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

By Andriy Norets[‡] and Justinas Pelenis[§]

Brown University and Vienna Institute for Advanced Studies

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 in the total variation distance. 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 optimal adaptive estimation of mixed discrete-continuous distributions.

1. Introduction. 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 (2017)) 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).

*First version: December 2017, current version: June 19, 2018.

†We thank participants of Harvard-MIT econometrics workshop, OBayes 2017, CIREQ 2018, and SBIES 2018 for helpful comments.

‡Associate Professor, Department of Economics, Brown University

§Assistant Professor, Vienna Institute for Advanced Studies

Keywords and phrases: Bayesian nonparametrics, adaptive rates, minimax rates, posterior contraction, discrete-continuous distribution, mixed scale, mixtures of normal distributions, latent variables.

In most applications, data contain both continuous and discrete variables. 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. The question we address in the present paper is whether a mixture of normal model that uses latent variables for modeling the discrete part of the distribution can deliver (near) optimal and adaptive posterior contraction rates for nonparametric estimation of discrete-continuous distributions.

Our contribution has two main parts. First, we derive lower bounds on the estimation rate for mixed multivariate discrete-continuous distributions under anisotropic smoothness conditions and potentially growing support of the discrete part of the distribution. Second, we study the posterior contraction rate for a mixture of normals model with a variable number of components 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 models can be used for (up to a log factor) optimal adaptive estimation of mixed discrete-continuous distributions. These results are obtained in a rich 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, smoothing is beneficial only for a subset of discrete variables with a quickly growing number of support points and/or high level of smoothness. In a sense, this subset is automatically and correctly selected by the mixture model. The obtained optimal posterior contraction rates are adaptive since the priors we consider do not depend on the number of support points and the smoothness of the data generating process.

Our results on lower bounds have independent value outside of the literature on Bayesian mixture models and their frequentist properties. Let us briefly review most relevant results on

lower bounds and place our results in that context. 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 shows that 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 tile of the present 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 potentially growing supports for discrete variables.

Some details of our settings and assumptions differ from those in Hall and Titterington (1987) and Efromovich (2011) because our original motivation was in understanding the behavior of the posterior in mixture models with latent variables. Specifically, we consider lower and upper estimation bounds in the total variation distance since posterior concentration in nonparametric settings is much better understood when the total variation distance is considered (Ghosal et al. (2000)). We also introduce a new definition of anisotropic smoothness that, on the one

hand, accommodates an extension of techniques for deriving lower bounds from Ibragimov and Hasminskii (1984) and, on the other hand, lets us exploit approximation results for mixtures of multivariate normal distributions developed by Shen et al. (2013).

The rest of the paper is organized as follows. In Section 2, we describe our framework and define notation. Section 3 presents our results on lower bounds for estimation rates. The results on the posterior contraction rates are given in Section 4. Appendix contains auxiliary results and some proofs.

2. Preliminaries and Notation. 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 $y = (y_1, \ldots, y_{d_y}) \in \mathcal{Y}$, let $A_y = \prod_{j=1}^{d_y} A_{y_j}$, where

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

and let us represent the data generating density-probability mass function as

$$p_0(y, x) = \int_{A_y} f_0(\tilde{y}, x) d\tilde{y}, \tag{2.1}$$

where f_0 belongs to $\mathcal{D}$, the set of probability density functions on $\mathbb{R}^{d_x + d_y}$ with respect to the Lebesgue measure. The representation of a mixed discrete-continuous distribution in (2.1) 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, $y \in \mathcal{Y}$.

In this paper, we consider 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 the generality of the representation in (2.1) can be lost when assumptions such as smoothness are imposed on f_0. Nevertheless, in what follows we do

impose a smoothness assumption on f_0. The interpretation of this assumption is that the values of discrete variables can be ordered and that borrowing of information from nearby discrete points can be useful in estimation.

To get more refined results, we allow N_j's to grow at different rates for different j's. For the same reason, we work with anisotropic smoothness. Let $\mathbb{Z}_+$ denote the set of non-negative integers. For smoothness coefficients $\beta_i > 0$, $i = 1, \ldots, d$, $d = d_x + d_y$, and an envelope function $L : \mathbb{R}^{2d} \to \mathbb{R}$, an anisotropic $(\beta_1, \ldots, \beta_d)$-Holder class, $\mathcal{C}^{\beta_1,\ldots,\beta_d,L}$, is defined as follows.

DEFINITION 2.1. *$f \in \mathcal{C}^{\beta_1,\ldots,\beta_d,L}$ if for any $k = (k_1, \ldots, k_d) \in \mathbb{Z}_+^d$, $\sum_{i=1}^d k_i/\beta_i < 1$, mixed partial derivative of order k, $D^k f$, is finite and*

$$|D^k f(z + \Delta z) - D^k f(z)| \leq L(z, \Delta z) \sum_{j=1}^{d} |\Delta z_j|^{\beta_j (1 - \sum_{i=1}^d k_i/\beta_i)}, \tag{2.2}$$

where $\Delta z_j = 0$ when $\sum_{i=1}^d k_i/\beta_i + 1/\beta_j < 1$.

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 ($\sum_{i=1}^d k_i/\beta_i < 1$ but $\sum_{i=1}^d k_i/\beta_i + 1/\beta_j \geq 1$). This definition slightly differs from definitions available in the literature on anisotropic smoothness that we found. Section 13.2 in Schumaker (2007) presents some 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 $|\Delta z_j|^{\beta_j(1 - \sum k_i/\beta_i)}$ in (2.2). 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 $\sum_{i=1}^d k_i/\beta + 1/\beta \geq 1$, $\beta_j(1 - \sum k_i/\beta_i) = \beta - \lfloor \beta \rfloor$, where $\lfloor \beta \rfloor$ is the largest integer that is strictly smaller than β, 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. Upper bounds on posterior contraction rates are derived under more general assumptions

on L as in Shen et al. (2013).

Some extra notation: for a multi-index $k = (k_1, \ldots, k_d) \in \mathbb{Z}_+^d$, $k! = \prod_{i=1}^d k_i!$, and for $z \in \mathbb{R}^d$, $z^k = \prod_{i=1}^d z_i^{k_i}$. The m-dimensional simplex is denoted by Δ^{m-1}. I_d stands for the $d \times d$ identity matrix. Let $\phi_{\mu,\sigma}(\cdot)$ and $\phi(\cdot; \mu, \sigma)$ denote a multivariate normal density with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\sigma^2 I_d$ (or a diagonal matrix with squared elements of σ on the diagonal when σ is a d-vector). For $z \in \mathbb{R}^d$ and $J \subset \{1, 2, \ldots, d\}$, z_J denotes sub-vector $\{z_i,\, i \in J\}$. Operator "$\lesssim$" denotes less or equal up to a multiplicative positive constant relation.

3. Lower Bounds on Estimation Rates. Let $\mathcal{A}$ denote a collection of all subsets of indices for discrete coordinates $\{1, \ldots, d_y\}$. For $J \in \mathcal{A}$, define $J^c = \{1, \ldots, d\} \setminus J$,

$$N_J = \prod_{i \in J} N_i, \qquad \beta_{J^c} = \left[\sum_{i \in J^c} \beta_i^{-1} \right]^{-1},$$

$N_\emptyset = 1$, $\beta_\emptyset = \infty$, and $\beta_\emptyset / (2\beta_\emptyset + 1) = 1/2$.

For a class of probability distributions $\mathcal{P}$, ζ is said to be a lower bound on the estimation error in metric ρ if

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

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

$$\mathcal{P} = \left\{ p :\ p(y,x) = \int_{A_y} f(\tilde{y}, x) d\tilde{y},\ f \in \mathcal{C}^{\beta_1, \ldots, \beta_d, L} \cap \mathcal{D} \right\}. \tag{3.1}$$

THEOREM 3.1. *For $\mathcal{P}$ defined in (3.1),*

$$\Gamma_n = \min_{J \in \mathcal{A}} \left[\frac{N_J}{n} \right]^{\frac{\beta_{J^c}}{2\beta_{J^c}+1}} = \left[\frac{N_{J_*}}{n} \right]^{\frac{\beta_{J_*^c}}{2\beta_{J_*^c}+1}} \tag{3.2}$$

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

One could recognize expression $[N_J/n]^{\frac{\beta_{J^c}}{2\beta_{J^c}+1}}$ in (3.2) 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 $[N_J/n]^{\frac{\beta_{J^c}}{2\beta_{J^c}+1}}$. 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 (3.2) should, in a sense, optimally choose the subset of discrete variables over which to perform smoothing.

We set up the notation and an outline of the proof of Theorem 3.1 below and delegate detailed calculations to lemmas in Appendix 5.1. The proof of the theorem is based on a general theorem from the literature on lower bounds, which we present next in a slightly simplified form.

LEMMA 3.1. *(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*

$$\sum_{j=1}^{M} KL(Q_j^n, Q_0^n)/M < \log(M)/8, \tag{3.3}$$

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 3.2. *(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 1)

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

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

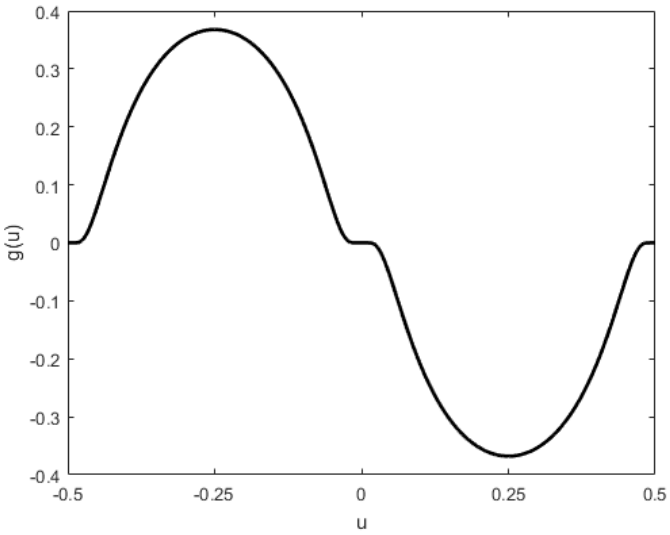

FIG 1. *Function g for $c_0 = 1$.*

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 5.5.

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 (3.2) 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 $c^r =$

$(c_1^r, \ldots, c_d^r)$, $c_i^r = h_i(k_{ir} - 1/2)$, $k_{ir} \in \{1, \ldots, m_i\}$. For $z \in \mathbb{R}^d$ and $r = 1, \ldots, \bar{m}$, 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

$$q_j(y, x) = \int_{A_y} \left[1_{[0,1]^d}(\tilde{y}, x) + \sum_{r=1}^{\bar{m}} w_r^j g_r(\tilde{y}, x) \right] d\tilde{y}, \tag{3.4}$$

where $w_r^j \in \{0, 1\}$, $j = 0, \ldots, M$, and M are defined in Lemma 3.2.

The rest of the proof is delegated to lemmas in Appendix 5.1, which show that q_k in (3.4) satisfy the sufficient conditions from Lemma 3.1. Specifically, Lemma 5.1 derives the lower bound on the total variation distance. Lemma 5.2 verifies condition (3.3) when $\bar{m} \geq 8$. Lemma 5.3, part (i) of Lemma 5.5, and the fact that q_k's are defined on $[0, 1]^d$ imply $q_j \in \mathcal{C}^{\beta_1, \ldots, \beta_d, L}$, $j = 0, \ldots, M$.

This argument (Lemma 5.2 specifically) requires $\bar{m} \geq 8$ as it relies on Lemma 3.2. 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.

Finally, note that we prove the lower bound results for a class of densities that includes densities that are in $\mathcal{C}^{\beta_1, \ldots, \beta_d, L}$ on $[0, 1]^d$. It is straightforward to modify the proof so that it works for a class of smooth densities on $\mathbb{R}^d$. To accomplish this we can replace $1_{[0,1]^d}(\cdot)$ in (3.4) with a smooth function on $\mathbb{R}^d$ that has a bounded support and is bounded away from zero on $[0, 1]^d$, for example,

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

multiplied by a normalization constant, where $IK_0(z_i) = \int_{-1}^{z_i} K_0(u)du / \int_{-1}^{1} K_0(u)du$. Then, proofs of Lemmas 5.1-5.3 go through with minor modifications.

4. Posterior Contraction Rates for a Mixture of Normals Model.

4.1. *Model and Prior.* In this section, we consider a Bayesian model for the data generating process in (2.1). We use a mixture of normal distributions with a variable number of components for modelling the joint distribution of $(\tilde{y}, x)$,

$$f(\tilde{y}, x|\theta, m) = \sum_{j=1}^{m} \alpha_j \phi(\tilde{y}, x; \mu_j, \sigma)$$
$$p(y, x|\theta, m) = \int_{A_y} f(\tilde{y}, x|\theta, m) d\tilde{y}, \tag{4.1}$$

where $\theta = (\mu_j^y, \mu_j^x, \alpha_j, j = 1, 2, \ldots, m; \sigma)$.

We assume the following conditions on the prior Π for (θ, m). For positive constants $a_1, a_2, \ldots, a_9$, for each $i \in \{1, \ldots, d\}$ the prior for σ_i satisfies

$$\Pi(\sigma_i^{-2} \geq s) \leq a_1 \exp\{-a_2 s^{a_3}\} \quad \text{for all sufficiently large } s > 0 \tag{4.2}$$
$$\Pi(\sigma_i^{-2} < s) \leq a_4 s^{a_5} \quad \text{for all sufficiently small } s > 0 \tag{4.3}$$
$$\Pi\{s < \sigma_i^{-2} < s(1+t)\} \geq a_6 s^{a_7} t^{a_8} \exp\{-a_9 s^{1/2}\}, \quad s > 0, \quad t \in (0, 1). \tag{4.4}$$

An example of a prior that satisfies (4.2)-(4.4) is the inverse Gamma prior for σ_i.

Prior for $(\alpha_1, \ldots, \alpha_m)$ conditional on m is Dirichlet$(a/m, \ldots, a/m)$, $a > 0$. Prior for the number of mixture components m is

$$\Pi(m = i) \propto \exp(-a_{10} i (\log i)^{\tau_1}), i = 2, 3, \ldots, \quad a_{10} > 0, \tau_1 \geq 0. \tag{4.5}$$

More generally, a prior that can be bounded above and below by functions in the form of the right hand side of (4.5), possibly with different constants, would also work.

A priori, the components of μ_j, $\mu_{j,i}$, $i = 1, \ldots, d$ are independent from each other, other parameters, and across j. Prior density for $\mu_{j,i}$ is bounded below for some $a_{12}, \tau_2 > 0$ by

$$a_{11} \exp(-a_{12}|\mu_{j,i}|^{\tau_2}), \tag{4.6}$$

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

$$\Pi(\mu_{j,i} \notin [-\mu, \mu]) \leq \exp(-a_{13}\mu^{\tau_3}). \tag{4.7}$$

4.2. *Assumptions on the Data Generating Process.* 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 $\left[\frac{N_J}{n}\right]^{\frac{\beta_{Jc}}{2\beta_{Jc}+1}}$ 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 (3.2) up to a log factor.

Without a loss of generality, let $J = \{1, \ldots, d_J\}$, $I = \{d_J + 1, \ldots, d_y\}$, $J^c = \{1, \ldots, d\} \setminus J$, 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 4.1. *There are positive finite constants $b, \bar{f}_0, \tau$ such that for any $y_J \in \mathcal{Y}_J$ and $\tilde{x} \in \tilde{\mathcal{X}}$*

$$
f_{0|J}(\tilde{x}|y_J) \leq \bar{f}_0 \exp\left(-b||\tilde{x}||^\tau\right). \tag{4.8}
$$

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

ASSUMPTION 4.2. *There exists a positive and finite $\bar{y}$ such that for any $(y_I, y_J) \in \mathcal{Y}$ and $x \in \mathcal{X}$*

$$
\int_{A_{y_I} \cap \{||\tilde{y}_I|| \leq \bar{y}\}} f_{0|J}(\tilde{y}_I, x|y_J) d\tilde{y}_I \geq \int_{A_{y_I} \cap \{||\tilde{y}_I|| > \bar{y}\}} f_{0|J}(\tilde{y}_I, x|y_J) d\tilde{y}_I. \tag{4.9}
$$

This assumption always holds for $A_{y_I} \subset [0,1]^{d_{J^c} - d_x}$. 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 4.3. *We assume that*

$$
f_{0|J} \in \mathcal{C}^{\beta_{d_J+1}, \ldots, \beta_d, L}, \tag{4.10}
$$

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

$$L(\tilde{x}, \Delta\tilde{x}) = \tilde{L}(\tilde{x}) \exp\left\{\tau_0 ||\Delta\tilde{x}||^2\right\}, \tag{4.11}$$

$$\tilde{L}(\tilde{x} + \Delta\tilde{x}) \leq \tilde{L}(\tilde{x}) \exp\left\{\tau_0 ||\Delta\tilde{x}||^2\right\}. \tag{4.12}$$

The smoothness assumption (4.10) 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 5.8.

Assumption 4.4. *There are positive finite constants* ε *and* $\bar{F}$*, such that for any* $y_J \in \mathcal{Y}_J$ *and* $k = \{k_i\}_{i \in J^c} \in \mathbb{N}_0^{d_{J^c}}$, $\sum_{i \in J^c} k_i/\beta_i < 1$,

$$\int \left[\frac{|D^k f_{0|J}(\tilde{x}|y_J)|}{f_{0|J}(\tilde{x}|y_J)}\right]^{\frac{(2+\varepsilon\beta_{J^c}^{-1} d_{J^c}^{-1})}{\sum_{i \in J^c} k_i/\beta_i}} f_{0|J}(\tilde{x}|y_J) d\tilde{x} < \bar{F}, \tag{4.13}$$

$$\int \left[\frac{\tilde{L}(\tilde{x})}{f_{0|J}(\tilde{x}|y_J)}\right]^{2+\varepsilon\beta_{J^c}^{-1} d_{J^c}^{-1}} f_{0|J}(\tilde{x}|y_J) d\tilde{x} < \bar{F}. \tag{4.14}$$

The envelope function and restrictions on its behaviour are mostly relevant for the case of unbounded support. Condition (4.14) suggests that the envelope function $\tilde{L}$ should be comparable to $f_{0|J}$.

Assumption 4.5. *For some small* $\nu > 0$,

$$N_J = o(n^{1-\nu}). \tag{4.15}$$

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 $\log n$.

4.3. *Posterior Contraction Rates.* Let

$$t_{J0} = \begin{cases} \frac{d_{J^c}[1+1/(\beta_{J^c} d_{J^c})+1/\tau]+\max\{\tau_1,1,\tau_2/\tau\}}{2+1/\beta_{J^c}} & \text{if } J^c \neq \emptyset \\ \max\{\tau_1, 1\}/2 & \text{if } J^c = \emptyset \end{cases} \tag{4.16}$$

where (τ, τ_1, τ_2) are defined in Sections 4.1-4.2.

Theorem 4.1. *Suppose the assumptions from Sections 4.1-4.2 hold for a given* $J \in \mathcal{A}$*. Let*

$$\epsilon_n = \left[\frac{N_J}{n}\right]^{\beta_{J^c}/(2\beta_{J^c}+1)} (\log n)^{t_J}, \tag{4.17}$$

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 3, 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 (4.17).

COROLLARY 4.1. *Suppose the assumptions from Sections 4.1-4.2 hold for every* $J \in \mathcal{A}$. *Let*

$$\epsilon_n = \min_{J \in \mathcal{A}} \left[\frac{N_J}{n}\right]^{\beta_{J^c}/(2\beta_{J^c}+1)} (\log n)^{t_J}, \tag{4.18}$$

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 4.1 delivers a valid upper bound on the posterior contraction rate for every $J \in \mathcal{A}$ including the one for which the minimum in (4.18) is attained. Hence, the corollary is an immediate implication of Theorem 4.1 whose proof is presented in the following section.

The results on lower bounds in Section 3 hold for any class of data generating densities that includes f_0 satisfying the following conditions: $f_0 \in \mathcal{C}^{\beta_1,\ldots,\beta_d,L}$, $f_0 = 0$ outside $[0,1]^d$, and $\overline{f} \geq f_0 \geq \underline{f} > 0$, where L, $\overline{f}$, and $\underline{f}$ are finite positive constants. It is worth pointing out that these conditions imply Assumptions 4.1-4.4, which combined with Assumption 4.5 for N_{J_*} and a prior specified in Section 4.1 would deliver the sufficient conditions of Corollary 4.1.

4.4. *Proof of Posterior Contraction Results.* To prove Theorem 4.1, 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 $M_e(\epsilon_n, \mathcal{F}_n, \rho)$

$$\log M_e(\epsilon_n, \mathcal{F}_n, \rho) \leq c_1 n\epsilon_n^2, \tag{4.19}$$

$$\Pi(\mathcal{F}_n^c) \leq c_3 \exp\{-(c_2+4)n\tilde{\epsilon}_n^2\}. \tag{4.20}$$

Suppose also that the prior thickness condition holds

$$\Pi(\mathcal{K}(p_0, \tilde{\epsilon}_n)) \geq c_4 \exp\{-c_2 n\tilde{\epsilon}_n^2\}, \tag{4.21}$$

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

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

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 verification of conditions (4.19) and (4.20) 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 5.18 in Appendix 5.2.2. Verification of the prior thickness condition is more involved and we formulate it as a separate result in the following theorem.

THEOREM 4.2. *Suppose the assumptions from Sections 4.1-4.2 hold for a given $J \in \mathcal{A}$. Let $t_J > t_{J0}$, where t_{J0} is defined in (4.16), and*

$$\tilde{\epsilon}_n = \left[\frac{N_J}{n} \right]^{\beta_{Jc}/(2\beta_{Jc}+1)} (\log n)^{t_J}. \tag{4.22}$$

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

$$\Pi(\mathcal{K}(p_0, \tilde{\epsilon}_n)) \geq \exp\{-Cn\tilde{\epsilon}_n^2\}. \tag{4.23}$$

Approximation results are key for showing the prior thickness condition (4.23). 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

$$f^\star_{|J}(\tilde{x}|y_J) = \sum_{j=1}^{K} \alpha^\star_{j|y_J} \phi(\tilde{x}; \mu^\star_{j|y_J}, \sigma^\star_{J^c}), \tag{4.24}$$

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 $\sigma^\star_J$. The following subsection presents proof details.

4.4.1. *Proof of Theorem 4.2 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 (4.13)-(4.14), b and τ defined in (4.8), 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 4.2.

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 $i \in J^c\}$,

$$d_H\left(f_{0|J}(\cdot|k), f^\star_{|J}(\cdot|k)\right) \leq C_2\sigma_n^\beta, \tag{4.25}$$

where $f^\star_{|J}$ is defined in (4.24). Constant C_2 is the same for all $k \in \mathcal{Y}_J$ since all the bounds on $f_{0|J}$ assumed in Section 4.2 are uniform over k.

Note also that our smoothness definition is different from the one used by Shen et al. (2013). In Lemmas 5.6 and 5.7 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 5.10, 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 $y_J \in \mathcal{Y}_J$

$$F_{0|J}\left(||\tilde{X}|| > a_{\sigma_n}|y_J\right) \leq B_0\sigma_n^{4\beta+2\varepsilon}\underline{\sigma}_n^8, \tag{4.26}$$

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:

$$\theta^\star = \left\{\{\mu_1^\star, \ldots, \mu_m^\star\} = \left\{(k, \mu^\star_{j|k}),\ j = 1, \ldots, K,\ k \in \mathcal{Y}_J\right\},\right.$$
$$\left.\{\alpha_1^\star, \ldots, \alpha_m^\star\} = \left\{\alpha^\star_{jk} = \alpha^\star_{j|k}\pi_{0J}(k),\ j = 1, \ldots, K,\ k \in \mathcal{Y}_J\right\},\right.$$

$$\sigma_J^{\star 2} = \{\sigma_i^{\star 2} = 1/[64 N_i^2 \beta \log(1/\sigma_n)],\ i \in J\}$$
$$\sigma_{J^c}^{\star} = \{\sigma_i^{\star} = \sigma_n^{\beta/\beta_i},\ i \in J^c\},\Big\}$$

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

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 5.9. 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 (4.24), 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 (4.25) and Lemma 5.4 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 $d_H^2(p_{|J}^\star(\cdot|\cdot)\pi_{0J}(\cdot), p(\cdot,\cdot|\theta,m))$ from Lemma 5.11 we obtain

$$d_H^2(p_0(\cdot,\cdot), p(\cdot,\cdot|\theta,m)) \lesssim \sigma_n^{2\beta}. \tag{4.27}$$

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 5.14 in the Appendix 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 $\|x\| \le a_{\sigma_n}$,

$$\frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} \ge C_3 \frac{\sigma_n^{2\beta}}{2m^2} \equiv \lambda_n, \tag{4.28}$$

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

$$\frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} \ge \exp\left\{-\frac{8||x||^2}{\underline{\sigma}_n^2} - C_4 \log n\right\}, \tag{4.29}$$

for some constant $C_4 > 0$. Consider all sufficiently large n such that $\lambda_n < e^{-1}$ and (4.28) and (4.29) hold. Then, for any $\theta \in S_{\theta^\star}$,

$$\begin{aligned}
&\sum_{y\in\mathcal{Y}} \int_{\mathcal{X}} \left(\log \frac{p_0(y_J, y_I, x)}{p(y_J, y_I, x|\theta, m)}\right)^2 \mathbf{1}\left\{\frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} < \lambda_n\right\} p_0(y_J, y_I, x)dx \\
&= \sum_{y\in\mathcal{Y}} \int_{\tilde{\mathcal{X}}} \left(\log \frac{p_0(y_J, y_I, x)}{p(y_J, y_I, x|\theta, m)}\right)^2 \mathbf{1}\left\{\frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} < \lambda_n\right\} \mathbf{1}\left\{\tilde{y}_I \in A_{y_I}\right\} f_{0J}(y_J, \tilde{x})d\tilde{x} \\
&= \sum_{y\in\mathcal{Y}} \int_{\tilde{\mathcal{X}}} \left(\log \frac{p_0(y_J, y_I, x)}{p(y_J, y_I, x|\theta, m)}\right)^2 \mathbf{1}\left\{\frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} < \lambda_n, ||x|| > a_{\sigma_n}, \tilde{y}_I \in A_{y_I}\right\} f_{0J}(y_J, \tilde{x})d\tilde{x} \\
&\le \sum_{y\in\mathcal{Y}} \int_{\{\tilde{x}:||x||>a_{\sigma_n}\}} \left(\log \frac{p_0(y_J, y_I, x)}{p(y_J, y_I, x|\theta, m)}\right)^2 \mathbf{1}\left\{\tilde{y}_I \in A_{y_I}\right\} f_{0J}(y_J, \tilde{x})d\tilde{x} \\
&\le \sum_{y\in\mathcal{Y}} \int_{\{\tilde{x}:||x||>a_{\sigma_n}\}} \left[\frac{128}{\underline{\sigma}_n^4}||x||^4 + 2(C_4 \log n)^2\right] f_{0|J}(\tilde{x}|y_J)\mathbf{1}\left\{\tilde{y}_I \in A_{y_I}\right\} d\tilde{x}\pi_{0J}(y_J) \\
&\le \sum_{y_J\in\mathcal{Y}_J} \int_{\{\tilde{x}:||\tilde{x}||>a_{\sigma_n}\}} \left[\frac{128}{\underline{\sigma}_n^4}||\tilde{x}||^4 + 2(C_4 \log n)^2\right] f_{0|J}(\tilde{x}|y_J)d\tilde{x}\pi_{0J}(y_J) \\
&\le \frac{128}{\underline{\sigma}_n^4} \sum_{y_J\in\mathcal{Y}_J} E_{0|y_J}\left(\left\|\tilde{X}\right\|^8\right)^{1/2} \left(F_{0|y_J}\left(\left\|\tilde{X}\right\| > a_{\sigma_n}\right)\right)^{1/2} \pi_{0J}(y_J) + 2(C_4 \log n)^2 B_0 \sigma_n^{4\beta+2\varepsilon} \underline{\sigma}_n^8 \\
&\le C_5 \sigma_n^{2\beta+\varepsilon}
\end{aligned} \tag{4.30}$$

for some constant $C_5 > 0$ and all sufficiently large n, where the last inequality holds by the tail condition in (4.8), (4.26), and $(\log n)^2 \sigma_n^{2\beta+\varepsilon} \underline{\sigma}_n^8 \to 0$.

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

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

and, therefore,

$$\sum_{y\in\mathcal{Y}} \int_{\mathcal{X}} \log \frac{p_0(y_J, y_I, x)}{p(y_J, y_I, x|\theta, m)} \mathbf{1}\left\{\frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} < \lambda_n\right\} p_0(y_J, y_I, x)dx \le C_5 \sigma_n^{2\beta+\varepsilon}. \tag{4.31}$$

Inequalities (4.27), (4.30), and (4.31) combined with Lemma 5.9 imply

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

for any $\theta\in S_{\theta^\star}$, $m=N_JK$, and some positive constant A (details are provided in Lemma 5.15 in the Appendix).

By Lemma 5.16 in the Appendix for all sufficiently large n, $s=1+1/\beta+1/\tau$, and some $C_6>0$,

$$\Pi(\mathcal{K}(p_0,\tilde\epsilon_n))\ge\Pi(m=N_JK,\theta\in S_{\theta^\star})\ge\exp\left[-C_6N_J\tilde\epsilon_n^{-d_{J^c}/\beta}\{\log(n)\}^{d_{J^c}s+\max\{\tau_1,1,\tau_2/\tau\}}\right].$$

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.

4.4.2. *Proof of Theorem 4.2 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

$$\theta^\star=\Big\{\{\mu_1^\star,\dots,\mu_m^\star\}=\{k,\ k\in\mathcal{Y}_J\},$$
$$\{\alpha_1^\star,\dots,\alpha_m^\star\}=\{\alpha_k^\star,\ k\in\mathcal{Y}_J\}=\{\pi_0(k)\}_{k\in\mathcal{Y}_J},$$
$$\sigma^{\star2}=\{\sigma_i^{\star2}=\frac{1}{64N_i^2\beta\log(1/\sigma_n)},\ i\in J\}\Big\},$$

$$S_{\theta^\star}=\Big\{\{\mu_1,\dots,\mu_m\}=\{\mu_k,\ k\in\mathcal{Y}_J\},\ \mu_{k,i}\in\left[k_i-\frac{1}{4N_i},k_i+\frac{1}{4N_i}\right],\ i=1,\dots,d_J,$$
$$\sigma=\{\sigma_i\in(0,\sigma_i^\star),\ i\in J\},$$
$$\{\alpha_j,\ j=1,\dots,m\}=\{\alpha_k,\ k\in\mathcal{Y}_J\}\in\Delta^{m-1},$$
$$\sum_{k\in\mathcal{Y}_J}|\alpha_k-\alpha_k^\star|\le2\sigma_n^{2\beta},\qquad\min_{k\in\mathcal{Y}_J}\alpha_k\ge\frac{\sigma_n^{2\beta}}{2m^2}\Big\}.$$

For $m=N_J$ and $\theta\in S_{\theta^\star}$, a simplification of the proof of Lemma 5.11 delivers

$$d_H^2(p_0(\cdot),p(\cdot|\theta,m))\le2\max_{k\in\mathcal{Y}_J}\int_{A_k^c}\phi(\tilde y_J;\mu_k,\sigma)d\tilde y_J+\sum_{k\in\mathcal{Y}_J}|\alpha_k^\star-\alpha_k|\lesssim\sigma_n^{2\beta}.$$

A simplification of derivations in Lemma 5.14 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 $\theta \in S_{\theta^\star}$

$$\sum_{y_J \in \mathcal{Y}_J} \left(\log \frac{p_0(y_J)}{p(y_J|\theta,m)} \right)^2 \mathbf{1}\left\{ \frac{p(y_J|\theta,m)}{p_0(y_J)} < \lambda_n \right\} p_0(y_J) = 0$$
$$\sum_{y_J \in \mathcal{Y}_J} \left(\log \frac{p_0(y_J)}{p(y_J|\theta,m)} \right) \mathbf{1}\left\{ \frac{p(y_J|\theta,m)}{p_0(y_J)} < \lambda_n \right\} p_0(y_J) = 0 \tag{4.32}$$

as $\frac{p(y_J|\theta,m)}{p_0(y_J)} \geq \lambda_n$ for all $y_J \in \mathcal{Y}_J$. As $\lambda_n \to 0$, by Lemma 5.9 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 $C_7 \log(1/\lambda_n)^2 \sigma_n^{2\beta} \leq A\tilde{\epsilon}_n^2$ for some constant A. By the simplification of Lemma 5.16 for this particular case for all sufficiently large n and some $C_8 > 0$,

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

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.

5. Future Work. It seems feasible to extend the results of this paper to conditional density estimation by covariate dependent mixtures along the lines of Norets and Pati (2017). We leave this to future work.

References.

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

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.

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. 232246.

Burman, P. (1987): "Smoothing Sparse Contingency Tables," *Sankhy?: 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'conomie et de Statistique*, 211–242.

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

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.

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.

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.

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

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.

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.

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.

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. (2017): "Optimal Auxiliary Priors and Reversible Jump Proposals for a Class of Variable Dimension

Models," Unpublished manuscript, Brown University.

Norets, A. and D. Pati (2017): "Adaptive Bayesian Estimation of Conditional Densities," *Econometric Theory*, 33, 9801012.

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.

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.

Shen, W., S. T. Tokdar, and S. Ghosal (2013): "Adaptive Bayesian multivariate density estimation with Dirichlet mixtures," *Biometrika*, 100, 623–640.

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

Appendix.

5.1. *Proofs and Auxiliary Results for Lower Bounds.*

Lemma 5.1. *For q_j, q_l, $i \neq l$ defined in (3.4), 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

$$g_r(\tilde{y}, x) = 0, \forall r \neq r(\tilde{y}, x). \tag{5.1}$$

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 $i\{1, \ldots, d_y\}$, 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,

$$d_{TV}(q_j, q_l) = \int \int \left| (w^j_{r(\tilde{y},x)} - w^l_{r(\tilde{y},x)}) g_{r(\tilde{y},x)}(\tilde{y}, x) \right| d\tilde{y} dx$$

$$= \sum_{r=1}^{\bar{m}} \int_{B_r} \left| (w^j_{r(z)} - w^l_{r(z)}) g_{r(z)}(z) \right| dz$$
$$= \sum_{r=1}^{\bar{m}} |w^j_r - w^l_r| \int_{B_r} |g_r(z)|\, dz. \tag{5.2}$$

Finally,

$$d_{TV}(q_j, q_l) = \sum_{r=1}^{\bar{m}} 1\{w^j_r \neq w^l_r\} \cdot \Gamma_n \cdot \prod_{i=1}^{d} h_i \cdot \left[\int_{-1/2}^{1/2} |g(u)| du \right]^d \quad \text{(change of variables in (5.2))}$$
$$\geq \Gamma_n \cdot \prod_{i=1}^{d} m_i h_i \cdot \left[\int_{-1/2}^{1/2} |g(u)| du \right]^d /8 \quad \text{(by Lemma 3.2)}$$
$$\geq \Gamma_n \cdot \left[\int_{-1/2}^{1/2} |g(u)| du/2 \right]^d /8 \quad \text{(since } m_i h_i > 1/2\text{)}.$$

□

LEMMA 5.2. *For $\Gamma_n \to 0$ and $\bar{m} \geq 8$ and a sufficiently small c_0 in the definition of g, condition (3.3) in Lemma (3.1) holds for all sufficiently large n.*

PROOF. By Lemma 3.2, it suffices to show that

$$d_{KL}(Q^n_j, Q^n_0) = n \cdot d_{KL}(q_j, q_0) < (\bar{m} \log 2)/64. \tag{5.3}$$

First, note that for a density $f \geq c > 0$ on $[0,1]^d$, $d_{KL}(f, 1_{[0,1]^d})$ is bounded above by

$$d_{KL}(f, 1_{[0,1]^d}) + d_{KL}(1_{[0,1]^d}, f) = \int_{[0,1]} (f-1) \log f \leq \frac{\int_{[0,1]} (f-1)^2}{c}. \tag{5.4}$$

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

$$1_{[0,1]^d}(z) + \sum_{r=1}^{\bar{m}} w^j_r g_r(z) \geq 1 - \Gamma_n \left[\max_{u \in [-1/2, 1/2]} g(u) \right]^d \geq 1/2 \tag{5.5}$$

for all sufficiently large n. Thus,

$$d_{KL}(q_j, q_0) \leq d_{KL}\left(1_{[0,1]^d} + \sum_{r=1}^{\bar{m}} w^j_r g_r, 1_{[0,1]^d} \right) \quad \text{(by (5.14))}$$
$$\leq 2 \int \left[\sum_{r=1}^{\bar{m}} w^j_r g_r(z) \right]^2 dz \quad \text{(by (5.4) and (5.5))}$$
$$= 2 \int \sum_{r=1}^{\bar{m}} w^j_r (g_r(z))^2 dz \quad \text{(since } g_r(z) g_l(z) = 0, \forall r \neq l\text{)}$$
$$\leq 2\bar{m} \int (g_1(z))^2 dz = 2\Gamma_n^2 \prod_i (m_i h_i) \left[\int_{-1/2}^{1/2} g(u)^2 du \right]^d$$

$$\leq 2\Gamma_n^2 \left[\int_{-1/2}^{1/2} g(u)^2 du\right]^d \leq 2\Gamma_n^2 c_0^{2d}. \tag{5.6}$$

Finally,

$$\begin{aligned}
\bar{m} &= \prod_{i=1}^d m_i \geq 2^{-d} \prod_{i=1}^d h_i^{-1} && \text{(by definitions of } \bar{m} \text{ and } m_i\text{)} \\
&= 2^{-d} \prod_{i \in J_*} (N_i/2) \cdot \prod_{i \in J_*^c, i \leq d_y} \left(\Gamma_n^{-\beta_i^{-1}}/\varrho_i\right) \cdot \prod_{i \in J_*^c, i > d_y} \left(\Gamma_n^{-\beta_i^{-1}}\right) && \text{(by definition of } h_i\text{)} \\
&\geq 2^{-d} \prod_{i \in J_*} (N_i/2) \cdot \prod_{i \in J_*^c, i \leq d_y} \left(\Gamma_n^{-\beta_i^{-1}}/2\right) \cdot \prod_{i \in J_*^c, i > d_y} \left(\Gamma_n^{-\beta_i^{-1}}\right) && \text{(by restrictions on } \varrho_i\text{)} \\
&= 2^{-d-d_y} \cdot N_{J_*} \cdot \Gamma_n^{-\beta_{J_*^c}^{-1}} = 2^{-d-d_y} n \Gamma_n^2 \\
&\geq 2^{-d-d_y} n \cdot d_{KL}(q_j, q_0)/(2c_0^{2d}) && \text{(by (5.6))}.
\end{aligned}$$

The last inequality implies (5.3) if

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

□

LEMMA 5.3. *For $j \in \{0, \ldots, M\}$, $q_j \in \mathcal{C}^{\beta_1^*, \ldots, \beta_d^*, L}$ with $L = 1$ for any sufficiently small constant c_0 in the definition of g.*

PROOF. For $j = 0$, the result is trivial. For $j \neq 0$, consider $k = (k_1, \ldots, k_d)$ and $z, \Delta z \in \mathbb{R}^d$ such that for some $i \in \{1, \ldots, d\}$, $\Delta z_i \neq 0$, for any $l \neq i$, $\Delta z_l = 0$, $\sum_{l=1}^d k_l/\beta_l^* < 1$, and $\sum_{l=1}^d k_l/\beta_l^* + 1/\beta_i^* \geq 1$ so that

$$0 \leq \beta_i^* (1 - \sum_{l=1}^d k_l/\beta_l^*) \leq 1. \tag{5.7}$$

For $r(\cdot)$ defined in (5.1),

$$\begin{aligned}
D^k q_j(z) &= 1\{k = (0, \ldots, 0)\} + w_{r(z)} \Gamma_n \prod_{l=1}^d g^{(k_l)}((z_l - c_l^{r(z)})/h_l)/h_l^{k_i} \\
&= 1\{k = (0, \ldots, 0)\} + B_i \cdot w_{r(z)} h_i^{\beta_i^*(1 - \sum_{l=1}^d k_l/\beta_l^*)} \prod_{l=1}^d g^{(k_l)}((z_l - c_l^{r(z)})/h_l),
\end{aligned} \tag{5.8}$$

where $B_i \in \{1, 1/2, \varrho_i^{-\beta_i^*}\} \subset (0, 1]$. In what follows we consider $k \neq (0, \ldots, 0)$ to simplify the notation; when $k = (0, \ldots, 0)$ the argument below goes through as the indicator function

$1\{k=(0,\ldots,0)\}$ is canceled out in the differences of derivatives. From Tsybakov (2008), (2.33)-(2.34), for any sufficiently small c_0 and $s \leq \max_l \beta_l^* + 1$,

$$\max_z |g^{(s)}(z)| \leq 1/8. \tag{5.9}$$

This imply that

$$|g^{(k_i)}((z_i + \Delta z_i - c_i^r)/h_i) - g^{(k_i)}((z_i - c_i^r)/h_i)| \leq |\Delta z_i|/(8h_i). \tag{5.10}$$

First, let us consider the case when $r(z) = r(z+\Delta z)$ and $|\Delta z_i| \leq h_i$. From (5.8), (5.9), and (5.10),

$$\begin{aligned} |D^k q_j(z+\Delta z) - D^k q_j(z)| &\leq h_i^{\beta_i^*(1-\sum_{l=1}^d k_l/\beta_l^*)} 8^{-d} |\Delta z_i/h_i| \\ &= 8^{-d} |\Delta z_i|^{\beta_i^*(1-\sum_{l=1}^d k_l/\beta_l^*)} \left|\frac{\Delta z_i}{h_i}\right|^{1-\beta_i^*(1-\sum_{l=1}^d k_l/\beta_l^*)} \\ &\leq |\Delta z_i|^{\beta_i^*(1-\sum_{l=1}^d k_l/\beta_l^*)}, \end{aligned} \tag{5.11}$$

where the last inequality follows from $\Delta z_i \leq h_i$ and (5.7).

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 (5.10),

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

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

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

If $w_{r(z)} \neq w_{r(z+\Delta z)}$ 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 $\Delta z^\star = (0,\ldots,0,\Delta z_i^\star,0,\ldots,0)$ be such that $z+\Delta z^\star$ is a boundary point of $B_{r(z)}$. Then, $D^k q_j(z+\Delta z^\star) = 0$ and (5.11) imply

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

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 q_j and g

$$|D^k q_j(z+\Delta z) - D^k q_j(z)| = |D^k q_j(z+\Delta z+0.5h_i) - D^k q_j(z+0.5h_i)| \leq |\Delta z_i|^{\beta_i^*(1-\sum_{l=1}^d k_l/\beta_l^*)},$$

where the last inequality follows from (5.11).

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

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

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

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

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

□

LEMMA 5.4. *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,*

$$d_{TV}(p_1, p_2) \leq d_{TV}(f_1, f_2) \tag{5.12}$$

$$d_H(p_1, p_2) \leq d_H(f_1, f_2) \tag{5.13}$$

$$d_{KL}(p_1, p_2) \leq d_{KL}(f_1, f_2). \tag{5.14}$$

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

$$\inf_{\tilde{y} \in A_y} \frac{f_1(\tilde{y}, x)}{f_2(\tilde{y}, x)} \leq \frac{p_1(y, x)}{p_2(y, x)} \leq \sup_{\tilde{y} \in A_y} \frac{f_1(\tilde{y}, x)}{f_2(\tilde{y}, x)}. \tag{5.15}$$

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,

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

$$\leq 2\left(1-\sum_y \int\int 1_{A_y}(\tilde{y})\sqrt{f_1(\tilde{y},x)f_2(\tilde{y},x)}d\lambda(\tilde{y})d\mu(x)\right) = d_H(f_1,f_2).$$

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 (5.14). The last claim follows from

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

□

LEMMA 5.5. *For Γ_n, h_i, ϱ_i, and β_i^* defined in Section 3, (i) $\beta_i^* \geq \beta_i$ for $i = 1,\dots,d$ and (ii) $\varrho_i \in (1,2]$ for $i \in J_*^c \cap \{1,\dots,d_y\}$.*

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 $\Gamma_n^{\beta_i^{-1}} N_i \geq 1$. □

5.2. *Proofs and Auxiliary Results for Posterior Contraction Rates.*

5.2.1. *Prior Thickness.*

LEMMA 5.6. *(Anisotropic Taylor Expansion) For* $f \in \mathcal{C}^{\beta_1,\ldots,\beta_d,L}$ *and* $r \in \{1,\ldots,d\}$

$$f(x_1+y_1,\ldots,x_d+y_d) = \sum_{k\in I^r} \frac{y^k}{k!} D^k f(x_1,\ldots,x_r,x_{r+1}+y_{r+1},\ldots,x_d+y_d) \tag{5.16}$$

$$+\sum_{l=1}^{r}\sum_{k\in \bar{I}^l} \frac{y^k}{k!}\Big(D^k f(x_1,\ldots,x_l+\zeta_l^k,x_{l+1}+y_{l+1},\ldots,x_d+y_d) \tag{5.17}$$

$$- D^k f(x_1,\ldots,x_l,x_{l+1}+y_{l+1},\ldots,x_d+y_d)\Big), \tag{5.18}$$

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

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

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

and the differences in derivatives in (5.17)-(5.18) *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 $r=1$, (5.16)-(5.18) 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 (5.16)-(5.18) holds for some $r \in \{1,\ldots,d\}$. Then, let us show that (5.16)-(5.18) holds for $r+1$. For that, consider a univariate Taylor expansion of $D^k f$ in (5.16). 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 $k_{r+1}^* = \lfloor \beta_{r+1}(1-\sum_{j=1}^r k_j/\beta_j)\rfloor$. Then,

$$\begin{aligned}
&D^k f(x_1,\ldots,x_r,x_{r+1}+y_{r+1},\ldots,x_d+y_d) = \\
&\sum_{k_{r+1}=0}^{k_{r+1}^*} \frac{y_{r+1}^{k_{r+1}}}{k_{r+1}!} D^{k+k_{r+1}\cdot e_{r+1}} f(x_1,\ldots,x_{r+1},x_{r+2}+y_{r+2},\ldots,x_d+y_d) \\
&+ \frac{y_{r+1}^{k_{r+1}^*}}{k_{r+1}^*!}\Big(D^{k+k_{r+1}^*\cdot e_{r+1}} f(x_1,\ldots,x_r,x_{r+1}+\zeta_{r+1}^{k+k_{r+1}^*\cdot e_{r+1}},x_{r+2}+y_{l+2},\ldots,x_d+y_d) \\
&- D^{k+k_{r+1}^*\cdot e_{r+1}} f(x_1,\ldots,x_r,x_{r+1},x_{r+2}+y_{l+2},\ldots,x_d+y_d)\Big).
\end{aligned}$$

Inserting this expansion into (5.16) delivers the result for $r+1$.

□

LEMMA 5.7. *Let $R(x,y)$ denote the remainder term in the anisotropic Taylor expansion ((5.17)-(5.18) for $r=d$). Suppose $f \in \mathcal{C}^{\beta_1,\ldots,\beta_d,L}$ and L satisfies (4.11)-(4.12). 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 $l \in \{1,\ldots,d\}$

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

where we used inequalities (2.2), (4.11), and (4.12) 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.

□

LEMMA 5.8. *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,

$$\begin{aligned}
&D^k f_{0|J}(\tilde{x}+\Delta\tilde{x}|y_J) - D^k f_{0|J}(\tilde{x}|y_J) = \\
&= \frac{1}{\pi_{0J}(y_J)}\int_{A_{y_J}}\left(D^{0,\ldots,0,k}f_0(\tilde{y}_J,\tilde{x}+\Delta\tilde{x}) - D^{0,\ldots,0,k}f_0(\tilde{y}_J,\tilde{x})\right)d\tilde{y}_J
\end{aligned}$$

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

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

LEMMA 5.9. *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). □

LEMMA 5.10. *Under the assumptions and notation of Section 4, 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 $\tilde{x} \in E_{\sigma_n}^*$,

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

As $a_{\sigma_n} > a$ and as $f_{0|J}(\tilde{x}|y_J) \geq \bar{f}_0 \exp\{-b a_{\sigma_n}^\tau\}$, then the tail condition (4.8) 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,

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

$$\leq B_0 \sigma_n^{4\beta+2\varepsilon+8\beta/\beta_{\min}} = B_0 \sigma_n^{4\beta+2\varepsilon} \underline{\sigma}_n^8$$

as desired since $\sigma_n^{\beta/\beta_{\min}} = \underline{\sigma}_n$ and the tail condition on $f_{0|J}(\cdot|y_J)$, (4.8), implies the existence of a $\delta > 0$ small enough such that $E_{0|J}(f_{0|J}^{-\delta}) \leq B_0 < \infty$ for any $y_J \in \mathcal{Y}_J$. □

LEMMA 5.11. *Under the assumptions and notation of Section 4, for $m = KN_J$ and any $\theta \in S_{\theta^\star}$*

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

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,

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

where the first inequality follows from $d_H^2(\cdot,\cdot) \leq d_{TV}(\cdot,\cdot)$, the second inequality holds by Lemma 5.4, 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,

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

where the fact that $\alpha^\star_{j,k} = 0$ for $j > N$ by design is used to get $j \leq N$ rather than $j \leq K$ in the max subscript. The last inequality is proved in Lemma 5.12.

Second,

$$\sum_{y_J\in\mathcal{Y}_J}\int_{\tilde{\mathcal{X}}}\left|\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\alpha^\star_{jk}\mathbf{1}\{k=y_J\}\phi(\tilde{x},\mu_{jk,J^c},\sigma_{J^c})-\alpha_{jk}\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\phi(\tilde{x},\mu_{jk,J^c},\sigma_{J^c})\right|d\tilde{x}$$
$$=\sum_{j=1}^{K}\left(\sum_{y_J\in\mathcal{Y}_J}\left|\sum_{k\in\mathcal{Y}_J}\alpha^\star_{jk}\mathbf{1}\{k=y_J\}-\alpha_{jk}\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\right|\int_{\tilde{\mathcal{X}}}\phi(\tilde{x},\mu_{jk,J^c},\sigma_{J^c})d\tilde{x}\right)$$
$$=\sum_{j=1}^{K}\sum_{y_J\in\mathcal{Y}_J}\left|\sum_{k\in\mathcal{Y}_J}\alpha^\star_{jk}\mathbf{1}\{k=y_J\}-\alpha_{jk}\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\right|$$
$$\le\sum_{y_J\in\mathcal{Y}_J}\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\left|\alpha^\star_{jk}\mathbf{1}\{k=y_J\}-\alpha^\star_{jk}\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\right|$$
$$+\sum_{y_J\in\mathcal{Y}_J}\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\left|\alpha^\star_{jk}\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J-\alpha_{jk}\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\right|$$
$$\le\sum_{y_J\in\mathcal{Y}_J}\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\alpha^\star_{jk}\left|\mathbf{1}\{k=y_J\}-\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\right|$$
$$+\sum_{y_J\in\mathcal{Y}_J}\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\left|\alpha^\star_{jk}-\alpha_{jk}\right|\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J$$
$$=\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\left(\alpha^\star_{jk}\sum_{y_J\in\mathcal{Y}_J}\left|\mathbf{1}\{k=y_J\}-\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\right|\right)$$
$$+\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\left(\left|\alpha^\star_{jk}-\alpha_{jk}\right|\sum_{y_J\in\mathcal{Y}_J}\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\right)$$
$$\le\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\alpha^\star_{jk}\left[\int_{A_k^c}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J+\sum_{y_J\neq k}\int_{A_{y_J}}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J\right]+\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\left|\alpha^\star_{jk}-\alpha_{jk}\right|$$
$$=\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\alpha^\star_{jk}\cdot 2\int_{A_k^c}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J+\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\left|\alpha^\star_{jk}-\alpha_{jk}\right|$$
$$\le 2\max_{j\le N,k\in\mathcal{Y}_J}\int_{A_k^c}\phi(\tilde{y}_J,\mu_{jk,J},\sigma_J)d\tilde{y}_J+\sum_{k\in\mathcal{Y}_J}\sum_{j=1}^{K}\left|\alpha^\star_{jk}-\alpha_{jk}\right|\lesssim\sigma_n^{2\beta}.$$

The last inequality follows from Lemma 5.13 and the definition of $S_{\theta^\star}$.

□

LEMMA 5.12. *Under the assumptions and notation of Section 4,*

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

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.

□

LEMMA 5.13. *Under the assumptions and notation of Section 4, 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 $\tilde{Y}_i \leq (4N_i\sigma_i^y)^{-1}$ and $\theta \leq 0$, therefore:

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

$$= \inf_{\theta \le 0} \exp\left\{-\theta(4N_i\sigma_i)^{-1}\right\} \exp\left\{\theta^2/2\right\} = \exp\left\{-(4N_i\sigma_i)^{-2}/2\right\}.$$

□

LEMMA 5.14. *Under the assumptions and notation of Section 4, for any $(y_J, y_I) \in \mathcal{Y}$, some constants $C_3, C_4 > 0$ and all sufficiently large n,*

$$\frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} \ge C_3 \frac{\sigma_n^{2\beta}}{m^2} \equiv \lambda_n, \tag{5.19}$$

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

$$\frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} \ge \exp\left\{-\frac{8||x||^2}{\underline{\sigma}_n^2} - C_4 \log n\right\} \tag{5.20}$$

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

PROOF. By assumption (4.8), $f_{0|J}(\tilde{x}|y_J) \le \bar{f}_0$, and $\pi_{0J}(y_J) \le 1$ for all $(\tilde{x}, y_J)$. Therefore,

$$\frac{f_J(y_J, \tilde{x}|\theta, m)}{f_{0|J}(\tilde{x}|y_J)\pi_{0J}(y_J)} \ge \bar{f}_0^{-1} f_J(\tilde{x}, y_J|\theta, m) \tag{5.21}$$

Let $k^\star = y_J$. Then, by Lemma 5.13, for any $j \in \{1, \ldots, K\}$,

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

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

For any $\tilde{x} \in \tilde{\mathcal{X}}$ with $\|\tilde{x}\| \le 2a_{\sigma_n}$, 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, $\sum_{i \in J^c} (\tilde{x}_i - \mu_{j^\star|k^\star,i})^2/\sigma_i^2 \le 4$. Then,

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

Thus,

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

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

$$\frac{f_J(y_J, \tilde{x}|\theta, m)}{f_{0|J}(\tilde{x}|y_J)\pi_{0J}(y_J)} \ge \bar{f}_0^{-1} \cdot \min_{j \le K, k \in \mathcal{Y}_J} \alpha_{jk} \cdot (2\pi)^{-d_{J^c}/2} \sigma_n^{-d_{J^c}} e^{-2} \cdot \frac{1}{2}$$

$$\geq 2C_3\frac{\sigma_n^{2\beta}}{m^2} = 2\lambda_n. \tag{5.22}$$

By assumption (4.9), for any $x \in \mathcal{X}$, any $y_J \in \mathcal{Y}_J$, and all sufficiently large n,

$$\int_{A_{y_I}} f_{0|J}(\tilde{x}|y_J)\pi_{0J}(y_J)d\tilde{y}_I \leq 2\int_{A_{y_I}\cap\{\tilde{y}_I:\,\|\tilde{y}_I\|\leq a_{\sigma_n}\}} f_{0|J}(\tilde{x}|y_J)\pi_{0J}(y_J)d\tilde{y}_I. \tag{5.23}$$

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

$$\begin{aligned}\frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} &= \frac{\int_{A_{y_I}} f_J(y_J, \tilde{x}|\theta, m)d\tilde{y}_I}{\int_{A_{y_I}} f_{0|J}(\tilde{x}|y_J)\pi_{0J}(y_J)d\tilde{y}_I} \\ &\geq \frac{\int_{A_{y_I}\cap\{\tilde{y}_I:\,\|\tilde{y}_I\|\leq a_{\sigma_n}\}} f_J(y_J, \tilde{x}|\theta, m)d\tilde{y}_I}{2\int_{A_{y_I}\cap\{\tilde{y}_I:\,\|\tilde{y}_I\|\leq a_{\sigma_n}\}} f_{0|J}(\tilde{x}|y_J)\pi_{0J}(y_J)d\tilde{y}_I} \geq \lambda_n,\end{aligned} \tag{5.24}$$

where the first inequality follows from (5.23) and the second one from (5.22) combined with Lemma 5.4.

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

$$\begin{aligned}\phi(\tilde{x}, \mu_{jk,J^c}, \sigma_{J^c}) &= (2\pi)^{-d_{J^c}/2}\prod_{i\in J^c}\sigma_i^{-1}\exp\left\{-0.5\sum_{i\in J^c}(\tilde{x}_i - \mu_{jk,i})^2/\sigma_i^2\right\} \\ &\geq (2\pi)^{-d_{J^c}/2}\sigma_n^{-d_{J^c}}\exp\left\{-\frac{8||x||^2}{\underline{\sigma}_n^2}\right\}.\end{aligned}$$

Then, for n large enough

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

Combining this inequality with (5.21), we get

$$\begin{aligned}\frac{f_J(y_J, \tilde{x}|\theta, m)}{f_{0|J}(\tilde{x}|y_J)\pi_{0J}(y_J)} &\geq \frac{1}{2}(2\pi)^{-d_{J^c}/2}\bar{f}_0^{-1}\sigma_n^{-d_{J^c}}K\frac{\sigma_n^{2\beta+d_{J^c}}}{2m^2}\exp\left\{-\frac{8||x||^2}{\underline{\sigma}_n^2}\right\} \\ &\geq \exp\left\{-\frac{8||x||^2}{\underline{\sigma}_n^2} - C_4\log n\right\}\end{aligned} \tag{5.25}$$

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

Thus, for $||x|| > a_{\sigma_n}$, (5.25) and the first inequality in (5.24), which holds for any $x \in \mathcal{X}$, deliver

$$\frac{p(y_J, y_I, x|\theta, m)}{p_0(y_J, y_I, x)} \geq \exp\left\{-\frac{8||x||^2}{\underline{\sigma}_n^2} - C_4 \log n\right\}. \tag{5.26}$$

□

LEMMA 5.15. *Under the assumptions and notation of Section 4, for $\lambda_n < \lambda_0$, where λ_0 is defined in Lemma 5.9,*

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

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 5.9 and penultimate inequality is derived using inequalities (4.27) and (4.31). Similarly,

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

Furthermore,

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

where the term multiplying $\tilde{\epsilon}_n^2$ on the right hand side is bounded by Assumption 4.5 ($N_J = o(n^{1-\nu})$) and definitions of $\tilde{\epsilon}_n$ and σ_n. □

LEMMA 5.16. *Under the assumptions and notation of Section 4, for all sufficiently large n, $s = 1 + 1/\beta + 1/\tau$, and some $C_6 > 0$*

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

PROOF. First, consider the prior probability of $m = N_J K$. By (4.5) for some $C_{61} > 0$,

$$\Pi(m = N_J K) \propto \exp[-a_{10} N_J K (\log N_J K)^{\tau_1}] \geq \exp[-C_{61} N_J \tilde{\epsilon}_n^{-d_{Jc}/\beta} \{\log(1/\tilde{\epsilon}_n)\}^{s d_{Jc}} (\log n)^{\tau_1}]$$
$$\geq \exp[-C_{61} N_J \tilde{\epsilon}_n^{-d_{Jc}/\beta} \{\log(n)\}^{s d_{Jc} + \tau_1}] \tag{5.27}$$

as $N_J = o(n^{1-\nu})$ by (4.15) and $\tilde{\epsilon}_n^{-1} < n$.

Second, consider the prior on $\{\alpha_{jk}\}$. There exist (j_0, k_0) such that $\alpha^\star_{j_0 k_0} \geq \frac{1}{m}$ 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,

$$\left|\alpha^\star_{j_0 k_0} - \alpha_{j_0 k_0}\right| = \left|\sum_{(jk) \neq (j_0 k_0)} \alpha^\star_{jk} - \alpha_{jk}\right| \leq (m-1)\frac{\sigma_n^{2\beta}}{m^2} \leq \frac{\sigma_n^{2\beta}}{m}$$

$$\alpha_{j_0 k_0} \geq \alpha^\star_{j_0 k_0} - \frac{\sigma_n^{2\beta}}{m} \geq \frac{1 - \sigma_n^{2\beta}}{m} \geq \frac{\sigma_n^{2\beta + d_{Jc}}}{2m^2}.$$

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

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

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

$$K \log(N_J K / \sigma_n^\beta) \leq \tilde{\epsilon}_n^{-d_{Jc}/\beta} \log(\tilde{\epsilon}_n^{-1})^{d_{Jc}s} \log(N_J \tilde{\epsilon}_n^{-d_{Jc}/\beta - 1} \log(\tilde{\epsilon}_n^{-1})^{d_{Jc}s+1})$$
$$\lesssim \tilde{\epsilon}_n^{-d_{Jc}/\beta} \log(n)^{d_{Jc}s+1}. \tag{5.28}$$

Assumption (4.4) on the prior for σ_i implies that for $i \in J$

$$\prod_{i=1}^{d_J} \Pi(\sigma_i^{-2} \geq 32 N_i^2 \beta \log \sigma_n^{-1}) \geq \prod_{i=1}^{d_J} \left(a_6 (64 N_i^2 \beta \log \sigma_n^{-1})^{a_7} \exp\left\{ -a_9 (64 N_i^2 \beta \log \sigma_n^{-1})^{1/2} \right\} \right)$$
$$\geq \exp\left\{ -C_{63} N_J \log(\sigma_n^{-1}) \right\} \geq \exp\left\{ -C_{64} N_J \log(n) \right\}, \tag{5.29}$$

and for $i \in J^c$,

$$\prod_{i=1}^{d_{J^c}} \Pi\left(\sigma_{i,n}^{-2} \leq \sigma_i^{-2} \leq \sigma_{i,n}^{-2}(1 + \sigma_n^{2\beta}) \right) \geq \prod_{i=1}^{d_{J^c}} \left(a_6 (\sigma_{i,n}^{-2})^{a_7} \sigma_n^{2a_8\beta} \exp\left\{ -a_9 \sigma_{i,n}^{-1} \right\} \right)$$
$$\geq \prod_{i=1}^{d_{J^c}} \exp\left\{ -C_{65} \sigma_{i,n}^{-1} \right\} = \prod_{i=1}^{d_{J^c}} \exp\left\{ -C_{65} \sigma_n^{-\beta/\beta_i} \right\} \geq \exp\left\{ -C_{65} d_{J^c} \sigma_n^{-d_{J^c}} \right\}$$
$$\geq \exp\left\{ -C_{66} \tilde{\epsilon}_n^{-d_{J^c}/\beta} \log(n)^{d_{J^c}/\beta} \right\}. \tag{5.30}$$

Assumption (4.6) on the prior for μ_{jk} implies

$$\prod_{j=1}^{K} \prod_{k \in \mathcal{Y}_J} \prod_{i \in J} \Pi\left(\mu_{jk,i} \in \left[k_i - \frac{1}{4N_i}, k_i + \frac{1}{4N_i} \right] \right) \geq \left(a_{11} 2^{-d_J} N_J^{-1} \exp\{-a_{12}\} \right)^{N_J K}$$
$$\geq \exp\left\{ -C_{67} N_J K \log(N_J) \right\}$$
$$\geq \exp\left\{ -C_{68} N_J \tilde{\epsilon}_n^{-d_{J^c}/\beta} \log(n)^{d_{J^c} s + 1} \right\} \tag{5.31}$$

and

$$\prod_{j=1}^{K} \prod_{k \in \mathcal{Y}_J} \Pi\left(\mu_{jk,J^c} \in U_{j|k} \right) \geq \left(a_{11} \exp\left\{ -a_{12} a_{\sigma_n}^{\tau_2} \right\} \min_{j,k} Vol(U_{j|k}) \right)^{N_J K}$$
$$= \left(a_{11} \exp\left\{ -a_{12} a_{\sigma_n}^{\tau_2} \right\} \sigma_n^{d_{J^c}} \tilde{\epsilon}_n^{2b_1 d_{J^c}} \right)^{N_J K}$$
$$\geq \exp\left\{ -C_{69} N_J \tilde{\epsilon}_n^{-d_{J^c}/\beta} \log(n)^{d_{J^c} s + \max\{1, \tau_2/\tau\}} \right\}. \tag{5.32}$$

It follows from (5.27) - (5.32), that for all sufficiently large n and some $C_6 > 0$,

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

□

5.2.2. *Sieve Construction and Entropy Bounds.*

LEMMA 5.17. *For $H \in \mathbb{N}$, $0 < \underline{\sigma} < \overline{\sigma}$, and $\overline{\mu} > 0$, let us define a sieve*

$$\mathcal{F} = \{ p(y, x|\theta, m) : \ m \leq H, \ \mu_j \in [-\overline{\mu}, \overline{\mu}]^d, j = 1, \ldots, m, \sigma_i \in [\underline{\sigma}, \overline{\sigma}], i = 1, \ldots, d \}. \tag{5.33}$$

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 $\underline{\sigma}$,

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

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

$$\begin{aligned}d_{TV}(f(\cdot|\theta, m), f(\cdot|\tilde{\theta}, m)) &\leq \sum_j |\alpha_j - \tilde{\alpha}_j| + 2 \max_{j=1,\ldots,m} ||\phi_{\mu_j,\sigma} - \phi_{\tilde{\mu}_j,\tilde{\sigma}}||_1 \\ &\leq \epsilon/3 + 4 \sum_{i=1}^d \left\{ \frac{|\mu_{ji} - \tilde{\mu}_{ji}|}{\sigma_i \wedge \tilde{\sigma}_i} + \frac{|\sigma_i - \tilde{\sigma}_i|}{\sigma_i \wedge \tilde{\sigma}_i} \right\} \leq \epsilon.\end{aligned}$$

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 (4.7) lead to dimension d appearing in front of some of the terms in the bound.

□

LEMMA 5.18. *Consider* $\epsilon_n = (N_J/n)^{\beta_{J^c}/(2\beta_{J^c}+1)}(\log n)^{t_J}$ *and* $\tilde{\epsilon}_n = (N_J/n)^{\beta_{J^c}/(2\beta_{J^c}+1)}(\log n)^{\tilde{t}_J}$ *with* $t_J > \tilde{t}_J + \max\{0, (1-\tau_1)/2\}$ *and* $\tilde{t}_J > t_{J0}$, *where* t_{J0} *is defined in* (4.16). *Define* $\mathcal{F}_n$ *as in* (5.33) *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* (4.19) *and* (4.20) *for all large* n.

PROOF. From Lemma 5.17,

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

Also,

$$\begin{aligned}\Pi(\mathcal{F}_n^c) \leq H^2 \exp\{-a_{13}n\} + \exp\{-a_{10}H(\log H)^{\tau_1}\} \\ + a_1 \exp\{-a_2 n\} + a_4 \exp\{-2a_5 n\}.\end{aligned}$$

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 $t_J > \tilde{t}_J + \max\{0, (1-\tau_1)/2\}$.

□

ECONOMICS DEPARTMENT,
BROWN UNIVERSITY, PROVIDENCE, RI 02912
E-MAIL: andriy_norets@brown.edu

INSTITUTE FOR ADVANCED STUDIES VIENNA,
JOSEFSTAEDTER STRASSE 39,
VIENNA 1080, AUSTRIA E-MAIL: pelenis@ihs.ac.at