim of the lemma follows by Lemma A.2. $\square$

Acknowledgments. The author is grateful to John Geweke and participants of seminars at Princeton, SBIES 09 and SITE 09 for helpful discussions. I thank Justinas Pelenis for pointing out shortcomings in several proofs. I thank an associate editor and anonymous referees for useful suggestions. All remaining errors are mine.

REFERENCES

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

Geweke, J. and Keane, M. (2007). Smoothly mixing regressions. *J. Econometrics* **138** 252–290. MR2380699

Ghosh, J. and Ramamoorthi, R. (2003). *Bayesian Nonparametrics*, 1st ed. Springer, New York. MR1992245

Hotz, J. and Miller, R. (1993). Conditional choice probabilities and the estimation of dynamic models. *Rev. Econom. Stud.* **60** 497–530. MR1236835

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Comput.* **3** 79–87. Available at http://dx.doi.org/10.1162/neco.1991.3.1.79.

Jansen, R. C. (1993). Maximum likelihood in a generalized linear finite mixture model by using the em algorithm. *Biometrics* **49** 227–231.

Jiang, W. and Tanner, M. (1999). Hierarchical mixtures-of-experts for exponential family regression models: Approximation and maximum likelihood estimation. *Ann. Statist.* **27** 987–1011. MR1724038

Jones, P. and McLachlan, G. J. (1992). Fitting finite mixture models in a regression context. *Aust. N. Z. J. Stat.* **34** 233–240.

Jordan, M. and Xu, L. (1995). Convergence results for the em approach to mixtures of experts architectures. *Neural Networks* **8** 1409–1431.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Comput.* **6** 181–214.

Kiefer, N. M. (1978). Discrete parameter variation: Efficient estimation of a switching regression model. *Econometrica* **46** 427–434. MR0483200

Li, J. Q. and Barron, A. R. (1999). Mixture density estimation. In *Advances in Neural Information Processing Systems* **12** 279–285. MIT Press, Cambridge, MA.

Maiorov, V. and Meir, R. (1998). Approximation bounds for smooth functions in c(rd) by neural and mixture networks. *Neural Networks, IEEE Transactions* **9** 969–978.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York. MR1789474

Norets, A. and Pelenis, J. (2009). Bayesian modeling of joint and conditional distributions. Unpublished manuscript, Princeton Univ.

Peng, F., Jacobs, R. A. and Tanner, M. A. (1996). Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *J. Amer. Statist. Assoc.* **91** 953–960.

Quandt, R. E. and Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *J. Amer. Statist. Assoc.* **73** 730–738. MR0521324

Roeder, K. and Wasserman, L. (1997). Practical bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.* **92** 894–902. MR1482121

Rust, J. (1996). Numerical dynamic programming in economics. In *Handbook of Computational Economics* (H. Amman, D. Kendrick and J. Rust, eds.). North-Holland, Amsterdam. Available at http://gemini.econ.umd.edu/jrust/sdp/ndp.pdf. MR1416619

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–540. MR0898357

Villani, M., Kohn, R. and Giordani, P. (2009). Regression density estimation using smooth adaptive Gaussian mixtures. *J. Econometrics* **153** 155–173.

Wedel, M. and DeSarbo, W. (1995). A mixture likelihood approach for generalized linear models. *J. Classification* **12** 21–55.

Wood, S., Jiang, W. and Tanner, M. (2002). Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika* **89** 513–528. MR1929159

Zeevi, A., Meir, R. and Maiorov, V. (1998). Error bounds for functional approximation and estimation using mixtures of experts. *IEEE Trans. Inform. Theory* **44** 1010–1025. MR1616675

Zeevi, A. J. and Meir, R. (1997). Density estimation through convex combinations of densities: Approximation and estimation bounds. *Neural Networks* **10** 99–109.

313 Fisher Hall
Department of Economics
Princeton University
Princeton, New Jersey 08544
USA
E-mail: anorets@princeton.edu