Submitted to Econometrica

ADAPTIVE BAYESIAN ESTIMATION OF DISCRETE-CONTINUOUS DISTRIBUTIONS UNDER SMOOTHNESS AND SPARSITY

ANDRIY NORETS[1]
AND JUSTINAS PELENIS

We consider nonparametric estimation of a mixed discrete-continuous distribution under anisotropic smoothness conditions and possibly increasing number of support points for the discrete part of the distribution. For these settings, we derive lower bounds on the estimation rates. Next, we consider a nonparametric mixture of normals model that uses continuous latent variables for the discrete part of the observations. We show that the posterior in this model contracts at rates that are equal to the derived lower bounds up to a log factor. Thus, Bayesian mixture of normals models can be used for (up to a log factor) optimal adaptive estimation of mixed discrete-continuous distributions. The proposed model demonstrates excellent performance in simulations mimicking first stage estimation of structural discrete choice models.

KEYWORDS: Bayesian nonparametrics, adaptive rates, minimax rates, anisotropic smoothness, posterior contraction, discrete-continuous distribution, mixed scale, mixtures of normal distributions, latent variables, discrete choice models.

.

1. INTRODUCTION

Nonparametric estimation methods have become more accessible and useful in empirical work due to availability of fast computers and very large datasets. The theory and practical implementation of nonparametric methods for continuous data are very well developed at this point. However, in most economic applications, the data contain both continuous and discrete variables. Nonparametric methods for multivariate discrete and mixed discrete-continuous distributions and their theoretical properties are less well understood and developed. We address this issue in the present paper.

The standard flexible approach to estimation of discrete distributions is to use sample frequencies as estimators of the corresponding probabilities. When the number of values that discrete variables can take is larger or comparable to the sample size, which we call here sparsity following Hall and Titterington (1987), the standard frequency estimators perform poorly. The sparsity in the multivariate case is rather a rule than an exception;

andriy_norets@brown.edu
pelenis@ihs.ac.at

[1]The first author gratefully acknowledges the support from NSF Grant SES-1851796.

for example, estimating a joint distribution of 5 discrete variables each taking 10 values would involve estimation of 10^5 probabilities by corresponding sample frequencies. The presence of continuous variables in addition to the discrete ones further exacerbates the problem. In economics, these issues often arise in the context of estimation of single-agent and game-theoretic static and dynamic discrete choice models. Popular two stage estimation procedures for these models pioneered by Hotz and Miller (1993) (and further developed by Aguirregabiria and Mira (2002), Pesendorfer and Schmidt-Dengler (2008), Bajari, Benkard, and Levin (2007) and Pakes, Ostrovsky, and Berry (2007) among others) deal with discrete dependent variables such as market entry decisions and discrete covariates such as the number of entrants currently in the market. A natural solution to this problem that appears to work well in practice (Aitchison and Aitken (1976), Li and Racine (2007), and the aforementioned references to the two stage discrete choice model estimation) is to smooth discrete data, hoping that probabilities at nearby discrete values are close or smoooth in some sense and that one could learn about a probability of a certain value from observations at nearby values. Of course, smoothing can only be beneficial if the underlying data have certain smoothness properties. Ideally, a procedure for estimation of discrete distributions should be able to optimally take advantage of smoothness in the data generating process if it is present and at the same time perform no worse than the standard frequency estimators if the data generating process is not (sufficiently) smooth.

In this paper, we formalize these ideas for multivariate mixed discrete-continuous distributions by setting up an asymptotic framework where the multivariate discrete part of the data generating distribution can have either a large or a small number of support points and it can be either very smooth or not, and these characteristics can differ from one discrete coordinate to another. In these settings, we derive optimal minimax rates for estimation of discrete-continuous distributions. We show that smoothing is beneficial only for a subset of discrete variables with a quickly growing number of support points and/or sufficiently high level of smoothness.

We propose an estimation procedure that adaptively (without a priori knowledge of smoothness levels of the data generating process) achieve the derived optimal convergence rates. The procedure is based on a Bayesian mixture of multivariate normal distributions. Mixture models have proven to be very useful for Bayesian nonparametric modeling of univariate and multivariate distributions of continuous variables. These models possess

outstanding asymptotic frequentist properties: in Bayesian nonparametric estimation of smooth densities the posterior in these models contracts at optimal adaptive rates up to a log factor (Rousseau (2010), Kruijer et al. (2010), Shen, Tokdar, and Ghosal (2013)). Tractable Markov chain Monte Carlo (MCMC) algorithms for exploring posterior distributions of these models are available (Escobar and West (1995), MacEachern and Muller (1998), Neal (2000), Miller and Harrison (2017), Norets (2020)) and they are widely used in empirical work (see Dey, Muller, and Sinha (1998), Chamberlain and Hirano (1999), Burda, Harding, and Hausman (2008), Chib and Greenberg (2010), and Jensen and Maheu (2014) among many others).

From the computational perspective, discrete variables can be easily accommodated through the use of continuous latent variables in Bayesian MCMC estimation (Albert and Chib (1993), McCulloch and Rossi (1994)). In nonparametric modelling of discrete-continuous data by mixtures, latent variables were used by Canale and Dunson (2011) and Norets and Pelenis (2012) among others. Some results on frequentist asymptotic properties of the posterior distribution in such models have also been established. Norets and Pelenis (2012) obtained approximation results in Kullback-Leibler distance and weak posterior consistency for mixture models with a prior on the number of mixture components. DeYoreo and Kottas (2017) establish weak posterior consistency for Dirichlet process mixtures. In similar settings, Canale and Dunson (2015) derived posterior contraction rates that are not optimal. In the present paper, we show that a mixture of normals model with a prior on the number of mixture components that uses latent variables for modeling the discrete part of the distribution can deliver optimal posterior contraction rates for nonparametric estimation of discrete-continuous distributions. The obtained optimal posterior contraction rates are adaptive since the priors we consider do not depend on the size of the support and the smoothness of the data generating process.

We illustrate our theoretical results in an application to the first stage estimation of discrete choice models briefly mentioned above. Specifically, we use data from Monte Carlo experiments in Pakes, Ostrovsky, and Berry (2007) who compare various two stage estimation procedures on a model of firm's entry decisions. Our procedure delivers 2.5 times reduction in the estimation error relative to the frequency estimator. Overall, our theoretical and simulation results suggest that models for discrete data based on mixtures and latent variables should be an important part of the econometric toolkit.

The rest of the paper is organized as follows. In Section 2, we describe our framework

and the Bayesian model. Section 3 presents simulation results and favorable comparisons with frequency and kernel estimators. The asymptotic theoretical results are presented in Section 4. MCMC algorithm for model estimation and proof outlines are given in Appendices. Auxiliary results and proof details are delegated to the online supplement.

2. DATA GENERATING PROCESS AND BAYESIAN MODEL

Let us denote the continuous part of observations by $x \in \mathcal{X} \subset \mathbb{R}^{d_x}$ and the discrete part by $y = (y_1, \ldots, y_{d_y}) \in \mathcal{Y}$, where

$$\mathcal{Y} = \prod_{j=1}^{d_y} \mathcal{Y}_j, \text{ with } \mathcal{Y}_j = \left\{ \frac{1 - 1/2}{N_j}, \frac{2 - 1/2}{N_j}, \ldots, \frac{N_j - 1/2}{N_j} \right\},$$

is a grid on $[0,1]^{d_y}$ (a product symbol Π applied to sets hereafter denotes a Cartesian product). The number of values that the discrete coordinates y_j can take, N_j, can potentially grow with the sample size or stay constant. For each discrete coordinate value $y_j \in \mathcal{Y}_j$, let

$$A_{y_j} = \begin{cases} (-\infty, y_j + 0.5/N_j] & \text{if } y_j = 0.5/N_j \\ (y_j - 0.5/N_j, \infty) & \text{if } y_j = 1 - 0.5/N_j \\ (y_j - 0.5/N_j, y_j + 0.5/N_j] & \text{otherwise} \end{cases}$$

be an interval that includes y_j and has a length of $1/N_j$, except for the first and the last intervals that are expanded to include the rest of the negative and positive parts of the real line correspondingly. Then, every value of the discrete part of observations $y = (y_1, \ldots, y_{d_y}) \in \mathcal{Y}$ can be associated with a hyper-rectangle $A_y = \prod_{j=1}^{d_y} A_{y_j}$. Let us represent the data generating density-probability mass function $p_0(y, x)$ as an integral of a latent density f_0 over A_y,

$$p_0(y, x) = \int_{A_y} f_0(\tilde{y}, x) d\tilde{y}, \tag{1}$$

where f_0 belongs to $\mathcal{D}$, the set of probability density functions on $\mathbb{R}^d$ with respect to the Lebesgue measure, and $d = d_x + d_y$. The representation of a mixed discrete-continuous distribution in (1) is so far without a loss of generality since for any given p_0 one could always define f_0 using a mixture of densities with non-overlapping supports included in A_y, $y \in \mathcal{Y}$.

We assume that the data available for estimation of p_0 are comprised of n independently identically distributed observations from p_0: $(Y^n, X^n) = (Y_1, X_1, \ldots, Y_n, X_n)$. Let P_0, E_0,

P_0^n, and E_0^n denote the probability measures and expectations corresponding to p_0 and its product p_0^n.

When N_j's grow with the sample size n the generality of the representation in (1) can be lost when assumptions such as smoothness are imposed on f_0. Nevertheless, in what follows we do allow for f_0 to be smooth. The interpretation of the smoothness is that the values of discrete variables can be ordered and that borrowing of information from nearby discrete points can be useful in estimation.

2.1. *Bayesian Model*

Our nonparametric Bayesian model for the data generating process in (1) is based on a mixture of normal distributions with a variable number of components for modelling the joint distribution of $(\tilde{y}, x)$,

$$f(\tilde{y}, x|\theta, m) = \sum_{k=1}^{m} \alpha_k \phi(\tilde{y}, x; \mu_k, \sigma \cdot \nu_k^{-1/2})$$

$$p(y, x|\theta, m) = \int_{A_y} f(\tilde{y}, x|\theta, m) d\tilde{y}, \tag{2}$$

where $\theta = (\mu_k, \nu_k, \alpha_k, k = 1, 2, \ldots; \sigma)$ and $\phi(\cdot; \mu_k, \sigma \cdot \nu_k^{-1/2})$ denotes a multivariate normal density with mean $\mu_k \in \mathbb{R}^d$ and a diagonal covariance matrix with the squared elements of vector $\sigma \cdot \nu_k^{-1/2} = (\sigma_1 \nu_{k1}^{-1/2}, \ldots, \sigma_d \nu_{kd}^{-1/2})$ on the diagonal.

We use the following prior for (θ, m). The prior for $(\alpha_1, \ldots, \alpha_m)$ conditional on m is Dirichlet$(a/m, \ldots, a/m)$, $a > 0$. It is a standard conjugate prior for discrete probability distributions, see, for example, Chamberlain and Imbens (2003) for applications in econometrics. The prior probability mass function for the number of mixture components m is

$$\Pi(m) \propto e^{-a_{10} m (\log m)^{\tau_1}}, m = 1, 2, \ldots, \quad a_{10} > 0, \tau_1 \geq 0, \tag{3}$$

where $\propto$ means "proportional to". The exponential tails of $\Pi(m)$ attain a tradeoff between putting just enough prior probability on the relevant finite mixture approximations of f_0 and putting appropriately small prior probabilities on rough mixtures that would overfit the data.

A popular alternative to specifying a prior on m and $(\alpha_1, \ldots, \alpha_m)$ is a Dirichlet process mixture (m is set to infinity and a "stick-breaking" prior (Sethuraman (1994)) is used for the infinite sequence of mixing weights $(\alpha_1, \ldots)$). This prior would deliver the same posterior contraction rates for continuous variables or settings where smoothing is important;

however, when smoothing is not beneficial, the Dirichlet process mixture prior does not seem to put sufficient weight on the relevant finite mixture approximations, and, hence, we focus on the mixtures of finite mixtures here.

The component specific scale parameters ν_k are not necessary for asymptotic results; it is a common practice in the literature to include them (see, for example, Geweke (2005)) and they seem to improve the finite sample performance. We use independent conditionally conjugate gamma-normal priors for (μ_{kj}, ν_{kj}). The common scale parameters σ are required to ensure that the prior puts sufficient probability on small values of the variances of all mixture components at once (the variances play a role of the bandwidth in asymptotic results). We use independent inverse Gamma priors for the components of σ. A detailed description of the model, priors, and the MCMC algorithm for model estimation is given in Appendix A. Section 4.3.1 provides more general conditions on the prior that deliver adaptive posterior contraction rates for the model in (2).

3. APPLICATION

In applied economics literature, nonparametric estimation of multivariate discrete or mixed discrete-continuous distributions is often used in the first stage of two stage estimation procedures for structural discrete choice models. Pakes, Ostrovsky, and Berry (2007) compare various two stage estimation procedures on a model of firm's entry decisions. Their Monte Carlo experiments provide convenient and realistic settings for demonstrating the performance of the mixture based models in practice.

The first stage in Pakes et al. (2007) requires estimation of entry and exit probabilities conditional on the number of entrants currently in the market and a discretized market size measure. These conditional probabilities are essentially obtained from the standard frequency estimator of the joint distribution for the four-dimensional vector of discrete random variables: the market size, the number of firms currently in the market, the number of new entrants, and the number of exiting firms. In what follows, we use the simulated data from Pakes et al. (2007) to compare our estimator with the standard frequency estimator and a classical kernel estimator with special discrete kernels from a publicly available R package *np* (Hayfield and Racine (2008)). The kernel bandwidth parameters are selected in the package by cross-validation as described in Li and Racine (2003); the latter authors provide simulation evidence that their methods outperform several other alternatives in the classical literature; the package *np* implements a wide variety of nonparametric methods presented in a textbook on nonparametric econometrics

by Li and Racine (2007).

Pakes et al. (2007) simulate a structural entry exit model to obtain one million draws for their Monte Carlo experiments. We use this one million simulated draws as a population distribution to estimate. The support of this population distribution consists of 2617 values of the four dimensional random vectors. The marginal population distributions of each vector component are depicted in Figure 1. From this population, we draw 50 random samples of size $n = 500$ (Pakes et al. (2007) use $n = 250$ and $n = 1000$ in their Monte Carlo experiments). For each sample, we compute the standard frequency estimator, the kernel estimator, and the mixture model estimators for a fixed $m \in \{1, \ldots, 30\}$ and a variable m. The MCMC algorithm for the fixed m model is standard in the literature (Diebolt and Robert (1994)). For the variable m model, we implemented two MCMC algorithms: an adaptation of a split-merge algorithm for Dirichlet process mixtures from Jain and Neal (2004) and an approximately optimal reversible jump algorithm from Norets (2020); they produce the same estimation results in the Monte Carlo experiments but the latter algorithm converges much faster. The reversible jump algorithm is described in detail in Appendix A.

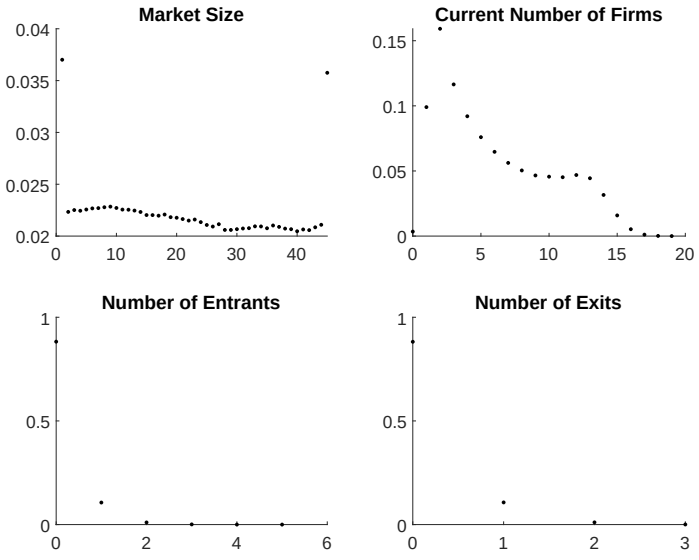

FIGURE 1.— Marginal population distributions

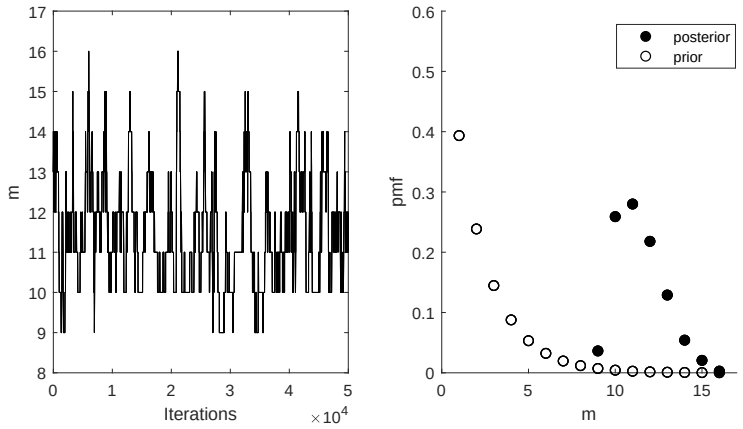

FIGURE 2.— MCMC trace plot and prior and posterior of m

Figure 2 presents the reversible jump MCMC draws and the prior and the posterior distributions of m for one of the samples used in Monte Carlo experiments. Estimation results for the fixed and variable m models are obtained from 10,000 and 50,000 MCMC draws correspondingly, as MCMC convergence is slower for the variable m models. As can be seen from the MCMC trace plot in the figure, the posterior simulator reliably explores the posterior distribution; MCMC results for other samples are similar.

The priors used in estimation experiments are roughly based on the first two sample moments: the prior for the location parameter μ_{kj} is centered at the corresponding sample average, $\bar{Y}_j = \sum_{i=1}^{n} Y_{ij}/n$ and has variance equal to the sample variance, $\hat{\sigma}_j^2 = \sum_{i=1}^{n}(Y_{ij} - \bar{Y}_j)^2/n$. The prior mode of the precision parameter σ_j^{-2} is set to the inverse of the sample variance, $\hat{\sigma}_j^{-2}$ and its variance is set to 1. The component specific scale parameters have prior mode and precision equal to 1. The Dirichlet parameter a is set to 15 and the prior hyper-parameters for m, $a_{10} = 0.5$ and $\tau_1 = 0$. The estimation results are not sensitive to moderate variation in prior hyper-parameters. These empirical Bayes priors are similar to unit variance priors centered at 0 for location parameters and 1 for scale parameters used in conjunction with standardized data.

The estimation errors in the total variation distance averaged over the 50 random samples are presented in Figure 3.

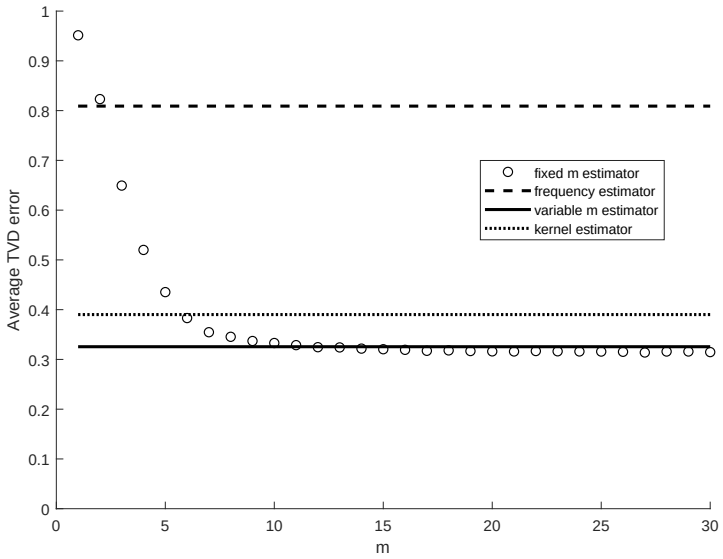

FIGURE 3.— Average Estimation Errors

As can be seen from the figure, the mixture based estimators match the average total variation error of the frequency estimator with just two mixture components and that of the kernel estimator with six mixture components. The use of a higher number of mixture components and a variable number of components further reduces the estimation error. The large improvements of both the kernel estimator and the mixture model over the standard frequency estimator are not very surprising given the smooth appearance of the probability mass functions in Figure 1, the sample size ($n = 500$), and the cardinality of the population support, 2617. The mixture models outperform the kernel estimator on average as shown in the figure and in each of the 50 random samples. Theoretical properties (beyond the consistency and the asymptotic normality for a fixed discrete support) are not known for the discrete kernel estimator in our asymptotic settings with smoothness and a possibly growing support. Our conjecture is that at least without considerable modifications this kernel estimator is unlikely to deliver the adaptive optimal estimation rates that are established for mixture models in the following section; and, perhaps, that is why the kernel estimator is outperformed by the mixture model in our simulations. A few other applications and favorable comparisons of a fixed m mixture model with standard parametric and nonparametric alternatives can be found in Norets and Pelenis (2012).

The performance of the variable m model is practically the same as the performance of models with large fixed m. Somewhat unexpectedly, the estimation results for the models with fixed m do not deteriorate when m is large ($m = 30$). The estimation error in the total variation distance is slightly more volatile for larger m, but on average, the

errors decrease in m as can be seen in Figure 3. Of course, the performance can be easily evaluated in simulation settings, when the data generating process is known. As far as we are aware, theoretically justified Bayesian procedures for choosing a fixed m have not been developed in nonparametric settings. Hence, the variable m model with the asymptotic guarantees obtained in the present paper is the preferred option, and the fixed m models should be used for sensitivity and robustness checks.

Overall, the Monte Carlo simulations presented in this section suggest that models for discrete data based on mixtures and latent variables should be an important part of the toolkit in empirical industrial organization and economics more generally. The following section presents asymptotic results that further justify this claim from the theoretical perspective.

4. ASYMPTOTIC FRAMEWORK AND RESULTS

To get more refined results and to accommodate discrete variables that are not ordered or "smooth", we allow N_j's to grow at different rates for different j's or to be constant for some j's. For the same reason, we work with anisotropic smoothness that allows for existence of derivatives of different orders along different coordinates.

4.1. *Anisotropic Smoothness*

For each coordinate $j \in \{1, \ldots, d\}$, we introduce a smoothness coefficient, $\beta_j > 0$, such that $\lfloor \beta_j \rfloor$ (the largest integer that is strictly smaller than β_j) is the highest possible order of the partial derivative with respect to the coordinate j. In the univariate case, $\lfloor \beta_j \rfloor$'th derivative is often assumed to satisfy a Holder condition with the exponent $\beta_j - \lfloor \beta_j \rfloor$ to accommodate noninteger smoothness coefficients and to deliver Taylor expansion approximations with remainders of the appropriate order. Different generalizations of these ideas to the multivariate case are possible. We introduce a generalization below that is suitable for our purposes. Let $\mathbb{Z}_+$ denote the set of non-negative integers. For smoothness coefficients $(\beta_1, \ldots, \beta_d)$ and an envelope constant L, an anisotropic $(\beta_1, \ldots, \beta_d)$-Holder class, $\mathcal{C}^{\beta_1,\ldots,\beta_d,L}$, is defined as follows.

DEFINITION 1 *$f \in \mathcal{C}^{\beta_1,\ldots,\beta_d,L}$ if for any $k = (k_1, \ldots, k_d) \in \mathbb{Z}_+^d$, $\sum_{l=1}^d k_l/\beta_l < 1$, mixed partial derivative of order k, $D^k f$, is finite and*

$$|D^k f(z + \Delta z) - D^k f(z)| \leq L \sum_{j=1}^{d} |\Delta z_j|^{\beta_j (1 - \sum_{l=1}^d k_l/\beta_l)}, \tag{4}$$

for any Δz *such that* $\Delta z_j = 0$ *when* $\sum_{l=1}^{d} k_l/\beta_l + 1/\beta_j < 1$.

In this definition, a Holder condition is imposed on $D^k f$ for a coordinate j when $D^k f$ cannot be differentiated with respect to z_j anymore ($\sum_{l=1}^{d} k_l/\beta_l < 1$ but $\sum_{l=1}^{d} k_l/\beta_l + 1/\beta_j \geq 1$). This definition slightly differs from definitions available in the literature on anisotropic smoothness that we found. Section 13.2 in Schumaker (2007) presents some very general anisotropic smoothness definitions but restricts attention to integer smoothness coefficients. Ibragimov and Hasminskii (1984), and most of the literature on minimax rates under anisotropic smoothness that followed including Barron et al. (1999) and Bhattacharya et al. (2014), do not restrict mixed derivatives. Shen et al. (2013) use $|\Delta z_j|^{\min(\beta_j - k_j, 1)}$ instead of $|\Delta z_j|^{\beta_j(1-\sum_l k_l/\beta_l)}$ in (4). Their requirement is stronger than ours for functions with bounded support, and it appears too strong for our derivation of lower bounds on the estimation rate. However, our definition is sufficiently strong to obtain a Taylor expansion with remainder terms that have the same order as those in Shen et al. (2013) (while the definitions that do not restrict mixed derivatives do not deliver such an expansion).

When $\beta_j = \beta$, $\forall j$ and $\sum_{l=1}^{d} k_l/\beta + 1/\beta \geq 1$, $\beta_j(1 - \sum_{l=1}^{d} k_l/\beta_l) = \beta - \lfloor \beta \rfloor$, and we get the standard definition of β-Holder smoothness for the isotropic case.

The envelope L can be assumed to be a function of $(z, \Delta z)$ to accommodate densities with unbounded support. We derive lower bounds on estimation rates for a constant envelope function; the derived bounds are applicable to functions with non-constant envelops as a constant envelop is just a special case of a non-constant one. Upper bounds on posterior contraction rates are derived under more general assumptions on L.

4.2. *Lower Bounds on Estimation Rates*

For a class of probability distributions $\mathcal{P}$, ζ is said to be a lower bound on the estimation error in metric ρ if there exists a positive constant c independent of n such that

$$\inf_{\hat{p}} \sup_{p \in \mathcal{P}} P\left(\rho(\hat{p}, p) \geq \zeta\right) \geq c > 0.$$

This definition means that there does not exist an estimator that asymptotically delivers an estimation error in ρ that is smaller than ζ for all data generating distributions in $\mathcal{P}$. If the estimation error for a given estimator for distributions in $\mathcal{P}$ matches (up to a multiplicative constant) a lower bound for $\mathcal{P}$, then this estimator is considered rate optimal. A good introduction into the theory of lower bounds can be found in Tsybakov

(2008). In this section, we present lower bounds for discrete continuous distributions that are matched with upper bounds on estimation errors for the mixture based models in Section 4.3.

We consider the following class of probability distributions: for a positive constant L, let

$$\mathcal{P} = \left\{ p : \; p(y,x) = \int_{A_y} f(\tilde{y},x) d\tilde{y}, \; f \in \mathcal{C}^{\beta_1,\ldots,\beta_d,L} \cap \mathcal{D} \right\}. \tag{5}$$

To define our lower bounds we need the following additional notation. Let $\mathcal{A}$ denote a collection of all subsets of indices for discrete coordinates $\{1,\ldots,d_y\}$. For $J \in \mathcal{A}$, let $J^c = \{1,\ldots,d\} \setminus J$ and y_J denotes the sub-vector $\{y_j, \, j \in J\}$ for a vector y. Then,

$$N_J = \prod_{j \in J} N_j$$

denotes the number of values a discrete subvector y_J can take and

$$\beta_{J^c} = \left[\sum_{j \in J^c} \beta_j^{-1} \right]^{-1}$$

denotes an aggregate smoothness coefficient for the subvector containing the coordinates of the continuous part of observations x and the continuous latent variables $\tilde{y}$ with indices in J^c. For $J = \emptyset$ or $J^c = \emptyset$, we set $N_\emptyset = 1$, $\beta_\emptyset = \infty$, and $\beta_\emptyset/(2\beta_\emptyset + 1) = 1/2$.

THEOREM 1 *For $\mathcal{P}$ defined in (5),*

$$\Gamma_n = \min_{J \in \mathcal{A}} \left[\frac{N_J}{n} \right]^{\frac{\beta_{Jc}}{2\beta_{Jc}+1}} = \left[\frac{N_{J_*}}{n} \right]^{\frac{\beta_{Jc_*}}{2\beta_{Jc_*}+1}} \tag{6}$$

multiplied by a positive constant is a lower bound on the estimation error in the total variation distance.

One could recognize expression $[N_J/n]^{\frac{\beta_{Jc}}{2\beta_{Jc}+1}}$ in (6) as the standard estimation rate for a $card(J^c)$-dimensional density with anisotropic smoothness coefficients $\{\beta_j, \, j \in J^c\}$ and the sample size n/N_J (Ibragimov and Hasminskii (1984)). One way to interpret this is that the density of $\{x, \tilde{y}_j, \, j \in J^c\}$ conditional on y_J is $\{\beta_j, \, j \in J^c\}$-smooth and the number of observations available for its estimation (observations with the same value of y_J) should be of the order n/N_J; also, the estimation rate for the marginal probability mass function for y_J is $[N_J/n]^{1/2}$, which is at least as fast as $[N_J/n]^{\frac{\beta_{Jc}}{2\beta_{Jc}+1}}$. In this interpretation, smoothing

is not performed over the discrete coordinates with indices in set J, and the lower bound is obtained when J minimizes $[N_J/n]^{\frac{\beta_{Jc}}{2\beta_{Jc}+1}}$. Thus, an estimator that delivers the rate in (6) should, in a sense, optimally choose the subset of discrete variables over which to perform smoothing.

It should be possible to extend the results on the lower bounds to other distances. However, suitable sufficient conditions in the Bayesian nonparametrics literature for the corresponding upper bounds appear to be currently available only for the total variation distance (or the Hellinger distance, which is equivalent); hence, we focus on that distance here.

The proof of Theorem 1 is given in Appendix B.

4.2.1. *Related Literature on Lower Bounds*

Let us briefly review most relevant results on lower bounds and place our results in that context. The most closely related results on minimax rates for anisotropic continuous distributions are developed in Ibragimov and Hasminskii (1984). The minimax estimation rates for mixed discrete continuous distributions appear to be studied first by Efromovich (2011). He considers discrete variables with a fixed support and no smoothness assumptions on the discrete part of the distribution. He shows that in these settings the optimal rates for discrete continuous distributions are equal to the optimal nonparametric rates for the continuous part of the distribution. Relaxing the assumption of the fixed support for the discrete part of the distribution is very desirable in nonparametric settings. It has been commonly observed at least since Aitchison and Aitken (1976) that smoothing discrete data in nonparametric estimation improves results in practice. Hall and Titterington (1987) introduced an asymptotic framework that provided a precise theoretical justification for improvements resulting from smoothing in the context of estimating a univariate discrete distribution with a support that can grow with the sample size. In their setup, the support is an ordered set and the probability mass function is β-smooth (in a sense that analogs of β-order Taylor expansions hold). They show that in their setup the minimax rate is the smaller one of the following two: (i) the optimal estimation rate for a continuous density with the smoothness level β, $n^{-\beta/(2\beta+1)}$, and (ii) the rate of convergence of the standard frequency estimator, $(N/n)^{1/2}$, where N is the cardinality of the support and n is the sample size. Hall and Titterington (1987) refer to their setup as "Sparse Multinomial Data" since N can be larger than n and this is the reason we refer

to sparsity in the title of the paper. Burman (1987) established similar results for $\beta = 2$. Subsequent literature in multivariate settings (e.g., Dong and Simonoff (1995), Aerts et al. (1997)) did not consider lower bounds but demonstrated that when the support of the discrete distribution grows sufficiently fast then estimators that employ smoothing can achieve the standard nonparametric rates for β-smooth densities on $\mathbb{R}^d$, $n^{-\beta/(2\beta+d)}$.

We generalize the results of Hall and Titterington (1987) on lower bounds for univariate discrete distributions to multivariate mixed discrete-continuous case and anisotropic smoothness. Alternatively, our results can be viewed as a generalization of results in Efromovich (2011) to settings with anisotropic smoothness and potentially growing supports for discrete variables.

4.3. *Posterior Contraction Rates for a Mixture of Normals Model*

4.3.1. *Assumptions on Prior*

The assumptions on the prior for model (2) in Section 2.1 can be slightly generalized as follows. A priori, the components of μ_k, μ_{kj}, $k = 1, \ldots, m$, $j = 1, \ldots, d$ are assumed independent from each other, other parameters, and across k. Prior density for μ_{kj} is bounded below for some $a_{12}, \tau_2 > 0$ by

$$a_{11} \exp(-a_{12}|\mu_{kj}|^{\tau_2}), \tag{7}$$

and for some $a_{13}, \tau_3 > 0$ and all sufficiently large $\mu_{kj} > 0$,

$$\Pi(\mu_{kj} \notin [-\mu, \mu]) \leq e^{-a_{13}\mu^{\tau_3}}. \tag{8}$$

Normal priors for μ_{kj} satisfy these conditions.

For positive constants $a_1, a_2, \ldots, a_9$, for each $j \in \{1, \ldots, d\}$, σ_j is assumed independent of other parameters a priori and the prior satisfies

$$\Pi(\sigma_j^{-2} \geq s) \leq a_1 e^{-a_2 s^{a_3}} \text{ for all sufficiently large } s > 0 \tag{9}$$

$$\Pi(\sigma_j^{-2} < s) \leq a_4 s^{a_5} \text{ for all sufficiently small } s > 0 \tag{10}$$

$$\Pi\{s < \sigma_j^{-2} < s(1+t)\} \geq a_6 s^{a_7} t^{a_8} e^{-a_9 s^{1/2}},\ s > 0,\ t \in (0,1). \tag{11}$$

The inverse Gamma prior for σ_i satisfies (9)-(11).

A prior on m that can be bounded above and below by functions in the form of the right hand side of (3), possibly with different constants, would work; to simplify the notation

we assume (3). We also set the component specific scale parameters ν_{ji} to 1. An extension of the posterior contraction results to variable ν_{kj}'s is straightforward, see, for example, Theorem A.5 in Norets and Pati (2017) for continuous variables, and it is not presented here for brevity.

4.3.2. *Posterior Contraction Rates*

This section presents upper bounds on the posterior contraction rates for the Bayesian mixture model that match the lower bounds in Section 4.2 up to a log factor. That means that the Bayesian mixture model deliver a rate optimal (up to a log) estimator for the data generating process in (1) under our smoothness assumptions. The estimator is adaptive since the prior and model specification do not depend on the smoothness of the data generating density and the fineness of the support relative to the sample size. To simplify the exposition we present the results below in Theorem 2 for the case when the data generating latent density f_0 has a bounded support.

THEOREM 2 *Assume the conditions on the prior in Sections 4.3.1. Suppose* $f_0 \in \mathcal{C}^{\beta_1,\ldots,\beta_d,L}$ *and* $\overline{f} \geq f_0 \geq \underline{f} > 0$ *holds on the support of* f_0*, where* L*,* $\overline{f}$*, and* $\underline{f}$ *are finite positive constants. Let*

$$\epsilon_n = \min_{J \in \mathcal{A}} \left[\frac{N_J}{n}\right]^{\beta_{J^c}/(2\beta_{J^c}+1)} (\log n)^{t_J} \tag{12}$$

where

$$t_J > \left(d_{J^c} + \beta_{J^c}^{-1} + \max\{\tau_1, 1\}\right) / \left(2 + \beta_{J^c}^{-1}\right) + \max\{0, (1-\tau_1)/2\}$$

and τ_1 *is a parameter in the prior on* m*. Suppose also* $n\epsilon_n^2 \to \infty$ *and for* J_* *that attains the minimum in (12),* $N_{J_*} = o(n^{1-\nu})$ *for some small* $\nu > 0$*. Then, the posterior contracts at the rate* ϵ_n*: there exists* $\bar{M} > 0$ *such that*

$$\Pi\left(p : d_{TV}(p, p_0) > \bar{M}\epsilon_n | Y^n, X^n\right) \overset{P_0^n}{\to} 0.$$

As in Section 4.2, when $J^c = \emptyset$, β_{J^c} can be defined to be infinity and $\beta_{J^c}/(2\beta_{J^c}+1) = 1/2$ in (12). The assumption $N_{J_*} = o(n^{1-\nu})$ excludes the cases with very slow (non-polynomial) rates as some parts of the proof require $\log(1/\epsilon_n)$ to be of order $\log n$.

The theorem is a special case of the results presented in Appendix C that can accommodate unbounded support for f_0. The proof of Theorem 2 follows from the discussion of the more general assumptions in the appendix as the bounded support case is used there to

illustrate the assumptions. Similarly to other papers on posterior contraction for mixtures of normal densities though, the more general sufficient conditions in the appendix require subexponential tails for f_0. The results for f_0 with an unbounded support also require the envelope function L in the smoothness definition to be comparable to f_0.

The proof of the posterior contraction results is based on the general sufficient conditions from Ghosal et al. (2000). It exploits approximations of smooth densities by mixtures of normal distributions developed in the Bayesian nonparametrics literature (Rousseau (2010), Kruijer et al. (2010), de Jonge and van Zanten (2010), and Shen, Tokdar, and Ghosal (2013)) and also develops appropriate approximations for nonsmooth discrete distributions. Posterior contraction rates for nonparametric density estimation by mixture models derived in the aforementioned papers also include a log factor similar to $(\log n)^{t_J}$ in (12). It is not known in the literature whether the log factor can be avoided; however, it is not a very important issue as the log factor is negligible compared to the polynomial part of the rate.

The results on the upper bounds in this section and lower bounds in Section 4.2 also hold for the data generating processes where f_0 is not smooth at all in some discrete coordinates. The resulting rates can be obtained from those we derive by setting the corresponding coordinates in β to (values arbitrarily close to) zero in (6), so that for the optimal rate, smoothing is effectively not performed for these coordinates. Thus, the proposed Bayesian model achieves the objective outlined in the introduction: it optimally takes advantage of smoothness in the data generating process if it is present and at the same time performs no worse than the standard frequency estimators if the data generating process is not (sufficiently) smooth. Simulations in Section 3 suggest that the model performs better in practice than available parametric and nonparametric alternatives and appears to live up to its excellent theoretical properties.

5. FUTURE WORK

In many applications, conditional rather than joint distributions are actually of interest. Of course, one could always estimate the joint distribution and then extract the conditional distributions of interest. When the smoothness of the joint and conditional distributions is the same then rate optimality of joint distribution estimator implies rate optimality for the corresponding conditional distribution estimator. However, when the conditional distribution is smoother then it could be beneficial to estimate the conditional distribution directly. In an ongoing work, we pursue an extension of our posterior contraction results

to conditional distribution models based on covariate dependent mixtures; the extension is similar to work by Norets and Pati (2017) on continuous distributions.

It would also be of interest to understand if other Bayesian nonparametric models (for example, those based on Gaussian process priors) or classical nonparametric methods based on higher order kernels or orthogonal series expansions can deliver estimators with adaptive optimal convergence rates in our asymptotic framework.

REFERENCES

Aerts, M., I. Augustyns, and P. Janssen (1997): "Local Polynomial Estimation of Contingency Table Cell Probabilities," *Statistics*, 30, 127–148.

Aguirregabiria, V. and P. Mira (2002): "Swapping the Nested Fixed Point Algorithm: A Class of Estimators for Discrete Markov Decision Models," *Econometrica*, 70, 1519–1543.

Aitchison, J. and C. G. G. Aitken (1976): "Multivariate binary discrimination by the kernel method," *Biometrika*, 63, 413–420.

Albert, J. H. and S. Chib (1993): "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Bajari, P., L. Benkard, and J. Levin (2007): "Estimating Dynamic Models of Imperfect Competition," *Econometrica*, 75, 1331–1370.

Barron, A., L. Birgé, and P. Massart (1999): "Risk bounds for model selection via penalization," *Probab. Theory Related Fields*, 113, 301–413.

Bhattacharya, A., D. Pati, and D. Dunson (2014): "Anisotropic function estimation using multi-bandwidth Gaussian processes," *The Annals of Statistics*, 42, 352–381.

Burda, M., M. Harding, and J. Hausman (2008): "A Bayesian Mixed Logit-Probit Model for Multinomial Choice," *Journal of Econometrics*, 147, pp. 232–246.

Burman, P. (1987): "Smoothing Sparse Contingency Tables," *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, 49, 24–36.

Canale, A. and D. B. Dunson (2011): "Bayesian Kernel Mixtures for Counts," *Journal of the American Statistical Association*, 106, 1528–1539.

——— (2015): "Bayesian multivariate mixed-scale density estimation," *Statistics and its Interface*, 8, 195–201.

Chamberlain, G. and K. Hirano (1999): "Predictive Distributions Based on Longitudinal Earnings Data," *Annales d'Economie et de Statistique*, 211–242.

Chamberlain, G. and G. W. Imbens (2003): "Nonparametric Applications of Bayesian Inference," *Journal of Business and Economic Statistics*, 21, 12–18.

Chib, S. and E. Greenberg (1995): "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, 49, 327–335.

——— (2010): "Additive cubic spline regression with Dirichlet process mixture errors," *Journal of Econometrics*, 156, 322–336.

de Jonge, R. and J. H. van Zanten (2010): "Adaptive nonparametric Bayesian inference using location-scale mixture priors," *The Annals of Statistics*, 38, 3300–3320.

Dey, D., P. Muller, and D. Sinha, eds. (1998): *Practical Nonparametric and Semiparametric Bayesian Statistics*, Lecture Notes in Statistics , Vol. 133, Springer.

DeYoreo, M. and A. Kottas (2017): "Bayesian Nonparametric Modeling for Multivariate Ordinal Regression," *Journal of Computational and Graphical Statistics*, 0, 1–14.

Diebolt, J. and C. P. Robert (1994): "Estimation of Finite Mixture Distributions through Bayesian Sampling," *Journal of the Royal Statistical Society. Series B (Methodological)*, 56, 363–375.

Dong, J. and J. S. Simonoff (1995): "A Geometric Combination Estimator for d-Dimensional Ordinal Sparse Contingency Tables," *Ann. Statist.*, 23, 1143–1159.

Efromovich, S. (2011): "Nonparametric estimation of the anisotropic probability density of mixed variables," *Journal of Multivariate Analysis*, 102, 468 – 481.

Escobar, M. and M. West (1995): "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Geweke, J. (2005): *Contemporary Bayesian Econometrics and Statistics*, Wiley-Interscience.

Ghosal, S., J. K. Ghosh, and A. W. v. d. Vaart (2000): "Convergence Rates of Posterior Distributions," *The Annals of Statistics*, 28, 500–531.

Ghosal, S. and A. van der Vaart (2007): "Posterior convergence rates of Dirichlet mixtures at smooth densities," *The Annals of Statistics*, 35, 697–723.

Ghosal, S. and A. W. van der Vaart (2001): "Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities," *The Annals of Statistics*, 29, 1233–1263.

Green, P. J. (1995): "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

Hall, P. and D. M. Titterington (1987): "On Smoothing Sparse Multinomial Data," *Australian Journal of Statistics*, 29, 19–37.

Hayfield, T. and J. S. Racine (2008): "Nonparametric Econometrics: The np Package," *Journal of Statistical Software*, 27, 1–32.

Hotz, J. and R. Miller (1993): "Conditional Choice Probabilities and the Estimation of Dynamic Models," *Review of Economic Studies*, 60, 497–530.

Ibragimov, I. and R. Hasminskii (1977): "Estimation of infinite-dimensional parameter in Gaussian white noise," *Doklady Akademii Nauk SSSR*, 236, 1053–1055.

Ibragimov, I. A. and R. Z. Hasminskii (1984): "More on the estimation of distribution densities," *Journal of Soviet Mathematics*, 25, 1155–1165.

Jain, S. and R. M. Neal (2004): "A Split-Merge Markov chain Monte Carlo Procedure for the Dirichlet Process Mixture Model," *Journal of Computational and Graphical Statistics*, 13, 158–182.

Jensen, M. J. and J. M. Maheu (2014): "Estimating a semiparametric asymmetric stochastic volatility model with a Dirichlet process mixture," *Journal of Econometrics*, 178, 523–538.

Kruijer, W., J. Rousseau, and A. van der Vaart (2010): "Adaptive Bayesian density estimation with location-scale mixtures," *Electronic Journal of Statistics*, 4, 1225–1257.

LI, Q. AND J. RACINE (2003): "Nonparametric estimation of distributions with categorical and continuous data," *Journal of Multivariate Analysis*, 86, 266–292.

LI, Q. AND J. S. RACINE (2007): *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

MACEACHERN, S. AND P. MULLER (1998): "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223–238.

MCCULLOCH, R. AND P. ROSSI (1994): "An exact likelihood analysis of the multinomial probit model," *Journal of Econometrics*, 64, 207–240.

METROPOLIS, N., A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER, AND E. TELLER (1953): "Equation of State Calculations by Fast Computing Machines," *The Journal of Chemical Physics*, 21, 1087–1092.

MILLER, J. W. AND M. T. HARRISON (2017): "Mixture Models With a Prior on the Number of Components," *Journal of the American Statistical Association*, 0, 1–17.

NEAL, R. (2000): "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 249–265.

NORETS, A. (2020): "Optimal Auxiliary Priors and Reversible Jump Proposals for a Class of Variable Dimension Models," *Econometric Theory*, Forthcoming.

NORETS, A. AND D. PATI (2017): "Adaptive Bayesian Estimation of Conditional Densities," *Econometric Theory*, 33, 980–1012.

NORETS, A. AND J. PELENIS (2012): "Bayesian modeling of joint and conditional distributions," *Journal of Econometrics*, 168, 332–346.

——— (2014): "Posterior Consistency in Conditional Density Estimation by Covariate Dependent Mixtures," *Econometric Theory*, 30, 606–646.

PAKES, A., M. OSTROVSKY, AND S. BERRY (2007): "Simple estimators for the parameters of discrete dynamic games (with entry/exit examples)," *the RAND Journal of Economics*, 38, 373–399.

PESENDORFER, M. AND P. SCHMIDT-DENGLER (2008): "Asymptotic Least Squares Estimators for Dynamic Games," *Review of Economic Studies*, 75, 901–928.

ROUSSEAU, J. (2010): "Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density," *The Annals of Statistics*, 38, 146–180.

SCHUMAKER, L. (2007): *Spline functions : basic theory*, Cambridge New York: Cambridge University Press.

SETHURAMAN, J. (1994): "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650.

SHEN, W., S. T. TOKDAR, AND S. GHOSAL (2013): "Adaptive Bayesian multivariate density estimation with Dirichlet mixtures," *Biometrika*, 100, 623–640.

SMITH, A. F. AND G. O. ROBERTS (1993): "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society: Series B (Methodological)*, 55, 3–23.

TSYBAKOV, A. B. (2008): *Introduction to Nonparametric Estimation (Springer Series in Statistics)*, Springer, New York, USA.

APPENDIX A: MODEL, PRIORS, AND MCMC ALGORITHM

A.1. *Model and Priors*

For the MCMC implementation and description, it is convenient to formulate the model in (2) using mixture allocation latent variables (Diebolt and Robert (1994)), $(s_1, \ldots, s_n)$, latent variables $(\tilde{Y}_1, \ldots, \tilde{Y}_n)$ corresponding to discrete observations, and precision parameters $h_j = \sigma_j^{-2}$ so that for each observation index $i \in \{1, \ldots, n\}$ and mixture component index $k \in \{1, \ldots, m\}$,

$$(\tilde{Y}_i, X_i)|s_i = k, \mu_k, h, \nu_j, m \sim \phi\left(\cdot; \mu_k, (h_1^{-1/2}\nu_{k1}^{-1/2}, \ldots, h_d^{-1/2}\nu_{kd}^{-1/2})\right),$$

$$p(s_i = k|\theta, m) = \alpha_k.$$

The joint distribution of observables and unobservables in the model is

$$\begin{aligned}(13)\qquad & p\left(Y_i, \tilde{Y}_i, X_i, s_i,\ i = 1, \ldots, n; \mu_1, \nu_1, \ldots, \mu_m, \nu_m; h, m\right) = \\ & \prod_{i=1}^{n} 1\{\tilde{Y}_i \in A_{Y_i}\}\phi\left(\tilde{Y}_i, X_i; \mu_{s_i}, (h_1^{-1/2}\nu_{s_i1}^{-1/2}, \ldots, h_d^{-1/2}\nu_{s_id}^{-1/2})\right)\ \alpha_{s_i} \\ & \cdot \Pi(\alpha_1, \ldots, \alpha_m|m) \cdot \prod_{j=1}^{d} \Pi(h_j) \prod_{k=1}^{m} \Pi(\mu_{kj}|\nu_{kj})\Pi(\nu_{kj}) \cdot \Pi(m).\end{aligned}$$

The common precision parameter, h_j, is a priori distributed as a square of a gamma distributed random variable with shape $\underline{A}_{h_j}$ and rate $\underline{B}_{h_j}$, which is consistent with the conditions in Section 4.3.1:

$$\Pi(h_j) \propto h_j^{\underline{A}_{h_j}/2-1} e^{-\underline{B}_{h_j} \cdot h_j^{1/2}}.$$

The priors for (ν_{kj}, μ_{kj}) are conditionally conjugate gamma-normal:

$$\Pi(\nu_{kj}) \propto \nu_{kj}^{\underline{A}_{\nu_j}-1} e^{-\underline{B}_{\nu_j} \cdot \nu_{kj}},$$

$$\Pi(\mu_{kj}|\nu_{kj}) \propto \nu_{kj}^{1/2} e^{-0.5\underline{h}_{\mu_j}\nu_{kj}(\mu_{kj}-\underline{\mu}_j)^2}.$$

The priors for mixing weights and m are as described in Section 2.1:

$$\Pi(\alpha_1, \ldots, \alpha_m|m) \propto \prod_{k=1}^{m} \alpha_k^{a/m-1},$$

$$\Pi(m) \propto e^{-a_{10}m(\log m)^{\tau_1}}.$$

A.2. *MCMC Algorithm*

We develop a Metropolis-within-Gibbs algorithm (Metropolis et al. (1953)) with a reversible jump step for m (Green (1995)) for exploring the posterior distribution. See, for example, Chib and Greenberg (1995) for an introduction to Metropolis-Hastings algorithm, Smith and Roberts (1993) for an introduction to Gibbs sampling, or Geweke (2005) for a textbook treatment of MCMC algorithms in general and for mixture models in particular.

Conditional on m, the distributions for the Gibbs sampler blocks of the parameters and the latent variables are proportional to (13) and can be written as follows:

$$\tilde{Y}_{ij}|\ldots \sim \phi\left(\tilde{Y}_{ij};\mu_{s_ij},h_j^{-1/2}\nu_{s_ij}^{-1/2}\right)\cdot 1\{\tilde{Y}_{ij}\in A_{Y_{ij}}\} \quad \text{(truncated normal)}$$

$$p(s_i=k|\ldots)\propto \phi\left(\tilde{Y}_i,X_i;\mu_k,(h_1^{-1/2}\nu_{k1}^{-1/2},\ldots,h_d^{-1/2}\nu_{kd}^{-1/2})\right)\ \alpha_k \quad \text{(multinomial)}$$

$$p(\alpha_1,\ldots,\alpha_m|\ldots)\propto\prod_{k=1}^m \alpha_k^{a/m+\sum_{i=1}^n 1\{s_i=k\}-1} \quad \text{(Dirichlet)}$$

$$p(\mu_{kj},\nu_{kj}|\ldots)\propto \nu_{kj}^{\bar{A}_{\nu_j}-1/2}e^{-\bar{B}_{\nu_j}\cdot\nu_{kj}-0.5\bar{h}_{\mu_j}\nu_{kj}(\mu_{kj}-\bar{\mu}_j)^2} \quad \text{(gamma-normal)}$$

with parameters

$$\bar{h}_{\mu_j}=\underline{h}_{\mu_j}+h_j\cdot\sum_{i=1}^n 1\{s_i=k\},\quad \bar{\mu}_j=\bar{h}_{\mu_j}^{-1}[\underline{h}_{\mu_j}\underline{\mu}_j+h_j\cdot\sum_{i:\,s_i=k}\tilde{Y}_{ij}],$$

$$\bar{A}_{\nu_j}=\underline{A}_{\nu_j}+0.5\sum_{i=1}^n 1\{s_i=k\},\quad \bar{B}_{\nu_j}=\underline{B}_{\nu_j}+0.5[h_j\sum_{i:\,s_i=k}\tilde{Y}_{ij}^2+\underline{h}_{\mu_j}\underline{\mu}_j^2-\bar{h}_{\mu_j}\bar{\mu}_j^2].$$

The block for h_j is simulated by the Metropolis-Hastings-within-Gibbs with a gamma proposal with shape parameter $\underline{A}_{h_j}/2+n/2$, rate parameter $0.5\sum_{i=1}^n\nu_{s_ij}(\tilde{Y}_{ij}-\mu_{s_ij})^2$ and the Metropolis-Hastings acceptance probability $\min\{1,e^{\underline{B}_{h_j}(h_j^{0.5}-(h_j^*)^{0.5})}\}$, where h_j^* is the proposal and h_j is the current value. In the descriptions of blocks for μ_{kj}, ν_{kj}, and h_j above, it was implicitly assumed that index j refers to discrete coordinates ($j\in\{1,\ldots,d_y\}$); for $j\geq d_y$, $\tilde{Y}_{ij}$ should be replaced by X_{ij} in the descriptions of these blocks.

For the model with variable m, a block for m is added to the MCMC algorithm. The update for m is performed by an approximately optimal reversible jump algorithm from Norets (2020). To apply the algorithm we first transform the mixing weights into unnormalized weights $\tilde{\alpha}_k$, $k=1,\ldots$, so that conditional on m, $\alpha_k=\tilde{\alpha}_k/\sum_{l=1}^m\tilde{\alpha}_l$ and the Dirichlet prior on $(\alpha_1,\ldots,\alpha_m)$ corresponds to a gamma prior for the unnormalized weights: $\tilde{\alpha}_k|m\sim Gamma(a/m,1)$, $k=1,\ldots,m$. Let $\theta_k=(\mu_k,\nu_k,\tilde{\alpha}_k)$, $\theta_{1m}=(h,\theta_1,\ldots,\theta_m)$,

$Y = \{Y_i, \tilde{Y}_i, X_i \, i = 1, \ldots, n\}$ and denote a proposal distribution for the parameter of a new mixture component $m+1$ by $\tilde{\pi}_{m+1}(\theta_{m+1}|Y, \theta_{1m})$. The algorithm works as follows. Simulate proposal m^* from $Pr(m^* = m+1|m) = Pr(m^* = m-1|m) = 1/2$. If $m^* = m+1$, then also simulate $\theta_{m+1} \sim \tilde{\pi}_{m+1}(\theta_{m+1}|Y, \theta_{1m})$. Accept the proposal with probability $\min\{1, \alpha(m^*, m)\}$, where

$$\alpha(m^*, m) = \frac{p(Y|m^*, \theta_{1m^*})\Pi(\theta_{1m^*}|m^*)\Pi(m^*)}{p(Y|m, \theta_{1m})\Pi(\theta_{1m}|m)\Pi(m)} \cdot \left(\frac{1\{m^* = m+1\}}{\tilde{\pi}_m(\theta_{m+1}|\theta_{1m}, Y)} + 1\{m^* = m-1\}\tilde{\pi}_{m-1}(\theta_m|\theta_{1m-1}, Y) \right). \tag{14}$$

Norets (2020) shows that an optimal choice of proposal $\tilde{\pi}_m$ is the conditional posterior $p(\theta_{m+1}|Y, m+1, \theta_{1m})$. The conditional posterior can be evaluated up to a normalization constant; however, it seems hard to directly simulate from it and compute the required normalization constant. Hence, we use a Gaussian approximation to $p(\theta_{m+1}|Y, m+1, \theta_{1m})$ as the proposal (with the mean equal to the conditional posterior mode, obtained by a Newton method, and the variance equal to the inverse of the negative of the Hessian evaluated at the mode).

From an initial value of parameters, $(\theta_{1m}^{(0)}, m^{(0)})$, the MCMC algorithm sequentially updates parameters by simulating from the algorithm blocks. The resulting Markov chain, $(\theta_{1m}^{(r)}, m^{(r)})$, $r = 1, \ldots, M$, is used to approximate posterior objects of interest such as the posterior predictive (or posterior mean) density-point mass

$$p(y, x|Y^n, X^n) \approx \frac{1}{M} \sum_{r=1}^{M} p(y, x|\theta_{1m}^{(r)}, m^{(r)}).$$

APPENDIX B: PROOF OUTLINE FOR LOWER BOUNDS

In this section, we set up the notation and an outline of the proof of Theorem 1. Detailed calculations are delegated to lemmas in the supplement. The proof is based on a general theorem from the literature on lower bounds, which we present next in a slightly simplified form.

LEMMA 1 *(Theorem 2.5 in Tsybakov (2008), see also Ibragimov and Hasminskii (1977))* *ζ is a lower bound on the estimation error in metric ρ for a class $\mathcal{Q}$ if there exist a positive integer $M \geq 2$ and $q_j, q_i \in \mathcal{Q}$, $0 \leq j < i \leq M$ such that $\rho(q_j, q_i) \geq 2\zeta$, $q_j << q_0$, $j = 1, \ldots, M$ and*

$$\sum_{j=1}^{M} KL(Q_j^n, Q_0^n)/M < \log(M)/8, \tag{15}$$

where KL is the Kullback-Leibler divergence and Q_j^n is the distribution of a random sample from q_j.

The following standard result on bounding the number of unequal elements in binary sequences is used in our construction of q_j, $j = 1, \ldots, M$.

LEMMA 2 *(Varshamov-Gilbert bound, Lemma 2.9 in Tsybakov (2008)) Consider the set of all binary sequences of length $\bar{m}$,*

$$\Omega = \{w = (w_1, \ldots, w_{\bar{m}}) : \ w_r \in \{0, 1\}\} = \{0, 1\}^{\bar{m}}.$$

Suppose $\bar{m} \geq 8$. Then there exists a subset $\{w^1, \ldots, w^M\}$ of Ω such that $w^0 = (0, \ldots, 0)$,

$$\sum_{r=1}^{\bar{m}} 1\{w_r^j \neq w_r^i\} \geq \bar{m}/8, \ \forall 0 \leq j < i \leq M,$$

and

$$M \geq 2^{\bar{m}/8}.$$

To define q_j's for our problem, we need some additional notation. Let

$$K_0(u) = \exp\{-1/(1 - u^2)\} \cdot 1\{|u| \leq 1\}.$$

This function has bounded derivatives of all orders and it smoothly decreases to zero at the boundary of its support. This type of kernel functions is usually used for constructing hypotheses for lower bounds, see Section 2.5 in Tsybakov (2008). Since we need to construct a smooth density that integrates to 1, we define (as illustrated in Figure 4)

$$g(u) = c_0[K_0(4(u + 1/4)) - K_0(4(u - 1/4))],$$

where $c_0 > 0$ is a sufficiently small constant that will be specified below.

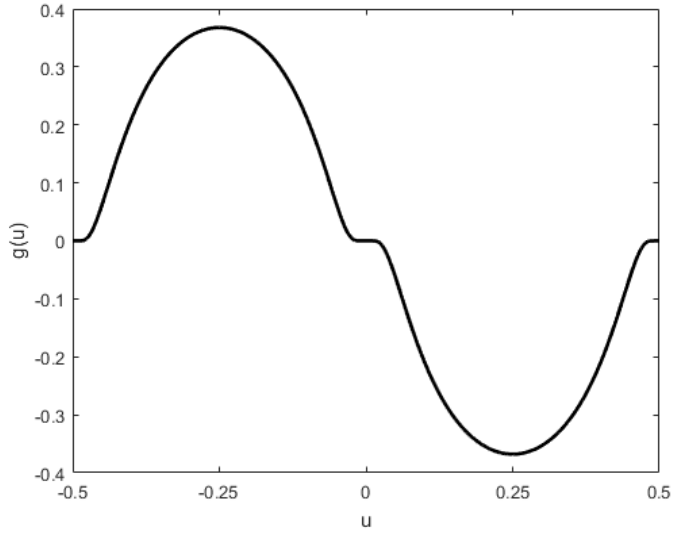

FIGURE 4.— Function g for $c_0 = 1$.

Function g will be used as a kernel in construction of q_k's. Let us define the bandwidth for these kernels first.

For the continuous coordinates, we define the bandwidth as in Ibragimov and Hasminskii (1984),

$$h_i = \Gamma_n^{1/\beta_i},\ i \in \{d_y + 1, \dots, d\}.$$

For the discrete ones, over which smoothing is beneficial, we define the bandwidth as

$$h_i = \varrho_i \cdot \Gamma_n^{1/\beta_i} = \frac{2}{N_i} \cdot R_i,\ i \in J_*^c \cap \{1, \dots, d_y\},$$

where $R_i = \lfloor \Gamma_n^{1/\beta_i} N_i/2 \rfloor + 1$ is a positive integer and $\varrho_i \in (1, 2]$ as shown in Lemma 7.

For the rest of the discrete coordinates, our innovation is to first define artificial anisotropic smoothness coefficients $\beta_i^* = -\log(\Gamma_n)/\log N_i$, $i \in J_*$, at which the rate in (6) would have the same value whether we smooth over y_i ($i \in J_*^c$) or not ($i \in J_*$). Then, we define the bandwidth as

$$h_i = 2 \cdot \Gamma_n^{1/\beta_i^*} = 2/N_i,\ i \in J_*.$$

To streamline the notation, we also define $\beta_i^* = \beta_i$ for $i \in J_*^c$.

Let m_i be the integer part of h_i^{-1}, $i = 1, \dots, d$. Let us consider $\bar{m} = \prod_{i=1}^d m_i$ adjacent rectangles in $[0,1]^d$, B_r, $r = 1, \dots, \bar{m}$, with the side lengths $(h_1, \dots, h_d)$ and centers $c^r = (c_1^r, \dots, c_d^r)$, $c_i^r = h_i(k_{ir} - 1/2)$, $k_{ir} \in \{1, \dots, m_i\}$. For $z \in \mathbb{R}^d$ and $r = 1, \dots, \bar{m}$, define

$$g_r(z) = \Gamma_n \prod_{i=1}^d g((z_i - c_i^r)/h_i),$$

which can be non-zero only on B_r. A set of hypotheses is defined by sequences of binary weights on g_r's as follows

$$q_j(y,x) = \int_{A_y} \left[g_0(\tilde{y},x) + \sum_{r=1}^{\bar{m}} w_r^j g_r(\tilde{y},x) \right] d\tilde{y}, \tag{16}$$

where $w_r^j \in \{0,1\}$, $j = 0,\ldots,M$, and M are defined in Lemma 2, and g_0 satisfies the following conditions: (i) it is a density on $\mathbb{R}^d$, (ii) it is bounded away from zero on $[0,1]^d$, (iii) it belongs to $\mathcal{C}^{\beta_1,\ldots,\beta_d,L/2}$ for some $L \geq 2$. Examples of g_0 include uniform ($g_0 = 1_{[0,1]^d}$), a normal density, and a smoothed to zero uniform that is proportional to

$$\prod_{i=1}^{d} \left[1_{[0,1]}(z_i) + IK_0(z_i+1)\cdot 1(z_i<0) + IK_0(2-z_i)\cdot 1(z_i>1) \right],$$

where $IK_0(z_i) = \int_{-1}^{z_i} K_0(u)du \Big/ \int_{-1}^{1} K_0(u)du$.

The rest of the proof is delegated to lemmas in the supplement, which show that q_k in (16) satisfy the sufficient conditions from Lemma 1. Specifically, Lemma 3 derives the lower bound on the total variation distance. Lemma 4 verifies condition (15) when $\bar{m} \geq 8$. Lemma 5, part (i) of Lemma 7, and the assumptions on g_0 imply that the latent densities in the definition of q_j belong to $\mathcal{C}^{\beta_1,\ldots,\beta_d,L}$, $j = 0,\ldots,M$.

This argument (Lemma 4 specifically) requires $\bar{m} \geq 8$ as it relies on Lemma 2. Observe that as $n \to \infty$, $\bar{m} \geq 8$ if there are continuous variables or there are discrete variables over which smoothing is beneficial ($J_*^c \neq \emptyset$). Thus, $\bar{m} < 8$ can happen only if there are no continuous variables and $N_{J_*} = N_1 \cdots N_d$ is bounded. This is just a problem of estimating a multinomial distribution with finite support and the standard results for parametric problems deliver the usual $n^{-1/2}$ rate.

APPENDIX C: POSTERIOR CONTRACTION RATES FOR UNBOUNDED SUPPORT

C.1. *Assumptions on the Data Generating Process for Unbounded Support*

In what follows, we consider a fixed subset of discrete indices $J \in \mathcal{A}$ and show that under regularity conditions, the posterior contraction rate is bounded above by $\left[\frac{N_J}{n}\right]^{\frac{\beta_{J^c}}{2\beta_{J^c}+1}}$ times a log factor. If the regularity conditions we describe below for a fixed J hold for every subset of $\mathcal{A}$, then the posterior contraction rate matches the lower bound in (6) up to a log factor.

Without a loss of generality, let $J = \{1,\ldots,d_J\}$, $I = \{d_J{+}1,\ldots,d_y\}$, $J^c = \{1,\ldots,d\}\backslash J$, and $d_{J^c} = card(J^c)$. Similarly to $\mathcal{Y}$ and A_y defined in Section 2, we define $\mathcal{Y}_J = \prod_{j\in J}\mathcal{Y}_j$ and $A_{y_J} = \prod_{i\in J} A_{y_i}$. Also, let $y_J = \{y_i\}_{i\in J}$, $\tilde{y}_I = \{\tilde{y}_i\}_{i\in I}$, $\tilde{x} = (\tilde{y}_I, x) \in \tilde{\mathcal{X}} = \mathbb{R}^{d_{J^c}}$.

To formulate the assumptions on the data generating process, we need additional notation,

$$
\begin{aligned}
f_{0J}(y_J, \tilde{x}) &= \int_{A_{y_J}} f_0(\tilde{y}_J, \tilde{x}) d\tilde{y}_J, \\
\pi_{0J}(y_J) &= \int_{\tilde{\mathcal{X}}} f_{0J}(y_J, \tilde{x}) d\tilde{x}, \\
f_{0|J}(\tilde{x}|y_J) &= \frac{f_{0J}(y_J, \tilde{x})}{\pi_{0J}(y_J)}, \\
p_{0|J}(y_I, x|y_J) &= \int_{A_{y_I}} f_{0|J}(\tilde{y}_I, x|y_J) d\tilde{y}_I.
\end{aligned}
$$

Also, let $F_{0|J}$ and $E_{0|J}$ denote the conditional probability and expectation corresponding to $f_{0|J}$. If $\pi_{0J}(y_J) = 0$ for a particular y_J, then we can define the conditional density $f_{0|J}(\tilde{x}|y_J)$ arbitrarily. We make the following assumptions on the data generating process.

ASSUMPTION 1 *There are positive finite constants $b, \bar{f}_0, \tau$ such that for any $y_J \in \mathcal{Y}_J$ and $\tilde{x} \in \tilde{\mathcal{X}}$*

$$
f_{0|J}(\tilde{x}|y_J) \leq \bar{f}_0 \exp\left(-b||\tilde{x}||^\tau\right). \tag{17}
$$

It appears that all the papers on (near) optimal posterior contraction rates for mixtures of normal densities impose similar tail conditions on the data generating densities.

ASSUMPTION 2 *There exists a positive and finite $\bar{y}$ such that for any $(y_I, y_J) \in \mathcal{Y}$ and $x \in \mathcal{X}$*

$$
\int_{A_{y_I} \cap \{||\tilde{y}_I|| \leq \bar{y}\}} f_{0|J}(\tilde{y}_I, x|y_J) d\tilde{y}_I \geq \int_{A_{y_I} \cap \{||\tilde{y}_I|| > \bar{y}\}} f_{0|J}(\tilde{y}_I, x|y_J) d\tilde{y}_I. \tag{18}
$$

This assumption always holds for $A_{y_I} \subset [0,1]^{d_{Jc} - d_x}$. When A_{y_I} is a rectangle with at least one infinite side, an interpretation of this assumption is that the tail probabilities for $\tilde{y}_I$ conditional on (x, y_J) decline uniformly in (x, y_J). Bounded support for $\tilde{y}_I$ is a sufficient condition for this assumption.

ASSUMPTION 3 *We assume that*

$$
f_{0|J} \in \mathcal{C}^{\beta_{d_J+1}, \ldots, \beta_d, L}, \tag{19}
$$

where for some $\tau_0 \geq 0$ and any $(\tilde{x}, \Delta\tilde{x}) \in \mathbb{R}^{2d_{Jc}}$

$$
L(\tilde{x}, \Delta\tilde{x}) = \tilde{L}(\tilde{x}) \exp\left\{\tau_0 ||\Delta\tilde{x}||^2\right\}, \tag{20}
$$

$$\tilde{L}(\tilde{x}+\Delta\tilde{x}) \leq \tilde{L}(\tilde{x}) \exp\left\{\tau_0 ||\Delta\tilde{x}||^2\right\}. \tag{21}$$

The smoothness assumption (19) on the conditional density $f_{0|J}$ is implied by the smoothness of the joint density f_0 at least under boundedness away from zero assumption, see Lemma 10 in Appendix D.2.3. A constant envelop function L used in the lower bound construction would satisfy the assumption.

ASSUMPTION 4 *There are positive finite constants ε and $\bar{F}$, such that for any $y_J \in \mathcal{Y}_J$ and $k = \{k_i\}_{i\in J^c} \in \mathbb{N}_0^{d_{J^c}}$, $\sum_{i\in J^c} k_i/\beta_i < 1$,*

$$\int \left[\frac{|D^k f_{0|J}(\tilde{x}|y_J)|}{f_{0|J}(\tilde{x}|y_J)}\right]^{\frac{(2+\varepsilon\beta_{J^c}^{-1}d_{J^c}^{-1})}{\sum_{i\in J^c} k_i/\beta_i}} f_{0|J}(\tilde{x}|y_J)d\tilde{x} < \bar{F}, \tag{22}$$

$$\int \left[\frac{\tilde{L}(\tilde{x})}{f_{0|J}(\tilde{x}|y_J)}\right]^{2+\varepsilon\beta_{J^c}^{-1}d_{J^c}^{-1}} f_{0|J}(\tilde{x}|y_J)d\tilde{x} < \bar{F}. \tag{23}$$

The envelope function and restrictions on its behaviour are mostly relevant for the case of unbounded support. Condition (23) suggests that the envelope function $\tilde{L}$ should be comparable to $f_{0|J}$.

ASSUMPTION 5 *For some small $\nu > 0$,*

$$N_J = o(n^{1-\nu}). \tag{24}$$

We impose this assumption to exclude from consideration the cases with very slow (non-polynomial) rates as some parts of the proof require $\log(1/\epsilon_n)$ to be of order $\log n$.

C.2. *Posterior Contraction Rates for Unbounded Support*

Let us define a constant that determines the power of the $\log n$ term in the upper bound on the posterior contraction rate derived below in Theorem 3,

$$t_{J0} = \begin{cases} \frac{d_{J^c}[1+1/(\beta_{J^c}d_{J^c})+1/\tau]+\max\{\tau_1,1,\tau_2/\tau\}}{2+1/\beta_{J^c}} & \text{if } J^c \neq \emptyset \\ \max\{\tau_1,1\}/2 & \text{if } J^c = \emptyset \end{cases} \tag{25}$$

where (τ, τ_1, τ_2) are defined in Sections 2.1, 4.3.1, and C.1.

THEOREM 3 *Suppose the assumptions from Sections 4.3.1 and C.1 hold for a given* $J \in \mathcal{A}$. *Let*

$$\epsilon_n = \left[\frac{N_J}{n}\right]^{\beta_{J^c}/(2\beta_{J^c}+1)} (\log n)^{t_J}, \tag{26}$$

where $t_J > t_{J0} + \max\{0, (1-\tau_1)/2\}$. *Suppose also* $n\epsilon_n^2 \to \infty$. *Then, there exists* $\bar{M} > 0$ *such that*

$$\Pi\left(p : d_{TV}(p, p_0) > \bar{M}\epsilon_n | Y^n, X^n\right) \overset{P_0^n}{\to} 0.$$

As in Section 4.2, when $J^c = \emptyset$, β_{J^c} can be defined to be infinity and $\beta_{J^c}/(2\beta_{J^c}+1) = 1/2$ in (26). Note that in the bounded support case, τ can be chosen arbitrarily large and a simplified expression in Theorem 2 can be used instead of t_{J0} in the lower bound on t_J.

COROLLARY 1 *Suppose the assumptions from Sections 4.3.1 and C.1 hold for every* $J \in \mathcal{A}$. *Let*

$$\epsilon_n = \min_{J \in \mathcal{A}} \left[\frac{N_J}{n}\right]^{\beta_{J^c}/(2\beta_{J^c}+1)} (\log n)^{t_J}, \tag{27}$$

where $t_J > t_{J0} + \max\{0, (1-\tau_1)/2\}$. *Suppose also* $n\epsilon_n^2 \to \infty$. *Then, there exists* $\bar{M} > 0$ *such that*

$$\Pi\left(p : d_{TV}(p, p_0) > \bar{M}\epsilon_n | Y^n, X^n\right) \overset{P_0^n}{\to} 0.$$

Under the assumptions of the corollary, Theorem 3 delivers a valid upper bound on the posterior contraction rate for every $J \in \mathcal{A}$ including the one for which the minimum in (27) is attained. Hence, the corollary is an immediate implication of Theorem 3. The proof of Theorem 3 is presented below.

C.3. *Proof Outline for Posterior Contraction Results*

To prove Theorem 3, we use the following sufficient conditions for posterior contraction from Theorem 2.1 in Ghosal and van der Vaart (2001). Let ϵ_n and $\tilde{\epsilon}_n$ be positive sequences with $\tilde{\epsilon}_n \leq \epsilon_n$, $\epsilon_n \to 0$, and $n\tilde{\epsilon}_n^2 \to \infty$, and c_1, c_2, c_3, and c_4 be some positive constants. Let ρ be Hellinger or total variation distance. Suppose $\mathcal{F}_n \subset \mathcal{F}$ is a sieve with the following bound on the metric entropy $M_e(\epsilon_n, \mathcal{F}_n, \rho)$

$$\log M_e(\epsilon_n, \mathcal{F}_n, \rho) \leq c_1 n\epsilon_n^2, \tag{28}$$

$$\Pi(\mathcal{F}_n^c) \leq c_3 \exp\{-(c_2+4)n\tilde{\epsilon}_n^2\}. \tag{29}$$

Suppose also that the prior thickness condition holds

$$\Pi(\mathcal{K}(p_0, \tilde{\epsilon}_n)) \geq c_4 \exp\{-c_2 n \tilde{\epsilon}_n^2\}, \tag{30}$$

where the generalized Kullback-Leibler neighborhood $\mathcal{K}(p_0, \tilde{\epsilon}_n)$ is defined by

$$\mathcal{K}(p_0, \epsilon) = \Bigg\{ p : \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} p_0(y, x) \log \frac{p_0(y, x)}{p(y, x)} dx < \epsilon^2, \\ \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} p_0(y, x) \left[\log \frac{p_0(y, x)}{p(y, x)} \right]^2 dx < \epsilon^2 \Bigg\}.$$

Then, there exists $\bar{M} > 0$ such that

$$\Pi\left(p : \rho(p, p_0) > \bar{M} \epsilon_n | Y^n, X^n\right) \overset{P_0^n}{\to} 0.$$

The definition of the sieve and a verification of conditions (28) and (29) closely follow analogous results in the literature on contraction rates for mixture models in the context of density estimation. The details are given in Lemma 20 in the supplement. Verification of the prior thickness condition is more involved and we formulate it as a separate result in the following theorem.

THEOREM 4 *Suppose the assumptions from Sections 4.3.1 and C.1 hold for a given $J \in \mathcal{A}$. Let $t_J > t_{J0}$, where t_{J0} is defined in (25), and*

$$\tilde{\epsilon}_n = \left[\frac{N_J}{n} \right]^{\beta_{J^c}/(2\beta_{J^c}+1)} (\log n)^{t_J}. \tag{31}$$

For any $C > 0$ and all sufficiently large n,

$$\Pi(\mathcal{K}(p_0, \tilde{\epsilon}_n)) \geq \exp\{-C n \tilde{\epsilon}_n^2\}. \tag{32}$$

Approximation results are key for showing the prior thickness condition (32). Appropriate approximation results for $f_{0J}(y_J, \tilde{x}) = f_{0|J}(\tilde{x}|y_J)\pi_{0J}(y_J)$ are obtained as follows. Based on approximation results for continuous densities by normal mixtures from Shen et al. (2013), we obtain approximations for $f_{0|J}(\cdot|y_J)$ for every y_J in the form

$$f^\star_{|J}(\tilde{x}|y_J) = \sum_{j=1}^{K} \alpha^\star_{j|y_J} \phi(\tilde{x}; \mu^\star_{j|y_J}, \sigma^\star_{J^c}), \tag{33}$$

where the parameters of the mixture will be defined precisely below. For the discrete variables over which smoothing is not performed, y_J, we show that $\pi_{0J}(y_J)$ can be appropriately approximated by

$$\int_{A_{y_J}} \sum_{y'_J} \pi_{0J}(y'_J) \phi(\tilde{y}_J; y'_J, \sigma^\star_J) d\tilde{y}_J,$$

where $\int_{A_{y_J}} \phi(\tilde{y}_J, y'_J, \sigma^\star_J) d\tilde{y}_J$ behaves like an indicator $1\{y_J = y'_J\}$ for sufficiently small $\sigma^\star_J$. Section D.2 in the supplement presents proof details.

APPENDIX D: SUPPLEMENT

D.1. *Proofs and Auxiliary Results for Lower Bounds*

LEMMA 3 *For q_j, q_l, $i \neq l$ defined in (16), the total variation distance is bounded below by const $\cdot \Gamma_n$.*

PROOF: Let us establish several facts about g_r in the definition of q_j. For any $(\tilde{y}, x) \in [0,1]^d$, there exists $r(\tilde{y}, x)$ such that

$$g_r(\tilde{y}, x) = 0, \forall r \neq r(\tilde{y}, x). \tag{34}$$

For $(\tilde{y}, x) \in B_r$, $r(\tilde{y}, x) = r$ and for $(\tilde{y}, x) \notin \cup_{r=1}^{\bar{m}} B_r$, $r(\tilde{y}, x)$ can have an arbitrary value. Thus,

$$\begin{aligned} d_{TV}(q_j, q_l) &= \sum_y \int \left| \int_{A_y} \left[\sum_{r=1}^{\bar{m}} (w_r^j - w_r^l) g_r(\tilde{y}, x) \right] d\tilde{y} \right| dx \\ &= \sum_y \int \left| \int_{A_y} (w^j_{r(\tilde{y},x)} - w^l_{r(\tilde{y},x)}) g_{r(\tilde{y},x)}(\tilde{y}, x) d\tilde{y} \right| dx. \end{aligned}$$

From $h_i = (2/N_i) \cdot R_i$ for $i \in \{1, \ldots, d_y\}$, where R_i is a positive integer, and the definitions of g, g_r, and A_y, it follows that for fixed $y \in \mathcal{Y}$ and $x \in [0,1]^{d_x}$, $(w^j_{r(\tilde{y},x)} - w^l_{r(\tilde{y},x)}) g_{r(\tilde{y},x)}(\tilde{y}, x)$ does not change the sign as $\tilde{y}$ changes within A_y ($r(\tilde{y}, x)$ is the same $\forall \tilde{y} \in A_y$ by the choice of c_i^r and h_i). Therefore,

$$\begin{aligned} d_{TV}(q_j, q_l) &= \int \int \left| (w^j_{r(\tilde{y},x)} - w^l_{r(\tilde{y},x)}) g_{r(\tilde{y},x)}(\tilde{y}, x) \right| d\tilde{y} dx \\ &= \sum_{r=1}^{\bar{m}} \int_{B_r} \left| (w^j_{r(z)} - w^l_{r(z)}) g_{r(z)}(z) \right| dz \\ &= \sum_{r=1}^{\bar{m}} |w_r^j - w_r^l| \int_{B_r} |g_r(z)| \, dz. \end{aligned} \tag{35}$$

Finally, using a change of variables in (35), Lemma 2, and $m_i h_i > 1/2$, we get

$$\begin{aligned} d_{TV}(q_j, q_l) &= \sum_{r=1}^{\bar{m}} 1\{w_r^j \neq w_r^l\} \cdot \Gamma_n \cdot \prod_{i=1}^{d} h_i \cdot \left[\int_{-1/2}^{1/2} |g(u)| du \right]^d \\ &\geq \Gamma_n \cdot \prod_{i=1}^{d} m_i h_i \cdot \left[\int_{-1/2}^{1/2} |g(u)| du \right]^d / 8 \\ &\geq \Gamma_n \cdot \left[\int_{-1/2}^{1/2} |g(u)| du / 2 \right]^d / 8. \end{aligned}$$

Q.E.D.

LEMMA 4 *For $\Gamma_n \to 0$ and $\bar{m} \geq 8$ and a sufficiently small c_0 in the definition of g, condition (15) in Lemma (1) holds for all sufficiently large n.*

PROOF: By Lemma 2, it suffices to show that

$$d_{KL}(Q_j^n, Q_0^n) = n \cdot d_{KL}(q_j, q_0) < (\bar{m} \log 2)/64. \tag{36}$$

First, note that for any $z \in [0,1]^d$, the density in the definition of q_j

$$g_0(z) + \sum_{r=1}^{\bar{m}} w_r^j g_r(z) \geq \underline{g}_0 - \Gamma_n \left[\max_{u \in [-1/2, 1/2]} g(u) \right]^d \geq \underline{g}_0/2 > 0 \tag{37}$$

for all sufficiently large n, where $\underline{g}_0 = \min_{z \in [0,1]^d} g_0(z) > 0$ by the assumption on g_0.

By (47) in Lemma 6 and non-negativity of the Kullback-Leibler divergence

$$\begin{aligned} d_{KL}(q_j, q_0) &\leq d_{KL}\left(g_0 + \sum_{r=1}^{\bar{m}} w_r^j g_r, g_0 \right) \\ &\leq d_{KL}\left(g_0 + \sum_{r=1}^{\bar{m}} w_r^j g_r, g_0 \right) + d_{KL}\left(g_0, g_0 + \sum_{r=1}^{\bar{m}} w_r^j g_r \right) \\ &= \int_{\mathbb{R}^d} \log\left(g_0(z) + \sum_{r=1}^{\bar{m}} w_r^j g_r(z) \right) \left(\sum_{r=1}^{\bar{m}} w_r^j g_r(z) \right) dz \\ &= \int_{[0,1]^d} \log\left(g_0(z) + \sum_{r=1}^{\bar{m}} w_r^j g_r(z) \right) \left(\sum_{r=1}^{\bar{m}} w_r^j g_r(z) \right) dz, \end{aligned} \tag{38}$$

where the last equality follows from $g_r(z) = 0$ outside $[0,1]^d$. The integrand of the last integral is bounded above by $2\underline{g}_0^{-1} \left(\sum_{r=1}^{\bar{m}} w_r^j g_r(z) \right)^2$, which follows from the logarithm inequality, $1 - 1/u \leq \log u \leq u - 1$, $\forall u > 0$, and (37). Thus,

$$\begin{aligned} d_{KL}(q_j, q_0) &\leq 2\underline{g}_0^{-1} \int \left[\sum_{r=1}^{\bar{m}} w_r^j g_r(z) \right]^2 dz \\ &= 2\underline{g}_0^{-1} \int \sum_{r=1}^{\bar{m}} w_r^j (g_r(z))^2 dz \\ &\leq 2\underline{g}_0^{-1} \bar{m} \int (g_1(z))^2 dz = 2\underline{g}_0^{-1} \Gamma_n^2 \prod_i (m_i h_i) \left[\int_{-1/2}^{1/2} g(u)^2 du \right]^d \\ &\leq 2\underline{g}_0^{-1} \Gamma_n^2 \left[\int_{-1/2}^{1/2} g(u)^2 du \right]^d \leq 2\underline{g}_0^{-1} \Gamma_n^2 c_0^{2d}, \end{aligned} \tag{39}$$

where the first equality holds since $g_r(z) g_l(z) = 0, \forall r \neq l$. Finally,

$$\bar{m} = \prod_{i=1}^d m_i \geq 2^{-d} \prod_{i=1}^d h_i^{-1}$$

$$
\begin{aligned}
&= 2^{-d}\prod_{i\in J_*}(N_i/2)\cdot\prod_{i\in J_*^c, i\le d_y}\left(\Gamma_n^{-\beta_i^{-1}}/\varrho_i\right)\cdot\prod_{i\in J_*^c, i>d_y}\left(\Gamma_n^{-\beta_i^{-1}}\right)\\
&\ge 2^{-d}\prod_{i\in J_*}(N_i/2)\cdot\prod_{i\in J_*^c, i\le d_y}\left(\Gamma_n^{-\beta_i^{-1}}/2\right)\cdot\prod_{i\in J_*^c, i>d_y}\left(\Gamma_n^{-\beta_i^{-1}}\right)\\
&= 2^{-d-d_y}\cdot N_{J_*}\cdot\Gamma_n^{-\beta_{J_*^c}^{-1}} = 2^{-d-d_y}n\Gamma_n^2\\
&\ge 2^{-d-d_y}n\cdot d_{KL}(q_j,q_0)/(2\underline{g}_0^{-1}c_0^{2d}),
\end{aligned}
$$

where the first inequality holds by definitions of $\bar{m}$ and m_i, the second equality by definition of h_i, the second inequality by restrictions on ϱ_i, and the last inequality by (39). The last inequality implies (36) if

$$c_0 \le [\underline{g}_0 2^{-(d+d_y+7)}\log 2]^{1/(2d)}.$$

Q.E.D.

LEMMA 5 *For $j \in \{1,\dots,M\}$, a part of the density in the definition of q_j, $f_j = \sum_{r=1}^{\bar{m}} w_r^j g_r \in \mathcal{C}^{\beta_1^*,\dots,\beta_d^*,L}$ with $L = 1$ for any sufficiently small constant c_0 in the definition of g.*

PROOF: Consider $k = (k_1,\dots,k_d)$ and $z, \Delta z \in \mathbb{R}^d$ such that for some $i \in \{1,\dots,d\}$, $\Delta z_i \neq 0$, for any $l \neq i$, $\Delta z_l = 0$, $\sum_{l=1}^d k_l/\beta_l^* < 1$, and $\sum_{l=1}^d k_l/\beta_l^* + 1/\beta_i^* \ge 1$ so that

$$0 \le \beta_i^*(1 - \sum_{l=1}^d k_l/\beta_l^*) \le 1. \tag{40}$$

For $r(\cdot)$ defined in (34),

$$
\begin{aligned}
D^k f_j(z) &= w_{r(z)}\Gamma_n\prod_{l=1}^d g^{(k_l)}((z_l - c_l^{r(z)})/h_l)/h_l^{k_i}\\
&= B_i\cdot w_{r(z)} h_i^{\beta_i^*(1-\sum_{l=1}^d k_l/\beta_l^*)}\prod_{l=1}^d g^{(k_l)}((z_l - c_l^{r(z)})/h_l), \qquad (41)
\end{aligned}
$$

where $B_i \in \{1, 1/2, \varrho_i^{-\beta_i^*}\} \subset (0,1]$. From Tsybakov (2008), (2.33)-(2.34), for any sufficiently small c_0 and $s \le \max_l \beta_l^* + 1$,

$$\max_z |g^{(s)}(z)| \le 1/8. \tag{42}$$

This imply that

$$|g^{(k_i)}((z_i + \Delta z_i - c_i^r)/h_i) - g^{(k_i)}((z_i - c_i^r)/h_i)| \le |\Delta z_i|/(8h_i). \tag{43}$$

First, let us consider the case when $r(z) = r(z + \Delta z)$ and $|\Delta z_i| \le h_i$. From (41), (42), and (43),

$$
\begin{aligned}
|D^k f_j(z + \Delta z) - D^k f_j(z)| &\le h_i^{\beta_i^*(1-\sum_{l=1}^d k_l/\beta_l^*)} 8^{-d} |\Delta z_i / h_i| \\
&= 8^{-d} |\Delta z_i|^{\beta_i^*(1-\sum_{l=1}^d k_l/\beta_l^*)} \left| \frac{\Delta z_i}{h_i} \right|^{1-\beta_i^*(1-\sum_{l=1}^d k_l/\beta_l^*)} \\
&\le |\Delta z_i|^{\beta_i^*(1-\sum_{l=1}^d k_l/\beta_l^*)}, \qquad (44)
\end{aligned}
$$

where the last inequality follows from $\Delta z_i \le h_i$ and (40).

Second, consider the case when $r(z) = r(z + \Delta z)$ and $|\Delta z_i| > h_i$. Similarly to the previous case but without using (43),

$$
|D^k f_j(z + \Delta z) - D^k f_j(z)| \le 2 \cdot 8^{-d} h_i^{\beta_i^*(1-\sum_{l=1}^d k_l/\beta_l^*)} \le |\Delta z_i|^{\beta_i^*(1-\sum_{l=1}^d k_l/\beta_l^*)}.
$$

Third, consider the case when $r(z) \ne r(z+\Delta z)$ and $|\Delta z_i| \le h_i/2$. If $w_{r(z)} = w_{r(z+\Delta z)} = 0$ or $z, z + \Delta z \notin \cup_{r=1}^{\bar m} B_r$

$$
|D^k f_j(z + \Delta z) - D^k f_j(z)| = D^k f_j(z + \Delta z) = D^k f_j(z) = 0.
$$

If $w_{r(z)} \ne w_{r(z+\Delta z)}$ or if one of z and $z + \Delta z$ is not in $\cup_{r=1}^{\bar m} B_r$, then without a loss of generality suppose that $w_{r(z)} = 1$ or that $z + \Delta z \notin \cup_{r=1}^{\bar m} B_r$. Let $|\Delta z_i^\star| \in [0, |\Delta z_i|]$ and $\Delta z^\star = (0, \dots, 0, \Delta z_i^\star, 0, \dots, 0)$ be such that $z + \Delta z^\star$ is a boundary point of $B_{r(z)}$. Then, $D^k f_j(z + \Delta z^\star) = 0$ and (44) imply

$$
\begin{aligned}
|D^k f_j(z + \Delta z) - D^k f_j(z)| &= |D^k f_j(z)| = |D^k f_j(z + \Delta z^\star) - D^k f_j(z)| \\
&\le |\Delta z_i^\star|^{\beta_i^*(1-\sum_{l=1}^d k_l/\beta_l^*)} \le |\Delta z_i|^{\beta_i^*(1-\sum_{l=1}^d k_l/\beta_l^*)}.
\end{aligned}
$$

If $w_{r(z)} = w_{r(z+\Delta z)} = 1$ and $z, z + \Delta z \in \cup_{r=1}^{\bar m} B_r$ then by construction of f_j and g

$$
\begin{aligned}
&|D^k f_j(z + \Delta z) - D^k f_j(z)| \qquad = |D^k f_j(z + \Delta z + 0.5 h_i) - D^k f_j(z + 0.5 h_i)| \\
&\le |\Delta z_i|^{\beta_i^*(1-\sum_{l=1}^d k_l/\beta_l^*)},
\end{aligned}
$$

where the last inequality follows from (44).

Finally, when $r(z) \ne r(z + \Delta z)$ and $\Delta z_i > h_i/2$,

$$
\begin{aligned}
&|D^k f_j(z + \Delta z) - D^k f_j(z)| \le |D^k f_j(z + \Delta z)| + |D^k f_j(z)| \\
&\le 2 \cdot 8^{-d} h_i^{\beta_i^*(1-\sum_{l=1}^d k_l/\beta_l^*)} \\
&\le |\Delta z_i|^{\beta_i^*(1-\sum_{l=1}^d k_l/\beta_l^*)}.
\end{aligned}
$$

Now, let us consider a general Δz such that for $\Delta z_i \neq 0$, $\sum_{l=1}^{d} k_l/\beta_l^* + 1/\beta_i^* \geq 1$.

$$
\begin{aligned}
&|D^k f_j(z+\Delta z) - D^k f_j(z)| \\
&\leq \sum_{i=1}^{d} \big| D^k f_j(z_1, \ldots, z_{i-1}, z_i + \Delta z_i, \ldots, z_d + \Delta z_d) \\
&\qquad - D^k f_j(z_1, \ldots, z_i, z_{i+1} + \Delta z_{i+1}, \ldots, z_d + \Delta z_d) \big|.
\end{aligned}
$$

The preceding argument applies to every term in this sum and, thus, $f_j \in \mathcal{C}^{\beta_1^*, \ldots, \beta_d^*, 1}$.

Q.E.D.

LEMMA 6 *Let $f_i : \tilde{\mathcal{Y}} \times \mathcal{X} \to \mathbb{R}$, $i \in \{1, 2\}$, be densities with respect to a product measure $\lambda \times \mu$ on $\tilde{\mathcal{Y}} \times \mathcal{X} \subset \mathbb{R}^d$. For a finite set $\mathcal{Y}$, let $\{A_y,\, y \in \mathcal{Y}\}$ be a partition of $\tilde{\mathcal{Y}}$ and let $p_i(y, x) = \int_{A_y} f_i(\tilde{y}, x) d\lambda(\tilde{y})$. Then,*

$$d_{TV}(p_1, p_2) \leq d_{TV}(f_1, f_2) \tag{45}$$

$$d_H(p_1, p_2) \leq d_H(f_1, f_2) \tag{46}$$

$$d_{KL}(p_1, p_2) \leq d_{KL}(f_1, f_2). \tag{47}$$

Also, if for given (y, x), $f_2(\tilde{y}, x) > 0$ for any $\tilde{y} \in A_y$, then

$$\inf_{\tilde{y} \in A_y} \frac{f_1(\tilde{y}, x)}{f_2(\tilde{y}, x)} \leq \frac{p_1(y, x)}{p_2(y, x)} \leq \sup_{\tilde{y} \in A_y} \frac{f_1(\tilde{y}, x)}{f_2(\tilde{y}, x)}. \tag{48}$$

PROOF: Trivially,

$$
\begin{aligned}
d_{TV}(p_1, p_2) &= \sum_y \int \left| \int_{A_y} (f_1(\tilde{y}, x) - f_2(\tilde{y}, x)) d\tilde{y} \right| d\mu(x) \\
&\leq \sum_y \int \int_{A_y} |f_1(\tilde{y}, x) - f_2(\tilde{y}, x)| d\lambda(\tilde{y}) d\mu(x) = d_{TV}(f_1, f_2).
\end{aligned}
$$

By Holder inequality,

$$
\begin{aligned}
d_H(p_1, p_2) &= 2 \left(1 - \sum_y \int \sqrt{\int 1_{A_y}(\tilde{y}_1) f_1(\tilde{y}_1, x) d\lambda(\tilde{y}_1) \cdot \int 1_{A_y}(\tilde{y}_2) f_2(\tilde{y}_2, x) d\lambda(\tilde{y}_2)} d\mu(x) \right) \\
&\leq 2 \left(1 - \sum_y \int \int 1_{A_y}(\tilde{y}) \sqrt{f_1(\tilde{y}, x) f_2(\tilde{y}, x)} d\lambda(\tilde{y}) d\mu(x) \right) = d_H(f_1, f_2).
\end{aligned}
$$

For fixed (y, x),

$$\int_{A_y} (f_1(\tilde{y}, x)/p_1(y, x)) \log \frac{f_1(\tilde{y}, x)/p_1(y, x)}{f_2(\tilde{y}, x)/p_2(y, x)} d\lambda(\tilde{y}) \geq 0$$

since the Kullback-Leibler divergence is nonnegative. Thus,

$$\int_{A_y} f_1(\tilde{y},x)\log\frac{f_1(\tilde{y},x)}{f_2(\tilde{y},x)}d\lambda(\tilde{y}) \geq \int_{A_y} f_1(\tilde{y},x)\log\frac{p_1(y,x)}{p_2(y,x)}d\lambda(\tilde{y}) = p_1(y,x)\log\frac{p_1(y,x)}{p_2(y,x)}.$$

This inequality integrated with respect to $d\mu(x)$ and summed over y implies (47). The last claim follows from

$$f_2(\tilde{y},x)\inf_{\tilde{z}\in A_y}\frac{f_1(\tilde{z},x)}{f_2(\tilde{z},x)} \leq f_1(\tilde{y},x) \leq f_2(\tilde{y},x)\sup_{\tilde{z}\in A_y}\frac{f_1(\tilde{z},x)}{f_2(\tilde{z},x)}.$$

Q.E.D.

LEMMA 7 *For Γ_n, h_i, ϱ_i, and β_i^* defined in Section 4.2, (i) $\beta_i^* \geq \beta_i$ for $i = 1,\ldots,d$ and (ii) $\varrho_i \in (1,2]$ for $i \in J_*^c \cap \{1,\ldots,d_y\}$.*

PROOF: For $i \notin J_*$, $\beta_i^* = \beta_i$ by definition. For $i \in J_*$, from the definition of Γ_n,

$$\Gamma_n \leq \left[\frac{N_{J_*}/N_i}{n}\right]^{\frac{1}{2+\beta_{J_*^c}^{-1}+\beta_i^{-1}}} = \Gamma_n^{\frac{2+\beta_{J_*^c}^{-1}}{2+\beta_{J_*^c}^{-1}+\beta_i^{-1}}} N_i^{\frac{-1}{2+\beta_{J_*^c}^{-1}+\beta_i^{-1}}},$$

which implies $N_i^{-\beta_i} \geq \Gamma_n$. By the definition of β_i^*, $N_i^{-\beta_i^*} = \Gamma_n$ and, thus, $\beta_i^* \geq \beta_i$.

For $i \in J_*^c$, from the definition of Γ_n,

$$\left[\frac{N_{J_*}N_i}{n}\right]^{\frac{1}{2+\beta_{J_*^c}^{-1}-\beta_i^{-1}}} \geq \left[\frac{N_{J_*}}{n}\right]^{\frac{1}{2+\beta_{J_*^c}^{-1}}},$$

which implies

$$N_i \geq \left[\frac{N_{J_*}}{n}\right]^{\frac{2+\beta_{J_*^c}^{-1}-\beta_i^{-1}}{2+\beta_{J_*^c}^{-1}}} = \Gamma_n^{-\beta_i^{-1}} \implies$$
$$\Gamma_n^{\beta_i^{-1}} \geq \frac{1}{N_i},$$

and, therefore, $\Gamma_n^{\beta_i^{-1}} N_i \geq 1$. Next, define

$$\varrho_i = \frac{\left\lfloor \Gamma_n^{\beta_i^{-1}} N_i/2 \right\rfloor + 1}{\Gamma_n^{\beta_i^{-1}} N_i/2}.$$

Then $\varrho_i \in (1,2]$ as $\Gamma_n^{\beta_i^{-1}} N_i \geq 1$. *Q.E.D.*

D.2. *Proofs of Posterior Contraction Results*

D.2.1. *Proof of Theorem 4 for $J^c \neq \emptyset$*

Define $\beta = d_{J^c}\left[\sum_{k\in J^c}\beta_k^{-1}\right]^{-1}$, $\beta_{\min} = \min_{j\in J^c}\beta_j$, and $\sigma_n = [\tilde{\epsilon}_n/\log(1/\tilde{\epsilon}_n)]^{1/\beta}$. For ε defined in (22)-(23), b and τ defined in (17), and a sufficiently small $\delta > 0$, let $a_0 = \{(8\beta + 4\varepsilon + 8 + 8\beta/\beta_{\min})/(b\delta)\}^{1/\tau}$, $a_{\sigma_n} = a_0\{\log(1/\sigma_n)\}^{1/\tau}$, and $b_1 > \max\{1, 1/2\beta\}$ satisfying $\tilde{\epsilon}_n^{b_1}\{\log(1/\tilde{\epsilon}_n)\}^{5/4} \leq \tilde{\epsilon}_n$. Then, the proofs of Theorems 4 and 6 in Shen et al. (2013) imply the following two claims for each $y_J = k \in \mathcal{Y}_J$ under the assumptions of Section C.1.

First, there exists a partition $\{U_{j|k}, j = 1, \ldots, K\}$ of $\{\tilde{x} \in \tilde{\mathcal{X}} : ||\tilde{x}|| \leq 2a_{\sigma_n}\}$, such that for $j = 1, \ldots, N$, $U_{j|k}$ is contained within an ellipsoid with center $\mu^\star_{j|k}$ and radii $\{\sigma_n^{\beta/\beta_i}\tilde{\epsilon}_n^{2b_1}, i \in J^c\}$

$$U_{j|k} \subset \left\{\tilde{x} : \sum_{i=1}^{d_{J^c}}\left[(\tilde{x}_i - \mu^\star_{j|k,i})/(\sigma_n^{\beta/\beta_{d_J+i}}\tilde{\epsilon}_n^{2b_1})\right]^2 \leq 1\right\};$$

for $j = N+1, \ldots, K$, $U_{j|k}$ is contained within an ellipsoid with radii $\{\sigma_n^{\beta/\beta_i}, i \in J^c\}$, and $1 \leq N < K \leq C_1\sigma_n^{-d_{J^c}}\{\log(1/\tilde{\epsilon}_n)\}^{d_{J^c}+d_{J^c}/\tau}$, where $C_1 > 0$ does not depend on n and y_J.

Second, for each $k \in \mathcal{Y}_J$ there exist $\alpha^\star_{j|k}$, $j = 1, \ldots, K$, with $\alpha^\star_{j|k} = 0$ for $j > N$, and $\mu^{x\star}_{jk} \in U_{j|k}$ for $j = N+1, \ldots, K$ such that for a positive constant C_2 and $\sigma^\star_{J^c} = \{\sigma_n^{\beta/\beta_i}$ for $i \in J^c\}$,

$$d_H\left(f_{0|J}(\cdot|k), f^\star_{|J}(\cdot|k)\right) \leq C_2\sigma_n^\beta, \tag{49}$$

where $f^\star_{|J}$ is defined in (33). Constant C_2 is the same for all $k \in \mathcal{Y}_J$ since all the bounds on $f_{0|J}$ assumed in Section C.1 are uniform over k.

Note also that our smoothness definition is different from the one used by Shen et al. (2013). In Lemmas 8 and 9 we show that our smoothness definition ($f_{0|J} \in \mathcal{C}^{L,\beta_{d_J+1},\ldots,\beta_d}$) delivers an anisotropic Taylor expansion with bounds on remainder terms such that the argument on p. 637 of Shen et al. (2013) goes through.

Third, by Lemma 12, which is an extension of a part of Proposition 1 in Shen et al. (2013), there exists a constant $B_0 > 0$ such that for all $y_J \in \mathcal{Y}_J$

$$F_{0|J}\left(||\tilde{X}|| > a_{\sigma_n}|y_J\right) \leq B_0\sigma_n^{4\beta+2\varepsilon}\underline{\sigma}_n^8, \tag{50}$$

where

$$\underline{\sigma}_n = \min_{i\in J^c}\sigma_n^{\beta/\beta_i}.$$

For $m = N_J K$ we define $\theta^\star$ and $S_{\theta^\star}$ as:

$$\theta^\star = \Big\{ \{\mu_1^\star, \ldots, \mu_m^\star\} = \{(k, \mu_{j|k}^\star),\ j = 1, \ldots, K,\ k \in \mathcal{Y}_J\},$$
$$\{\alpha_1^\star, \ldots, \alpha_m^\star\} = \{\alpha_{jk}^\star = \alpha_{j|k}^\star \pi_{0J}(k),\ j = 1, \ldots, K,\ k \in \mathcal{Y}_J\},$$
$$\sigma_J^{\star 2} = \{\sigma_i^{\star 2} = 1/[64 N_i^2 \beta \log(1/\sigma_n)],\ i \in J\}$$
$$\sigma_{J^c}^\star = \{\sigma_i^\star = \sigma_n^{\beta/\beta_i},\ i \in J^c\}, \Big\}$$

$$S_{\theta^\star} = \Big\{ \{\mu_1, \ldots, \mu_m\} = \{(\mu_{jk,J}, \mu_{jk,J^c}),\ j = 1, \ldots, K,\ k \in \mathcal{Y}_J\},$$
$$\mu_{jk,J^c} \in U_{j|k}, \quad \mu_{jk,i} \in \left[k_i - \frac{1}{4N_i}, k_i + \frac{1}{4N_i}\right],\ i \in J,$$
$$\sigma_i^2 \in \left(0, \sigma_i^{\star 2}\right),\ i \in J,$$
$$\sigma_i^2 \in \left(\sigma_i^{\star 2}(1 + \sigma_n^{2\beta})^{-1}, \sigma_i^{\star 2}\right),\ i \in J^c,$$
$$(\alpha_1, \ldots, \alpha_m) = \{\alpha_{jk},\ j = 1, \ldots, K,\ k \in \mathcal{Y}_J\} \in \Delta^{m-1},$$
$$\sum_{r=1}^m |\alpha_r - \alpha_r^\star| \le 2\sigma_n^{2\beta}, \quad \min_{j \le K, k \in \mathcal{Y}_J} \alpha_{jk} \ge \frac{\sigma_n^{2\beta + d_{J^c}}}{2m^2} \Big\},$$

where Δ^{m-1} denotes the m-dimensional simplex.

The rest of the proof of the Kullback-Leibler thickness condition follows the general argument developed for mixture models in Ghosal and van der Vaart (2007) and Shen et al. (2013) among others. First, we will show that for $m = N_J K$ and $\theta \in S_{\theta^\star}$, the Hellinger distance $d_H^2(p_0(\cdot,\cdot), p(\cdot,\cdot|\theta, m))$ can be bounded by $\sigma_n^{2\beta}$ up to a multiplicative constant. Second, we construct bounds on the ratios $p(\cdot,\cdot|\theta, m)/p_0(\cdot,\cdot)$ and combine them with the bound on the Hellinger distance using Lemma 11. Finally, we will show that the prior puts sufficient probability on $m = N_J K$ and $S_{\theta^\star}$.

For $f_{|J}^\star$ defined in (33), let us define

$$p_{|J}^\star(y_I, x|y_J) = \int_{A_{y_I}} f_{|J}^\star(\tilde{y}_I, x|y_J) d\tilde{y}_I.$$

For $m = N_J K$ and $\theta \in S_{\theta^\star}$, we can bound the Hellinger distance between the DGP and the model as follows,

$$d_H^2(p_0(\cdot,\cdot), p(\cdot,\cdot|\theta, m)) = d_H^2(p_{0|J}(\cdot|\cdot)\pi_0(\cdot), p(\cdot,\cdot|\theta, m))$$
$$\le d_H^2(p_{0|J}(\cdot|\cdot)\pi_{0J}(\cdot), p_{|J}^\star(\cdot|\cdot)\pi_{0J}(\cdot)) + d_H^2(p_{|J}^\star(\cdot|\cdot)\pi_{0J}(\cdot), p(\cdot,\cdot|\theta, m)).$$

It follows from (49) and Lemma 6 linking distances between probability mass functions and corresponding latent variable densities that the first term on the right hand side of this inequality is bounded by $(C_2)^2\sigma_n^{2\beta}$. Combining this result with the bound on $d_H^2(p^\star_{|J}(\cdot|\cdot)\pi_{0J}(\cdot), p(\cdot,\cdot|\theta,m))$ from Lemma 13 we obtain

$$d_H^2(p_0(\cdot,\cdot), p(\cdot,\cdot|\theta,m)) \lesssim \sigma_n^{2\beta}, \tag{51}$$

where "$\lesssim$" denotes less or equal up to a multiplicative positive constant relation.

Next, for $\theta \in S_{\theta^\star}$ and $m = N_J K$, let us consider lower bounds on the ratio $p(y_J, y_I, x|\theta, m)/p_0(y_J, y_I, x)$. In Lemma 16, we show that lower bounds on the ratio $f_J(y_J, \tilde{x}|\theta, m)/f_{0|J}(\tilde{x}|y_J)\pi_0(y_J)$ imply the following bounds for all sufficiently large n: for any $x \in \mathcal{X}$ with $\|x\| \le a_{\sigma_n}$,

$$\frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} \ge C_3 \frac{\sigma_n^{2\beta}}{2m^2} \equiv \lambda_n, \tag{52}$$

for some constant $C_3 > 0$; and for any $x \in \mathcal{X}$ with $\|x\| > a_{\sigma_n}$,

$$\frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} \ge \exp\left\{-\frac{8||x||^2}{\underline{\sigma}_n^2} - C_4 \log n\right\}, \tag{53}$$

for some constant $C_4 > 0$. Consider all sufficiently large n such that $\lambda_n < e^{-1}$ and (52) and (53) hold. Then, for any $\theta \in S_{\theta^\star}$,

$$\begin{aligned}
&\sum_{y\in\mathcal{Y}} \int_{\mathcal{X}} \left(\log \frac{p_0(y_J, y_I, x)}{p(y_J, y_I, x|\theta, m)}\right)^2 \mathbf{1}\left\{\frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} < \lambda_n\right\} p_0(y_J, y_I, x)dx \\
&= \sum_{y\in\mathcal{Y}} \int_{\tilde{\mathcal{X}}} \left(\log \frac{p_0(y_J, y_I, x)}{p(y_J, y_I, x|\theta, m)}\right)^2 \\
&\qquad \mathbf{1}\left\{\frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} < \lambda_n\right\} \mathbf{1}\left\{\tilde{y}_I \in A_{y_I}\right\} f_{0J}(y_J, \tilde{x})d\tilde{x} \\
&= \sum_{y\in\mathcal{Y}} \int_{\tilde{\mathcal{X}}} \left(\log \frac{p_0(y_J, y_I, x)}{p(y_J, y_I, x|\theta, m)}\right)^2 \\
&\qquad \mathbf{1}\left\{\frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} < \lambda_n, ||x|| > a_{\sigma_n}, \tilde{y}_I \in A_{y_I}\right\} f_{0J}(y_J, \tilde{x})d\tilde{x} \\
&\le \sum_{y\in\mathcal{Y}} \int_{\{\tilde{x}:||x||>a_{\sigma_n}\}} \left(\log \frac{p_0(y_J, y_I, x)}{p(y_J, y_I, x|\theta, m)}\right)^2 \mathbf{1}\left\{\tilde{y}_I \in A_{y_I}\right\} f_{0J}(y_J, \tilde{x})d\tilde{x} \\
&\le \sum_{y\in\mathcal{Y}} \int_{\{\tilde{x}:||x||>a_{\sigma_n}\}} \left[\frac{128}{\underline{\sigma}_n^4}||x||^4 + 2(C_4 \log n)^2\right] f_{0|J}(\tilde{x}|y_J)\mathbf{1}\left\{\tilde{y}_I \in A_{y_I}\right\} d\tilde{x}\pi_{0J}(y_J) \\
&\le \sum_{y_J\in\mathcal{Y}_J} \int_{\{\tilde{x}:||\tilde{x}||>a_{\sigma_n}\}} \left[\frac{128}{\underline{\sigma}_n^4}||\tilde{x}||^4 + 2(C_4 \log n)^2\right] f_{0|J}(\tilde{x}|y_J)d\tilde{x}\pi_{0J}(y_J)
\end{aligned} \tag{54}$$

$$\leq \frac{128}{\underline{\sigma}_n^4} \sum_{y_J \in \mathcal{Y}_J} E_{0|y_J} \left(\left\| \tilde{X} \right\|^8 \right)^{1/2} \left(F_{0|y_J} \left(\left\| \tilde{X} \right\| > a_{\sigma_n} \right) \right)^{1/2} \pi_{0J}(y_J)$$
$$+ 2(C_4 \log n)^2 B_0 \sigma_n^{4\beta+2\varepsilon} \underline{\sigma}_n^8$$
$$\leq C_5 \sigma_n^{2\beta+\varepsilon}$$

for some constant $C_5 > 0$ and all sufficiently large n, where the last inequality holds by the tail condition in (17), (50), and $(\log n)^2 \sigma_n^{2\beta+\varepsilon} \underline{\sigma}_n^8 \to 0$.

Furthermore, as $\lambda_n < e^{-1}$,

$$\log \frac{p_0(y_J, y_I, x)}{p(y_J, y_I, x|\theta, m)} \mathbf{1}\left\{ \frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} < \lambda_n \right\}$$
$$\leq \left(\log \frac{p_0(y_J, y_I, x)}{p(y_J, y_I, x|\theta, m)} \right)^2 \mathbf{1}\left\{ \frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} < \lambda_n \right\}$$

and, therefore,

$$\sum_{y \in \mathcal{Y}} \int_{\mathcal{X}} \log \frac{p_0(y_J, y_I, x)}{p(y_J, y_I, x|\theta, m)} \mathbf{1}\left\{ \frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} < \lambda_n \right\} p_0(y_J, y_I, x) dx \tag{55}$$
$$\leq C_5 \sigma_n^{2\beta+\varepsilon}.$$

Inequalities (51), (54), and (55) combined with Lemma 11 imply

$$E_0 \left(\log \frac{p_0(y_J, y_I, x)}{p(y_J, y_I, x|\theta, m)} \right) \leq A\tilde{\epsilon}_n^2, \; E_0 \left(\left[\log \frac{p_0(y_J, y_I, x)}{p(y_J, y_I, x|\theta, m)} \right]^2 \right) \leq A\tilde{\epsilon}_n^2$$

for any $\theta \in S_{\theta^\star}$, $m = N_J K$, and some positive constant A (details are provided in Lemma 17).

By Lemma 18 for all sufficiently large n, $s = 1 + 1/\beta + 1/\tau$, and some $C_6 > 0$,

$$\Pi(\mathcal{K}(p_0, \tilde{\epsilon}_n)) \geq \Pi(m = N_J K, \theta \in S_{\theta^\star})$$
$$\geq \exp\left[-C_6 N_J \tilde{\epsilon}_n^{-d_{J^c}/\beta} \{\log(n)\}^{d_{J^c} s + \max\{\tau_1, 1, \tau_2/\tau\}} \right].$$

The last expression of the above display is bounded below by $\exp\{-Cn\tilde{\epsilon}_n^2\}$ for any $C > 0$, $\tilde{\epsilon}_n = \left[\frac{N_J}{n}\right]^{\beta/(2\beta+d_{J^c})} (\log n)^{t_J}$, any $t_J > (d_{J^c} s + \max\{\tau_1, 1, \tau_2/\tau\})/(2 + d_{J^c}/\beta)$, and all sufficiently large n. Since the inequality in the definition of t_J is strict, the claim of the theorem follows.

When $J = \emptyset$ and $N_J = 1$, the preceding argument delivers the claim of the theorem if we add an artificial discrete coordinate with only one possible value to the vector of observables.

D.2.2. *Proof of Theorem 4 for $J^c = \emptyset$*

In this case, the proof from the previous subsection can be simplified as follows. For $m = N_J$ and for any $\beta > 0$ we define $\theta^\star$ and $S_{\theta^\star}$ as

$$\theta^\star = \Big\{ \{\mu_1^\star, \dots, \mu_m^\star\} = \{k,\ k \in \mathcal{Y}_J\}, \\ \{\alpha_1^\star, \dots, \alpha_m^\star\} = \{\alpha_k^\star,\ k \in \mathcal{Y}_J\} = \{\pi_0(k)\}_{k \in \mathcal{Y}_J}, \\ \sigma^{\star 2} = \{\sigma_i^{\star 2} = \frac{1}{64 N_i^2 \beta \log(1/\sigma_n)},\ i \in J\} \Big\},$$

$$S_{\theta^\star} = \Big\{ \{\mu_1, \dots, \mu_m\} = \{\mu_k,\ k \in \mathcal{Y}_J\},\ \mu_{k,i} \in \Big[k_i - \frac{1}{4N_i}, k_i + \frac{1}{4N_i}\Big],\ i \in J, \\ \sigma = \{\sigma_i \in (0, \sigma_i^\star),\ i \in J\}, \\ \{\alpha_j,\ j = 1, \dots, m\} = \{\alpha_k,\ k \in \mathcal{Y}_J\} \in \Delta^{m-1}, \\ \sum_{k \in \mathcal{Y}_J} |\alpha_k - \alpha_k^\star| \le 2\sigma_n^{2\beta}, \quad \min_{k \in \mathcal{Y}_J} \alpha_k \ge \frac{\sigma_n^{2\beta}}{2m^2} \Big\}.$$

For $m = N_J$ and $\theta \in S_{\theta^\star}$, a simplification of the proof of Lemma 13 delivers

$$d_H^2(p_0(\cdot), p(\cdot|\theta, m)) \le 2 \max_{k \in \mathcal{Y}_J} \int_{A_k^c} \phi(\tilde{y}_J; \mu_k, \sigma) d\tilde{y}_J + \sum_{k \in \mathcal{Y}_J} |\alpha_k^\star - \alpha_k| \lesssim \sigma_n^{2\beta}.$$

A simplification of derivations in Lemma 16 show that for all $y_J \in \mathcal{Y}_J$

$$\frac{p(y_J|\theta, m)}{p_0(y_J)} \ge \frac{1}{2} \frac{\sigma_n^{2\beta}}{2m^2} \equiv \lambda_n.$$

Then, for any $\theta \in S_{\theta^\star}$

$$\sum_{y_J \in \mathcal{Y}_J} \Big(\log \frac{p_0(y_J)}{p(y_J|\theta, m)} \Big)^2 \mathbf{1}\Big\{ \frac{p(y_J|\theta, m)}{p_0(y_J)} < \lambda_n \Big\} p_0(y_J) = 0$$

$$\sum_{y_J \in \mathcal{Y}_J} \Big(\log \frac{p_0(y_J)}{p(y_J|\theta, m)} \Big) \mathbf{1}\Big\{ \frac{p(y_J|\theta, m)}{p_0(y_J)} < \lambda_n \Big\} p_0(y_J) = 0$$

as $\frac{p(y_J|\theta,m)}{p_0(y_J)} \ge \lambda_n$ for all $y_J \in \mathcal{Y}_J$. As $\lambda_n \to 0$, by Lemma 11 for $\lambda_n < \lambda_0$, both $E_0(\log \frac{p_0(y_J)}{p(y_J|\theta,m)})$ and $E_0([\log \frac{p_0(y_J)}{p(y_J|\theta,m)}]^2)$ are bounded by $C_7 \log(1/\lambda_n)^2 \sigma_n^{2\beta} \le A\tilde{\epsilon}_n^2$ for some constant A. By the simplification of Lemma 18 for this particular case for all sufficiently large n and some $C_8 > 0$,

$$\Pi(\mathcal{K}(p_0, \tilde{\epsilon}_n)) \ge \Pi(m = N_J, \theta \in S_{\theta^\star}) \ge \exp\left[-C_8 N_J \{\log(n)\}^{\max\{\tau_1, 1\}} \right].$$

The last expression of the above display is bounded below by $\exp\{-Cn\tilde{\epsilon}_n^2\}$ for any $C > 0$, $\tilde{\epsilon}_n = \left[\frac{N_J}{n}\right]^{1/2} (\log n)^{t_J}$, any $t_J > \max\{\tau_1, 1\}/2$, and all sufficiently large n. Since the inequality in the definition of t_J is strict, the claim of the theorem follows.

D.2.3. *Auxiliary Results for Posterior Contraction Rates*

For a multi-index $k = (k_1, \ldots, k_d) \in \mathbb{Z}_+^d$, let $k! = \prod_{i=1}^d k_i!$, and for $z \in \mathbb{R}^d$, let $z^k = \prod_{i=1}^d z_i^{k_i}$.

LEMMA 8 *(Anisotropic Taylor Expansion) For* $f \in \mathcal{C}^{\beta_1,\ldots,\beta_d,L}$ *and* $r \in \{1, \ldots, d\}$

$$f(x_1 + y_1, \ldots, x_d + y_d) = \sum_{k \in I^r} \frac{y^k}{k!} D^k f(x_1, \ldots, x_r, x_{r+1} + y_{r+1}, \ldots, x_d + y_d) \tag{56}$$

$$+ \sum_{l=1}^{r} \sum_{k \in \bar{I}^l} \frac{y^k}{k!} \Big(D^k f(x_1, \ldots, x_l + \zeta_l^k, x_{l+1} + y_{l+1}, \ldots, x_d + y_d) \tag{57}$$

$$- D^k f(x_1, \ldots, x_l, x_{l+1} + y_{l+1}, \ldots, x_d + y_d) \Big), \tag{58}$$

where $\zeta_l^k \in [x_l, x_l + y_l] \cup [x_l + y_l, x_l]$,

$$I^l = \Big\{ k = (k_1, \ldots, k_l, 0, \ldots, 0) \in \mathbb{Z}_+^d : \; k_i \leq \big\lfloor \beta_i (1 - \sum_{j=1}^{i-1} k_j/\beta_j) \big\rfloor_s, \; i = 1, \ldots, l \Big\},$$

$$\bar{I}^l = \Big\{ k \in I^l : \; k_l = \big\lfloor \beta_l (1 - \sum_{j=1}^{l-1} k_j/\beta_j) \big\rfloor_s \Big\},$$

and the differences in derivatives in (57)-(58) *are bounded by*

$$L \left| \zeta_l^k \right|^{\beta_l (1 - \sum_{i=1}^d k_i/\beta_i)}.$$

PROOF: The lemma is proved by induction. For $r = 1$, (56)-(58) is a standard univariate Taylor expansion of $f(x+y)$ in the first argument around $(x_1, x_2+y_2, \ldots, x_d+y_d)$. Suppose (56)-(58) holds for some $r \in \{1, \ldots, d\}$. Then, let us show that (56)-(58) holds for $r+1$. For that, consider a univariate Taylor expansion of $D^k f$ in (56). The following notation will be useful. Let $e_i \in \mathbb{R}^d$, $i = 1, \ldots, d$, be such that $e_{ij} = 1$ for $i = j$ and $e_{ij} = 0$ for $i \neq j$ and $k_{r+1}^* = \lfloor \beta_{r+1} (1 - \sum_{j=1}^r k_j/\beta_j) \rfloor_s$. Then,

$$D^k f(x_1, \ldots, x_r, x_{r+1} + y_{r+1}, \ldots, x_d + y_d) =$$

$$\sum_{k_{r+1}=0}^{k_{r+1}^*} \frac{y_{r+1}^{k_{r+1}}}{k_{r+1}!} D^{k + k_{r+1} \cdot e_{r+1}} f(x_1, \ldots, x_{r+1}, x_{r+2} + y_{r+2}, \ldots, x_d + y_d)$$

$$+ \frac{y_{r+1}^{k_{r+1}^*}}{k_{r+1}^*!} \Big(D^{k + k_{r+1}^* \cdot e_{r+1}} f(x_1, \ldots, x_r, x_{r+1} + \zeta_{r+1}^{k + k_{r+1}^* \cdot e_{r+1}}, x_{r+2} + y_{l+2}, \ldots, x_d + y_d)$$

$$- D^{k + k_{r+1}^* \cdot e_{r+1}} f(x_1, \ldots, x_r, x_{r+1}, x_{r+2} + y_{l+2}, \ldots, x_d + y_d) \Big).$$

Inserting this expansion into (56) delivers the result for $r+1$.

Q.E.D.

LEMMA 9 *Let $R(x,y)$ denote the remainder term in the anisotropic Taylor expansion ((57)-(58) for $r=d$). Suppose $f \in \mathcal{C}^{\beta_1,\ldots,\beta_d,L}$ and L satisfies (20)-(21). Let $\sigma = \{\sigma_i = \sigma_n^{\beta/\beta_i},\ i=1,\ldots,d\}$ and $\sigma_n \to 0$. Then, for all sufficiently large n,*

$$\int |R(x,y)|\phi(y;0,\sigma)dy \lesssim L(x)\sigma_n^{\beta}.$$

PROOF: Note that $|R(x,y)|$ is bounded by a sum of the following terms over $k \in \bar{I}^l$ and $l \in \{1,\ldots,d\}$

$$\begin{aligned}
&\frac{y^k}{k!}\Big|D^k f(x_1,\ldots,x_l+\zeta_l^k, x_{l+1}+y_{l+1},\ldots,x_d+y_d) \\
&\qquad - D^k f(x_1,\ldots,x_l, x_{l+1}+y_{l+1},\ldots,x_d+y_d)\Big| \\
&\le \frac{y^k}{k!} L\left(x+(0,\ldots,0,y_{l+1:d}),\zeta_l^k e_l\right)\left|\zeta_l^k\right|^{\beta_l(1-\sum_{i=1}^d k_i/\beta_i)} \\
&\le \tilde{L}(x)\exp\left\{\tau_0||y_{l+1:d}||^2\right\}\exp\left\{\tau_0||\zeta_l^k||^2\right\}\left|\zeta_l^k\right|^{\beta_l(1-\sum_{i=1}^d k_i/\beta_i)} \\
&\le \tilde{L}(x)\frac{y^k}{k!}\exp\left\{\tau_0||y||^2\right\}|y_l|^{\beta_l(1-\sum_{i=1}^d k_i/\beta_i)},
\end{aligned}$$

where we used inequalities (4), (20), and (21) and that $\left|\zeta_l^k\right| \le |y_l|$.

For all sufficiently large n such that $\tau_0 < 0.5/\max_i \sigma_i^2$,

$$\begin{aligned}
&\int \left|\tilde{L}(x)\frac{y^k}{k!}\exp\left\{\tau_0||y||^2\right\}|y_l|^{\beta_l(1-\sum_{i=1}^d k_i/\beta_i)}\right|\phi(y;0,\sigma)dy \\
&\lesssim \tilde{L}(x)\prod_{i=1}^{l-1}\int |y_i|^{k_i}\phi(y_i;0;\sigma_i\sqrt{2})dy_i \cdot \int y_l^{k_l}|y_l|^{\beta_l(1-\sum_{i=1}^d k_i/\beta_i)}\phi(y_l;0;\sigma_l\sqrt{2})dy_l \\
&\lesssim \tilde{L}(x)\sigma_1^{k_1}\cdots\sigma_{l-1}^{k_{l-1}}\sigma_l^{k_l+\beta_l(1-\sum_{i=1}^d k_i/\beta_i)} \\
&= \tilde{L}(x)\sigma_n^{k_1\beta/\beta_1}\cdots\sigma_n^{k_l\beta/\beta_l}\sigma_n^{\frac{\beta}{\beta_l}\beta_l(1-\sum_{i=1}^d k_i/\beta_i)} = \tilde{L}(x)K_2\sigma_n^{\beta},
\end{aligned}$$

where we use $\int |z|^\rho\phi(z,0,\omega)dz \lesssim \omega^\rho$ and $k_{l+1} = \cdots = k_d = 0$ for $k \in \bar{I}_l$. Thus, the claim of the lemma follows.

Q.E.D.

LEMMA 10 *Suppose density $f_0 \in \mathcal{C}^{\beta_1,\ldots,\beta_d,L}$ with a constant envelope L has support on $[0,1]^d$ and $f_0(z) \ge \underline{f} > 0$. Then, $f_{0|J} \in \mathcal{C}^{\beta_{d_{Jc}},\ldots,\beta_d,L/\underline{f}}$.*

PROOF: For $\tilde{x}, \Delta\tilde{x} \in \mathcal{X}$, $y_J \in \mathcal{Y}_J$, and some $\tilde{y}_J^* \in A_{y_J}$, by the mean value theorem,

$$D^k f_{0|J}(\tilde{x}+\Delta\tilde{x}|y_J) - D^k f_{0|J}(\tilde{x}|y_J) =$$

$$= \frac{1}{\pi_{0J}(y_J)} \int_{A_{y_J}} \left(D^{0,\ldots,0,k} f_0(\tilde{y}_J, \tilde{x} + \Delta\tilde{x}) - D^{0,\ldots,0,k} f_0(\tilde{y}_J, \tilde{x}) \right) d\tilde{y}_J$$
$$= \frac{1/N_J}{\pi_{0J}(y_J)} \left(D^{0,\ldots,0,k} f_0(\tilde{y}_J^*, \tilde{x} + \Delta\tilde{x}) - D^{0,\ldots,0,k} f_0(\tilde{y}_J^*, \tilde{x}) \right)$$

and the claim of the lemma follows from the definition of $\mathcal{C}^{\beta_1,\ldots,\beta_d,L}$ and $\pi_{0J}(y_J) \geq \underline{f}/N_J$. *Q.E.D.*

LEMMA 11 *There is a $\lambda_0 \in (0,1)$ such that for any $\lambda \in (0, \lambda_0)$ and any two conditional densities $p, q \in \mathcal{F}$, a probability measure P on $\mathcal{Z}$ that has a conditional density equal to p, and d_h defined with the distribution on $\mathcal{X}$ implied by P,*

$$P \log \frac{p}{q} \leq d_h^2(p,q) \left(1 + 2 \log \frac{1}{\lambda} \right) + 2P \left\{ \left(\log \frac{p}{q} \right) 1 \left(\frac{q}{p} \leq \lambda \right) \right\},$$
$$P \left(\log \frac{p}{q} \right)^2 \leq d_h^2(p,q) \left(12 + 2 \left(\log \frac{1}{\lambda} \right)^2 \right) + 8P \left\{ \left(\log \frac{p}{q} \right)^2 1 \left(\frac{q}{p} \leq \lambda \right) \right\},$$

PROOF: The proof is exactly the same as the proof of Lemma 4 of Shen et al. (2013), which in turn, follows the proof of Lemma 7 in Ghosal and van der Vaart (2007). *Q.E.D.*

LEMMA 12 *Under the assumptions and notation of Section 4.3, for for some $B_0 \in (0, \infty)$ and any $y_J \in \mathcal{Y}_J$,*

$$F_{0|J} \left(||\tilde{X}|| > a_{\sigma_n} | y_J \right) \leq B_0 \sigma_n^{4\beta+2\varepsilon} \underline{\sigma}_n^8.$$

PROOF: Note that in the proof of Proposition 1 of Shen et al. (2013) it is shown that $a_{\sigma_n}^{STG} > a$, where $a_0^{STG} = \{(8\beta + 4\varepsilon + 16)/(b\delta)\}^{1/\tau}$ and $a_{\sigma_n}^{STG} = a_0^{STG} \log(1/\sigma_n)^{1/\tau}$. As $a_0 > a_0^{STG}$ and $a_{\sigma_n} > a_{\sigma_n}^{STG}$, therefore $a_{\sigma_n} > a$. Define $E_{\sigma_n}^* = \left\{ \tilde{x} \in \mathbb{R}^{d_{J^c}} : f_{0|J}(\tilde{x}|y_J) \geq \sigma_n^{(4\beta+2\varepsilon+8\beta/\beta_{\min})/\delta} \right\}$. Note that by construction of s_2 in proof of Proposition 1 of Shen et al. (2013) and as $\sigma_n < s_2$ it follows that

$$\frac{(4\beta + 2\varepsilon + 8)}{b\delta} \log \left(\frac{1}{\sigma_n} \right) \geq \frac{1}{b} \log \bar{f}_0 \implies \sigma_n^{-\frac{(4\beta+2\varepsilon+8)}{\delta}} \geq \bar{f}_0.$$

For $\tilde{x} \in E_{\sigma_n}^*$,

$$f_{0|J}(\tilde{x}|y_J) \geq \sigma_n^{(4\beta+2\varepsilon+8\beta/\beta_{\min})/\delta} = \sigma_n^{(8\beta+4\varepsilon+8\beta/\beta_{\min}+8)/\delta} \sigma_n^{-(4\beta+2\varepsilon+8)/\delta}$$
$$\geq \bar{f}_0 \sigma_n^{(8\beta+4\varepsilon+8\beta/\beta_{\min}+8)/\delta} = \bar{f}_0 \sigma_n^{a_0^\tau b} = \bar{f}_0 \exp \left\{ -b a_0^\tau \log(\frac{1}{\sigma_n}) \right\}$$
$$= \bar{f}_0 \exp \left\{ -b \left(a_0 (\log(\frac{1}{\sigma_n})^{1/\tau}) \right)^\tau \right\} = \bar{f}_0 \exp \left\{ -b a_{\sigma_n}^\tau \right\}.$$

As $a_{\sigma_n} > a$ and as $f_{0|J}(\tilde{x}|y_J) \geq \bar{f}_0 \exp\{-ba_{\sigma_n}^{\tau}\}$, then the tail condition (17) is satisfied only if $||\tilde{x}|| < a_{\sigma_n}$. Therefore, $E_{\sigma_n}^* \subset \left\{\tilde{x} \in \mathbb{R}^{d_J} \ : \ ||\tilde{x}|| \leq a_{\sigma_n}\right\}$. As in the proof of Proposition 1 of Shen et al. (2013), by Markov's inequality,

$$
\begin{aligned}
&F_{0|J}\left(||\tilde{X}|| > a_{\sigma_n}|y_J\right) \leq F_{0|J}(E_{\sigma_n}^{*,c}|y_J) \\
&= F_{0|J}\left(f_{0|J}(\tilde{x}|y_J)^{-\delta} > \sigma_n^{-(4\beta+2\varepsilon+8\beta/\beta_{\min})}|y_J\right) \\
&\leq B_0 \sigma_n^{4\beta+2\varepsilon+8\beta/\beta_{\min}} = B_0 \sigma_n^{4\beta+2\varepsilon} \underline{\sigma}_n^8
\end{aligned}
$$

as desired since $\sigma_n^{\beta/\beta_{\min}} = \underline{\sigma}_n$ and the tail condition on $f_{0|J}(\cdot|y_J)$, (17), implies the existence of a $\delta > 0$ small enough such that $E_{0|J}(f_{0|J}^{-\delta}) \leq B_0 < \infty$ for any $y_J \in \mathcal{Y}_J$. *Q.E.D.*

LEMMA 13 *Under the assumptions and notation of Section 4.3, for $m = KN_J$ and any $\theta \in S_{\theta^\star}$*

$$
d_H^2(p_{|J}^\star(\cdot|\cdot)\pi_0(\cdot), p(\cdot,\cdot|\theta,m)) \lesssim \sigma_n^{2\beta}.
$$

PROOF: Let us define

$$
f_J(y_J, \tilde{x}|\theta, m) = \int_{A_{y_J}} f(\tilde{y}_J, \tilde{x}|\theta, m) d\tilde{y}_J.
$$

Then,

$$
\begin{aligned}
&d_H^2(p_{|J}^\star(\cdot|\cdot)\pi_0(\cdot), p(\cdot,\cdot|\theta,m)) \leq d_{TV}(p_{|J}^\star(\cdot|\cdot)\pi_0(\cdot), p(\cdot,\cdot|\theta,m)) \\
&\leq d_{TV}(f_{|J}^\star(\cdot|\cdot)\pi_0(\cdot), f_J(\cdot,\cdot|\theta,m)) \\
&= \sum_{y_J \in \mathcal{Y}_J} \int_{\tilde{\mathcal{X}}} \Bigg| \sum_{k \in \mathcal{Y}_J} \sum_{j=1}^{K} \alpha_{j|k}^\star \pi_0(k) \mathbf{1}\{k = y_J\} \phi(\tilde{x}, \mu_{j|k}^\star, \sigma_{J^c}^\star) \\
&\qquad\qquad - \alpha_{jk} \int_{A_{y_J}} \phi(\tilde{y}_J, \mu_{jk,J}, \sigma_J) d\tilde{y}_J \cdot \phi(\tilde{x}, \mu_{jk,J^c}, \sigma_{J^c}) \Bigg| d\tilde{x} \\
&\leq \sum_{y_J \in \mathcal{Y}_J} \int_{\tilde{\mathcal{X}}} \left| \sum_{k \in \mathcal{Y}_J} \sum_{j=1}^{K} \alpha_{jk}^\star \mathbf{1}\{k = y_J\} \phi(\tilde{x}, \mu_{j|k}^\star, \sigma_{J^c}^\star) - \alpha_{jk}^\star \mathbf{1}\{k = y_J\} \phi(\tilde{x}, \mu_{jk,J^c}, \sigma_{J^c}) \right| d\tilde{x} \\
&+ \sum_{y_J \in \mathcal{Y}_J} \int_{\tilde{\mathcal{X}}} \Bigg| \sum_{k \in \mathcal{Y}_J} \sum_{j=1}^{K} \alpha_{jk}^\star \mathbf{1}\{k = y_J\} \phi(\tilde{x}, \mu_{jk,J^c}, \sigma_{J^c}) \\
&\qquad\qquad - \alpha_{jk} \int_{A_{y_J}} \phi(\tilde{y}, \mu_{jk,J}, \sigma_J) d\tilde{y} \phi(\tilde{x}, \mu_{jk,J^c}, \sigma_{J^c}) \Bigg| d\tilde{x},
\end{aligned}
$$

where the first inequality follows from $d_H^2(\cdot,\cdot) \leq d_{TV}(\cdot,\cdot)$, the second inequality holds by Lemma 6, and the last inequality is obtained by the triangle inequality.

Let's explore the two parts of the right hand side in the last inequality independently. First,

$$\begin{aligned}
&\sum_{y_J\in\mathcal{Y}_J}\int_{\tilde{\mathcal{X}}}\left|\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\alpha^\star_{jk}\mathbf{1}\{k=y_J\}\phi(\tilde{x},\mu^\star_{j|k},\sigma^\star_{J^c})-\alpha^\star_{jk}\mathbf{1}\{k=y_J\}\phi(\tilde{x},\mu_{jk,J^c},\sigma_{J^c})\right|d\tilde{x}\\
&\le\sum_{y_J\in\mathcal{Y}_J}\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\alpha^\star_{jk}\mathbf{1}\{k=y_J\}\int_{\tilde{\mathcal{X}}}\left|\phi(\tilde{x},\mu^\star_{j|k},\sigma^\star_{J^c})-\phi(\tilde{x},\mu_{jk,J^c},\sigma_{J^c})\right|d\tilde{x}\\
&\le\max_{j\le N,k\in\mathcal{Y}_J}d_{TV}(\phi(\cdot;\mu^\star_{j|k},\sigma^\star_{J^c}),\phi(\cdot,\mu_{jk,J^c},\sigma_{J^c}))\lesssim\sigma_n^{2\beta},
\end{aligned}$$

where the fact that $\alpha^\star_{j,k}=0$ for $j>N$ by design is used to get $j\le N$ rather than $j\le K$ in the max subscript. The last inequality is proved in Lemma 14.

Second,

$$\begin{aligned}
&\sum_{y_J\in\mathcal{Y}_J}\int_{\tilde{\mathcal{X}}}\left|\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\alpha^\star_{jk}\mathbf{1}\{k=y_J\}\phi(\tilde{x},\mu_{jk,J^c},\sigma_{J^c})\right.\\
&\qquad\left.-\alpha_{jk}\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\phi(\tilde{x},\mu_{jk,J^c},\sigma_{J^c})\right|d\tilde{x}\\
&=\sum_{j=1}^{K}\left(\sum_{y_J\in\mathcal{Y}_J}\left|\sum_{k\in\mathcal{Y}_J}\alpha^\star_{jk}\mathbf{1}\{k=y_J\}-\alpha_{jk}\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\right|\int_{\tilde{\mathcal{X}}}\phi(\tilde{x},\mu_{jk,J^c},\sigma_{J^c})d\tilde{x}\right)\\
&=\sum_{j=1}^{K}\sum_{y_J\in\mathcal{Y}_J}\left|\sum_{k\in\mathcal{Y}_J}\alpha^\star_{jk}\mathbf{1}\{k=y_J\}-\alpha_{jk}\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\right|\\
&\le\sum_{y_J\in\mathcal{Y}_J}\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\left|\alpha^\star_{jk}\mathbf{1}\{k=y_J\}-\alpha^\star_{jk}\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\right|\\
&+\sum_{y_J\in\mathcal{Y}_J}\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\left|\alpha^\star_{jk}\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J-\alpha_{jk}\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\right|\\
&\le\sum_{y_J\in\mathcal{Y}_J}\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\alpha^\star_{jk}\left|\mathbf{1}\{k=y_J\}-\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\right|\\
&+\sum_{y_J\in\mathcal{Y}_J}\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\left|\alpha^\star_{jk}-\alpha_{jk}\right|\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\\
&=\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\left(\alpha^\star_{jk}\sum_{y_J\in\mathcal{Y}_J}\left|\mathbf{1}\{k=y_J\}-\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\right|\right)\\
&+\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\left(\left|\alpha^\star_{jk}-\alpha_{jk}\right|\sum_{y_J\in\mathcal{Y}_J}\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\right)\\
&\le\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\alpha^\star_{jk}\left[\int_{A_k^c}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J+\sum_{y_J\neq k}\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\right]
\end{aligned}$$

$$+\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\left|\alpha^\star_{jk}-\alpha_{jk}\right|$$

$$=\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\alpha^\star_{jk}\cdot 2\int_{A_k^c}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J+\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\left|\alpha^\star_{jk}-\alpha_{jk}\right|$$

$$\leq 2\max_{j\leq N,k\in\mathcal{Y}_J}\int_{A_k^c}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J+\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\left|\alpha^\star_{jk}-\alpha_{jk}\right|\lesssim\sigma_n^{2\beta}.$$

The last inequality follows from Lemma 15 and the definition of $S_{\theta^\star}$.

Q.E.D.

LEMMA 14 *Under the assumptions and notation of Section 4.3,*

$$\max_{j\leq N,k\in\mathcal{Y}_J} d_{TV}(\phi(\cdot;\mu^\star_{j|k},\sigma^\star_{J^c}),\phi(\cdot,\mu_{jk,J^c},\sigma_{J^c}))\lesssim\sigma_n^{2\beta}.$$

PROOF: Fix some $j\leq N$ and $k\in\mathcal{Y}_J$. It is known that

$$d_{TV}(\phi(\cdot;\mu^\star_{j|k},\sigma^\star_{J^c}),\phi(\cdot,\mu_{jk,J^c},\sigma_{J^c}))\leq 2\sqrt{d_{KL}(\phi(\cdot;\mu^\star_{j|k},\sigma^\star_{J^c}),\phi(\cdot,\mu_{jk,J^c},\sigma_{J^c}))}$$

and

$$d_{KL}(\phi(\cdot;\mu^\star_{j|k},\sigma^\star_{J^c}),\phi(\cdot,\mu_{jk,J^c},\sigma_{J^c}))=\sum_{i\in J^c}\frac{\sigma_i^2}{\sigma_i^{\star 2}}-1-\log\frac{\sigma_i^2}{\sigma_i^{\star 2}}+\frac{(\mu^\star_{j|k,i}-\mu_{jk,i})^2}{\sigma_i^{\star 2}}.$$

From the definition of $S_{\theta^\star}$,

$$\sum_{i\in J^c}\frac{(\mu^\star_{j|k,i}-\mu_{jk,i})^2}{\sigma_i^{\star 2}}\leq\tilde{\epsilon}_n^{4b_1}\leq\sigma_n^{4\beta}.$$

Since $\sigma_i^2\in(\sigma_i^{\star 2}(1+\sigma_n^{2\beta})^{-1},\sigma_i^{\star 2})$ and the fact that $|z-1-\log z|\lesssim|z-1|^2$ for z in a neighborhood of 1, we have for all sufficiently large n

$$\left|\frac{\sigma_i^2}{\sigma_i^{\star 2}}-1-\log\frac{\sigma_i^2}{\sigma_i^{\star 2}}\right|\lesssim\left(1-\frac{\sigma_i^2}{\sigma_i^{\star 2}}\right)^2\lesssim\sigma_n^{4\beta}.$$

The three inequalities derived above imply the claim of the lemma.

Q.E.D.

LEMMA 15 *Under the assumptions and notation of Section 4.3, for $\theta\in S_{\theta^\star}$,*

$$\max_{j\leq N,k\in\mathcal{Y}_J}\int_{A_k^c}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\lesssim\sigma_n^{2\beta}.$$

PROOF: Fix $j \leq N$, $k \in \mathcal{Y}_J$, and $\theta \in S_{\theta^\star}$. Since $\mu_{jk,i} \in \left[k_i - \frac{1}{4N_i}, k_i + \frac{1}{4N_i}\right]$,

$$\begin{aligned}
\int_{A_k^c} \phi(\tilde{y}_J, \mu_{jk,J}, \sigma_J) d\tilde{y}_J &\leq \sum_{i \in J} Pr\left(\tilde{y}_i \notin \left[k_i - \frac{1}{2N_i}, k_i + \frac{1}{2N_i}\right]\right) \\
&\leq \sum_{i \in J} Pr\left(\tilde{y}_i \notin \left[\mu_{jk,i} - \frac{1}{4N_i}, \mu_{jk,i} + \frac{1}{4N_i}\right]\right) \\
&= 2 \sum_{i \in J} \int_{-\infty}^{-\frac{1}{4N_i\sigma_i}} \phi(\tilde{y}_i, 0, 1) d\tilde{y}_i \\
&\leq 2 \sum_{i \in J} \exp\left\{-\frac{1}{2(4N_i\sigma_i)^2}\right\} \leq 2 \sum_{i \in J} \sigma_n^{2\beta} \lesssim \sigma_n^{2\beta},
\end{aligned}$$

where the last inequality follows from the restrictions on σ_J in $S_{\theta^\star}$ and the penultimate inequality follows from a bound on the normal tail probability derived below.

If $\tilde{Y}_i$ has $N(0,1)$ distribution, then the moment generating function is $M(\theta) = \exp\{\theta^2/2\}$. Note that $\exp\{\theta(\tilde{Y}_i - (4N_i\sigma_i)^{-1})\} \geq 1$ when $\tilde{Y}_i \leq (4N_i\sigma_i)^{-1}$ and $\theta \leq 0$, therefore,

$$\begin{aligned}
&\int_{-\infty}^{-\frac{1}{4N_i\sigma_i}} \phi(\tilde{y}_i, 0, 1) d\tilde{y}_i \leq \inf_{\theta \leq 0} \mathbb{P} \exp\left\{\theta(\tilde{Y}_i - (4N_i\sigma_i)^{-1})\right\} \\
&= \inf_{\theta \leq 0} \exp\left\{-\theta(4N_i\sigma_i)^{-1}\right\} M(\theta) \\
&= \inf_{\theta \leq 0} \exp\left\{-\theta(4N_i\sigma_i)^{-1}\right\} \exp\left\{\theta^2/2\right\} = \exp\left\{-(4N_i\sigma_i)^{-2}/2\right\}.
\end{aligned}$$

Q.E.D.

LEMMA 16 *Under the assumptions and notation of Section 4.3, for any $(y_J, y_I) \in \mathcal{Y}$, some constants $C_3, C_4 > 0$ and all sufficiently large n,*

$$\frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} \geq C_3 \frac{\sigma_n^{2\beta}}{m^2} \equiv \lambda_n, \tag{59}$$

when $\|x\| \leq a_{\sigma_n}$ and

$$\frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} \geq \exp\left\{-\frac{8||x||^2}{\underline{\sigma}_n^2} - C_4 \log n\right\} \tag{60}$$

when $\|x\| > a_{\sigma_n}$.

PROOF: By assumption (17), $f_{0|J}(\tilde{x}|y_J) \leq \bar{f}_0$, and $\pi_{0J}(y_J) \leq 1$ for all $(\tilde{x}, y_J)$. Therefore,

$$\frac{f_J(y_J, \tilde{x}|\theta, m)}{f_{0|J}(\tilde{x}|y_J)\pi_{0J}(y_J)} \geq \bar{f}_0^{-1} f_J(y_J, \tilde{x}|\theta, m) \tag{61}$$

Let $k^\star = y_J$. Then, by Lemma 15, for any $j \in \{1, \ldots, K\}$,

$$\int_{A_{y_J}} \phi(\tilde{y}_J; \mu_{jk^*,J}, \sigma_J) d\tilde{y}_J \geq \frac{1}{2}$$

for all n large enough as $\sigma_n \to 0$.

For any $\tilde{x} \in \tilde{\mathcal{X}}$ with $\|\tilde{x}\| \leq 2a_{\sigma_n}$, by the construction of sets $U_{j|k^\star}$, there exists $j^\star \in \{1,\ldots,K\}$ such that $\tilde{x}, \mu_{j^\star|k^\star} \in U_{j^\star|k^\star}$ and for all sufficiently large n, $\sum_{i\in J^c}(\tilde{x}_i - \mu_{j^\star|k^\star,i})^2/\sigma_i^2 \leq 4$. Then,

$$\begin{aligned}\phi(\tilde{x}, \mu_{j^\star|k^\star}, \sigma_{J^c}) &= (2\pi)^{-d_{J^c}/2}\prod_{i\in J^c}\sigma_i^{-1}\exp\left\{-0.5\sum_{i\in J^c}(\tilde{x}_i - \mu_{j^\star|k^\star,i})^2/\sigma_i^2\right\}\\ &\geq (2\pi)^{-d_{J^c}/2}\sigma_n^{-d_{J^c}}e^{-2}.\end{aligned}$$

Thus,

$$\begin{aligned}f_J(y_J,\tilde{x}|\theta) &= \sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\alpha_{jk}\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\phi(\tilde{x},\mu_{jk,J^c},\sigma_{J^c})\\ &\geq \alpha_{j^\star k^*}\phi(\tilde{x},\mu_{j^\star k^\star,J^c},\sigma_{J^c})\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{j^\star k^\star,J},\sigma_J)d\tilde{y}_J\end{aligned}$$

and for $C_3 = \bar{f}_0^{-1}(2\pi)^{-d_{J^c}/2}e^{-2}/8$,

$$\begin{aligned}\frac{f_J(y_J,\tilde{x}|\theta,m)}{f_{0|J}(\tilde{x}|y_J)\pi_{0J}(y_J)} &\geq \bar{f}_0^{-1}\cdot\min_{j\leq K,k\in\mathcal{Y}_J}\alpha_{jk}\cdot(2\pi)^{-d_{J^c}/2}\sigma_n^{-d_{J^c}}e^{-2}\cdot\frac{1}{2}\\ &\geq 2C_3\frac{\sigma_n^{2\beta}}{m^2} = 2\lambda_n. \qquad (62)\end{aligned}$$

By assumption (18), for any $x \in \mathcal{X}$, any $y_J \in \mathcal{Y}_J$, and all sufficiently large n,

$$\int_{A_{y_I}} f_{0|J}(\tilde{x}|y_J)\pi_{0J}(y_J)d\tilde{y}_I \leq 2\int_{A_{y_I}\cap\{\tilde{y}_I:\|\tilde{y}_I\|\leq a_{\sigma_n}\}} f_{0|J}(\tilde{x}|y_J)\pi_{0J}(y_J)d\tilde{y}_I. \qquad (63)$$

For any $x \in \mathcal{X}$ with $\|x\| \leq a_{\sigma_n}$ and $\tilde{y}_I \in A_{y_I}\cap\{\tilde{y}_I : \|\tilde{y}_I\| \leq a_{\sigma_n}\}$, we have $\|\tilde{x}\| \leq 2a_{\sigma_n}$ and

$$\begin{aligned}\frac{p(y_J,y_I,x|\theta,m)}{p_0(y_J,y_I,x)} &= \frac{\int_{A_{y_I}} f_J(y_J,\tilde{x}|\theta,m)d\tilde{y}_I}{\int_{A_{y_I}} f_{0|J}(\tilde{x}|y_J)\pi_{0J}(y_J)d\tilde{y}_I}\\ &\geq \frac{\int_{A_{y_I}\cap\{\tilde{y}_I:\|\tilde{y}_I\|\leq a_{\sigma_n}\}} f_J(y_J,\tilde{x}|\theta,m)d\tilde{y}_I}{2\int_{A_{y_I}\cap\{\tilde{y}_I:\|\tilde{y}_I\|\leq a_{\sigma_n}\}} f_{0|J}(\tilde{x}|y_J)\pi_{0J}(y_J)d\tilde{y}_I} \geq \lambda_n, \qquad (64)\end{aligned}$$

where the first inequality follows from (63) and the second one from (62) combined with Lemma 6.

Next, let us bound $f_J(y_J,\tilde{x}|\theta,m)/f_{0|J}(\tilde{x}|y_J)\pi_0(y_J)$ from below for $\tilde{x} \in \tilde{\mathcal{X}}$ such that $\|x\| > a_{\sigma_n}$ and $\|\tilde{y}_I\| \leq a_{\sigma_n}$. For any $j \leq K$ and $k \in \mathcal{Y}_J$, $||\tilde{x} - \mu_{jk,J^c}||^2 \leq 2(||\tilde{x}||^2 + ||\mu_{jk,J^c}||^2) \leq 16||x||^2$ as $||\mu_{jk,J^c}|| \leq 2a_{\sigma_n}$ by construction of $U_{j|k}$ and $2||x|| > ||\tilde{x}||$. Then

$$\phi(\tilde{x},\mu_{jk,J^c},\sigma_{J^c}) = (2\pi)^{-d_{J^c}/2}\prod_{i\in J^c}\sigma_i^{-1}\exp\left\{-0.5\sum_{i\in J^c}(\tilde{x}_i - \mu_{jk,i})^2/\sigma_i^2\right\}$$

$$\geq (2\pi)^{-d_{J^c}/2}\sigma_n^{-d_{J^c}} \exp\left\{-\frac{8||x||^2}{\underline{\sigma}_n^2}\right\}.$$

Then, for n large enough

$$f_J(y_J,\tilde{x}|\theta,m) = \sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\alpha_{jk}\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\phi(\tilde{x},\mu_{jk,J^c},\sigma_{J^c})$$
$$\geq (2\pi)^{-d_{J^c}/2}\sigma_n^{-d_{J^c}} \exp\left\{-\frac{8||x||^2}{\underline{\sigma}_n^2}\right\}\sum_{j=1}^{K}\alpha_{jk}\sum_{k\in\mathcal{Y}_J}\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J$$
$$\geq (2\pi)^{-d_{J^c}/2}\sigma_n^{-d_{J^c}} \exp\left\{-\frac{8||x||^2}{\underline{\sigma}_n^2}\right\}\frac{1}{2}K\min_{j,k}\alpha_{jk}.$$

Combining this inequality with (61), we get

$$\frac{f_J(y_J,\tilde{x}|\theta,m)}{f_{0|J}(\tilde{x}|y_J)\pi_{0J}(y_J)} \geq \frac{1}{2}(2\pi)^{-d_{J^c}/2}\bar{f}_0^{-1}\sigma_n^{-d_{J^c}}K\frac{\sigma_n^{2\beta+d_{J^c}}}{2m^2}\exp\left\{-\frac{8||x||^2}{\underline{\sigma}_n^2}\right\}$$
$$\geq \exp\left\{-\frac{8||x||^2}{\underline{\sigma}_n^2} - C_4\log n\right\} \tag{65}$$

for sufficiently large C_4 because $|\log\left[K\sigma_n^{2\beta}/m^2\right]| \lesssim \log n$.

Thus, for $||x|| > a_{\sigma_n}$, (65) and the first inequality in (64), which holds for any $x \in \mathcal{X}$, deliver

$$\frac{p(y_J,y_I,x|\theta,m)}{p_0(y_J,y_I,x)} \geq \exp\left\{-\frac{8||x||^2}{\underline{\sigma}_n^2} - C_4\log n\right\}. \tag{66}$$

Q.E.D.

LEMMA 17 *Under the assumptions and notation of Section 4.3, for $\lambda_n < \lambda_0$, where λ_0 is defined in Lemma 11,*

$$E_0\left(\left[\log\frac{p_0(y_J,y_I,x)}{p(y_J,y_I,x|\theta,m)}\right]^2\right) \leq A\tilde{\epsilon}_n^2$$
$$E_0\left(\left[\log\frac{p_0(y_J,y_I,x)}{p(y_J,y_I,x|\theta,m)}\right]\right) \leq A\tilde{\epsilon}_n^2$$

PROOF:

$$E_0\left(\left[\log\frac{p_0(y_J,y_I,x)}{p(y_J,y_I,x|\theta,m)}\right]^2\right)$$
$$\leq d_H^2(p_0(\cdot,\cdot),p(\cdot,\cdot|\theta,m))\left(12 + 2\left(\log\frac{1}{\lambda_n}\right)^2\right)$$
$$+ 8P\left\{\left(\log\frac{p_0(\cdot,\cdot)}{p(\cdot,\cdot|\theta,m)}\right)^2\mathbf{1}\left\{\frac{p(\cdot,\cdot|\theta,m)}{p_0(\cdot,\cdot)} < \lambda_n\right\}\right\}$$

$$\lesssim \sigma_n^{2\beta}(12 + 2\log(1/\lambda_n)^2) + \sigma_n^{2\beta+\epsilon} \lesssim \log(1/\lambda_n)^2\sigma_n^{2\beta},$$

where first inequality is derived using Lemma 11 and penultimate inequality is derived using inequalities (51) and (55). Similarly,

$$\begin{aligned}
&E_0\left(\log\frac{p_0(y_J, y_I, x)}{p(y_J, y_I, x|\theta, m)}\right) \\
&\leq d_H^2(p_0(\cdot,\cdot), p(\cdot,\cdot|\theta, m))\left(1 + 2\left(\log\frac{1}{\lambda_n}\right)\right) \\
&\quad + 2P\left\{\left(\log\frac{p_0(\cdot,\cdot)}{p(\cdot,\cdot|\theta, m)}\right)\mathbf{1}\left\{\frac{p(\cdot,\cdot|\theta, m)}{p_0(\cdot,\cdot)} < \lambda_n\right\}\right\} \\
&\lesssim \sigma_n^{2\beta}(1 + 2\log(1/\lambda_n)) + \sigma_n^{2\beta+\epsilon} \lesssim \log(1/\lambda_n)\sigma_n^{2\beta}.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\log(1/\lambda_n)\sigma_n^{2\beta} \leq \log(1/\lambda_n)^2\sigma_n^{2\beta} &= \log\left(\frac{2N_JK^2}{\sigma_n^{2\beta}}\right)^2 \tilde{\epsilon}_n^2(\log(\tilde{\epsilon}_n^{-1}))^{-2} \\
&\leq \left(\frac{\log[2N_J^2(C_1\sigma_n^{-d_{Jc}}\{\log(\tilde{\epsilon}_n^{-1})\}^{d_{Jc}+d_{Jc}/\tau})^2\sigma_n^{-2\beta}]}{\log(\tilde{\epsilon}_n^{-1})}\right)^2 \tilde{\epsilon}_n^2,
\end{aligned}$$

where the term multiplying $\tilde{\epsilon}_n^2$ on the right hand side is bounded by Assumption 5 ($N_J = o(n^{1-\nu})$) and definitions of $\tilde{\epsilon}_n$ and σ_n. *Q.E.D.*

LEMMA 18 *Under the assumptions and notation of Section 4.3, for all sufficiently large n, $s = 1 + 1/\beta + 1/\tau$, and some $C_6 > 0$*

$$\Pi(m = N_JK, \theta \in S_{\theta^\star}) \geq \exp\left[-C_6N_J\tilde{\epsilon}_n^{-d_{Jc}/\beta}\{\log(n)\}^{d_{Jc}s+\max\{\tau_1,1,\tau_2/\tau\}}\right].$$

PROOF: First, consider the prior probability of $m = N_JK$. By (3) for some $C_{61} > 0$,

$$\begin{aligned}
(67)\qquad &\Pi(m = N_JK) \propto \exp[-a_{10}N_JK(\log N_JK)^{\tau_1}] \\
&\geq \exp[-C_{61}N_J\tilde{\epsilon}_n^{-d_{Jc}/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{sd_{Jc}}(\log n)^{\tau_1}] \\
&\geq \exp[-C_{61}N_J\tilde{\epsilon}_n^{-d_{Jc}/\beta}\{\log(n)\}^{sd_{Jc}+\tau_1}]
\end{aligned}$$

as $N_J = o(n^{1-\nu})$ by (24) and $\tilde{\epsilon}_n^{-1} < n$.

Second, consider the prior on $\{\alpha_{jk}\}$. There exist (j_0, k_0) such that $\alpha^\star_{j_0k_0} \geq \frac{1}{m}$ and suppose that $|\alpha^\star_{jk} - \alpha_{jk}| \leq \frac{\sigma_n^{2\beta}}{m^2}$ for all $(j,k) \neq (j_0, k_0)$. Then,

$$\left|\alpha^\star_{j_0k_0} - \alpha_{j_0k_0}\right| = \left|\sum_{(jk)\neq(j_0k_0)} \alpha^\star_{jk} - \alpha_{jk}\right| \leq (m-1)\frac{\sigma_n^{2\beta}}{m^2} \leq \frac{\sigma_n^{2\beta}}{m}$$

$$\alpha_{j_0k_0} \geq \alpha^\star_{j_0k_0} - \frac{\sigma_n^{2\beta}}{m} \geq \frac{1-\sigma_n^{2\beta}}{m} \geq \frac{\sigma_n^{2\beta+d_{J^c}}}{2m^2}.$$

Furthermore,

$$\sum_{j=1}^{K}\sum_{k\in\mathcal{Y}_J} |\alpha_{jk} - \alpha^\star_{jk}| \leq (m-1)\frac{\sigma_n^{2\beta}}{m^2} + \frac{\sigma_n^{2\beta}}{m} \leq 2\sigma_n^{2\beta}.$$

It then follows that

$$\begin{aligned}
&\Pi\left(\sum_{j=1}^{K}\sum_{k\in\mathcal{Y}_J} |\alpha_{jk} - \alpha^\star_{jk}| \leq 2\sigma_n^{2\beta}, \min_{j\leq K, k\in\mathcal{Y}_J} \alpha_{jk} \geq \frac{\sigma_n^{2\beta+d_{J^c}}}{2m^2}\right) \\
&\geq \Pi\left(|\alpha_{jk} - \alpha^\star_{jk}| \leq \frac{\sigma_n^{2\beta}}{m^2}, \alpha_{jk} \geq \frac{\sigma_n^{2\beta}}{2m^2},\ (j,k) \in \{1,\ldots,K\}\times\mathcal{Y}_J \setminus \{(j_0,k_0)\}\right) \\
&\geq \exp\left\{-C_{62}N_J K \log(N_J K/\sigma_n^\beta)\right\},
\end{aligned}$$

where the last inequality is derived in the proof of Lemma 10 in Ghosal and van der Vaart (2007) for some $C_{62} > 0$ (see, also, Lemma 6.1 in Ghosal et al. (2000)). Note that

$$\begin{aligned}
&K\log(N_J K/\sigma_n^\beta) \\
&\leq \tilde{\epsilon}_n^{-d_{J^c}/\beta}\log(\tilde{\epsilon}_n^{-1})^{d_{J^c}s}\log(N_J\tilde{\epsilon}_n^{-d_{J^c}/\beta-1}\log(\tilde{\epsilon}_n^{-1})^{d_{J^c}s+1}) \\
&\lesssim \tilde{\epsilon}_n^{-d_{J^c}/\beta}\log(n)^{d_{J^c}s+1}.
\end{aligned} \tag{68}$$

Assumption (11) on the prior for σ_i implies that for $i \in J$

$$\begin{aligned}
&\prod_{i=1}^{d_J}\Pi(\sigma_i^{-2} \geq 32N_i^2\beta\log\sigma_n^{-1}) \\
&\geq \prod_{i=1}^{d_J}\left(a_6(64N_i^2\beta\log\sigma_n^{-1})^{a_7}\exp\left\{-a_9(64N_i^2\beta\log\sigma_n^{-1})^{1/2}\right\}\right) \\
&\geq \exp\left\{-C_{63}N_J\log(\sigma_n^{-1})\right\} \geq \exp\left\{-C_{64}N_J\log(n)\right\},
\end{aligned} \tag{69}$$

and for $i \in J^c$,

$$\begin{aligned}
&\prod_{i=1}^{d_{J^c}}\Pi\left(\sigma_{i,n}^{-2} \leq \sigma_i^{-2} \leq \sigma_{i,n}^{-2}(1+\sigma_n^{2\beta})\right) \geq \prod_{i=1}^{d_{J^c}}\left(a_6(\sigma_{i,n}^{-2})^{a_7}\sigma_n^{2a_8\beta}\exp\left\{-a_9\sigma_{i,n}^{-1}\right\}\right) \\
&\geq \prod_{i=1}^{d_{J^c}}\exp\left\{-C_{65}\sigma_{i,n}^{-1}\right\} = \prod_{i=1}^{d_{J^c}}\exp\left\{-C_{65}\sigma_n^{-\beta/\beta_i}\right\} \geq \exp\left\{-C_{65}d_{J^c}\sigma_n^{-d_{J^c}}\right\} \\
&\geq \exp\left\{-C_{66}\tilde{\epsilon}_n^{-d_{J^c}/\beta}\log(n)^{d_{J^c}/\beta}\right\}.
\end{aligned} \tag{70}$$

Assumption (7) on the prior for μ_{jk} implies

$$\prod_{j=1}^{K}\prod_{k\in\mathcal{Y}_J}\prod_{i\in J}\Pi\left(\mu_{jk,i} \in \left[k_i - \frac{1}{4N_i}, k_i + \frac{1}{4N_i}\right]\right) \tag{71}$$

$$\geq \left(a_{11}2^{-d_J}N_J^{-1}\exp\{-a_{12}\}\right)^{N_JK}$$
$$\geq \exp\{-C_{67}N_JK\log(N_J)\}$$
$$\geq \exp\left\{-C_{68}N_J\tilde{\epsilon}_n^{-d_{Jc}/\beta}\log(n)^{d_{Jc}s+1}\right\}$$

and

$$\prod_{j=1}^{K}\prod_{k\in\mathcal{Y}_J}\Pi\left(\mu_{jk,J^c}\in U_{j|k}\right) \geq \left(a_{11}\exp\left\{-a_{12}a_{\sigma_n}^{\tau_2}\right\}\min_{j,k}Vol(U_{j|k})\right)^{N_JK} \tag{72}$$
$$= \left(a_{11}\exp\left\{-a_{12}a_{\sigma_n}^{\tau_2}\right\}\sigma_n^{d_{Jc}}\tilde{\epsilon}_n^{2b_1d_{Jc}}\right)^{N_JK}$$
$$\geq \exp\left\{-C_{69}N_J\tilde{\epsilon}_n^{-d_{Jc}/\beta}\log(n)^{d_{Jc}s+\max\{1,\tau_2/\tau\}}\right\}.$$

It follows from (67) - (72), that for all sufficiently large n and some $C_6 > 0$,

$$\Pi(\mathcal{K}(p_0,\tilde{\epsilon}_n)) \geq \Pi(m = N_JK, \theta\in S_{\theta^\star})$$
$$\geq \exp[-C_6N_J\tilde{\epsilon}_n^{-d_{Jc}/\beta}\{\log(n)\}^{d_{Jc}s+\max\{\tau_1,1,\tau_2/\tau\}}].$$

Q.E.D.

LEMMA 19 *For $H\in\mathbb{N}$, $0<\underline{\sigma}<\overline{\sigma}$, and $\overline{\mu}>0$, let us define a sieve*

$$\mathcal{F} = \{p(y,x|\theta,m) : m\leq H,\ \mu_j\in[-\overline{\mu},\overline{\mu}]^d, j=1,\ldots,m, \tag{73}$$
$$\sigma_i\in[\underline{\sigma},\overline{\sigma}], i=1,\ldots,d\}.$$

For $0<\epsilon<1$ and $\underline{\sigma}\leq 1$,

$$M_e(\epsilon,\mathcal{F},d_{TV}) \leq H\cdot\left\lceil\frac{12\overline{\mu}d}{\underline{\sigma}\epsilon}\right\rceil^{Hd}\cdot\left\lceil\frac{15}{\epsilon}\right\rceil^{H}\cdot\left\lceil\frac{\log(\overline{\sigma}/\underline{\sigma})}{\log(1+\epsilon/[12d])}\right\rceil^{d}.$$

For all sufficiently large H, large $\overline{\sigma}$ and small $\underline{\sigma}$,

$$\Pi(\mathcal{F}^c) \leq H^2d\exp\{-a_{13}\overline{\mu}^{\tau_3}\} + \exp\{-a_{10}H(\log H)^{\tau_1}\}$$
$$+ da_1\exp\{-a_2\underline{\sigma}^{-2a_3}\} + da_4\exp\{-2a_5\log\overline{\sigma}\}.$$

PROOF: The proof is similar to proofs of related results in Norets and Pati (2017), Shen et al. (2013), and Ghosal and van der Vaart (2001) among others.

Let us begin with the first claim. For a fixed value of m, define set S_μ^m to contain centers of $|S_\mu^m| = \lceil 12\overline{\mu}d/(\underline{\sigma}\epsilon)\rceil$ equal length intervals partitioning $[-\overline{\mu},\overline{\mu}]$. Let S_α^m be an $\epsilon/3$-net of Δ^{m-1} in total variation distance ($\forall\alpha\in\Delta^{m-1}$, $\exists\tilde{\alpha}\in S_\alpha^m$, $d_{TV}(\alpha,\tilde{\alpha})\leq\epsilon/3$). From Lemma A.4 in Ghosal and van der Vaart (2001), the cardinality of S_α^m, is bounded as follows

$$|S_\alpha^m| \leq [15/\epsilon]^m.$$

Define $S_\sigma = \{\sigma^l, l = 1, \ldots, \lceil \log(\overline{\sigma}/\underline{\sigma})/(\log(1+\epsilon/(12d)\rceil, \sigma^1 = \underline{\sigma}, (\sigma^{l+1} - \sigma^l)/\sigma^l = \epsilon/(12d)\}$.

Let us show that

$$S_{\mathcal{F}} = \{p(y, x|\theta, m) : \ m \leq H, \ \alpha \in S_\alpha^m, \ \sigma_i \in S_\sigma, \ \mu_{ji} \in S_\mu^m, \ j \leq m, \ i \leq d\}$$

is an ϵ-net for $\mathcal{F}$ in d_{TV}. For a given $p(\cdot|\theta, m) \in \mathcal{F}$ with $\sigma^{l_i} \leq \sigma_i \leq \sigma^{l_i+1}$, $i = 1, \ldots, d$, find $\tilde{\alpha} \in S_\alpha^m$, $\tilde{\mu}_{ji} \in S_{\mu^x}^m$, and $\tilde{\sigma}_i = \sigma_{l_i} \in S_\sigma$ such that for all $j = 1, \ldots, m$ and $i = 1, \ldots, d$

$$|\mu_{ji} - \tilde{\mu}_{ji}| \leq \frac{\underline{\sigma}\epsilon}{12d}, \ \sum_j |\alpha_j - \tilde{\alpha}_j| \leq \frac{\epsilon}{3}, \ \frac{|\sigma_i - \tilde{\sigma}_i|}{\tilde{\sigma}_i} \leq \frac{\epsilon}{12d}.$$

By Lemma 6, $d_{TV}(p(\cdot|\theta, m), p(\cdot|\tilde{\theta}, m)) \leq d_{TV}(f(\cdot|\theta, m), f(\cdot|\tilde{\theta}, m))$. Similarly to the proof of Proposition 3.1 in Norets and Pelenis (2014) or Theorem 4.1 in Norets and Pati (2017),

$$d_{TV}(f(\cdot|\theta, m), f(\cdot|\tilde{\theta}, m)) \leq \sum_j |\alpha_j - \tilde{\alpha}_j| + 2 \max_{j=1,\ldots,m} ||\phi(\cdot; \mu_j, \sigma) - \phi(\cdot; \tilde{\mu}_j, \tilde{\sigma}||_1$$

$$\leq \epsilon/3 + 4 \sum_{i=1}^d \left\{ \frac{|\mu_{ji} - \tilde{\mu}_{ji}|}{\min(\sigma_i, \tilde{\sigma}_i)} + \frac{|\sigma_i - \tilde{\sigma}_i|}{\min(\sigma_i, \tilde{\sigma}_i)} \right\} \leq \epsilon.$$

This concludes the proof for the covering number.

The proof of the upper bound on $\Pi(\mathcal{F}^c)$ is the same as the corresponding proof of Theorem 4.1 in Norets and Pati (2017), except here the coordinate specific scale parameters and slightly different notation for the prior tail condition (8) lead to dimension d appearing in front of some of the terms in the bound.

Q.E.D.

LEMMA 20 *Consider* $\epsilon_n = (N_J/n)^{\beta_{Jc}/(2\beta_{Jc}+1)}(\log n)^{t_J}$ *and* $\tilde{\epsilon}_n = (N_J/n)^{\beta_{Jc}/(2\beta_{Jc}+1)}(\log n)^{\tilde{t}_J}$ *with* $t_J > \tilde{t}_J + \max\{0, (1-\tau_1)/2\}$ *and* $\tilde{t}_J > t_{J0}$, *where* t_{J0} *is defined in* (25). *Define* $\mathcal{F}_n$ *as in* (73) *with* $\epsilon = \epsilon_n, H = n\epsilon_n^2/(\log n)$, $\underline{\alpha} = e^{-nH}$, $\underline{\sigma} = n^{-1/(2a_3)}$, $\overline{\sigma} = e^n$, *and* $\overline{\mu} = n^{1/\tau_3}$. *Then, for some constants* $c_1, c_3 > 0$ *and every* $c_2 > 0$, $\mathcal{F}_n$ *satisfies* (28) *and* (29) *for all large* n.

PROOF: From Lemma 19,

$$\log M_e(\epsilon_n, \mathcal{F}_n, \rho) \leq c_1 H \log n = c_1 n \epsilon_n^2.$$

Also,

$$\Pi(\mathcal{F}_n^c) \leq H^2 \exp\{-a_{13} n\} + \exp\{-a_{10} H (\log H)^{\tau_1}\}$$

$$+ a_1 \exp\{-a_2 n\} + a_4 \exp\{-2a_5 n\}.$$

Hence, $\Pi(\mathcal{F}_n^c) \leq e^{-(c_2+4)n\tilde{\epsilon}_n^2}$ for any c_2 if $\epsilon_n^2 (\log n)^{\tau_1 - 1}/\tilde{\epsilon}_n^2 \to \infty$, which holds for $t_J > \tilde{t}_J + \max\{0, (1-\tau_1)/2\}$.

Q.E.D.