

Adaptive Bayesian multivariate density estimation with Dirichlet mixtures

BY WEINING SHEN AND SUBHASHIS GHOSAL

Department of Statistics, North Carolina State University, 5217 SAS Hall, 2311 Stinson Drive,

Raleigh, North Carolina 27695-8203, USA

wshen2@ncsu.edu sghosal@ncsu.edu

SUMMARY

We consider Bayesian multivariate density estimation using a Dirichlet mixture of normal kernel as the prior distribution. By representing a Dirichlet process as a stick-breaking process, we are able to extend convergence results beyond finitely supported mixtures priors to Dirichlet mixtures. Thus our results have new implications in the univariate situation as well. Assuming that the true density satisfies Hölder smoothness and exponential tail conditions, we show the rates of posterior convergence are minimax-optimal up to a logarithmic factor. This procedure is fully adaptive since the priors are constructed without using the knowledge of the smoothness level.

Some key words: Bayesian density estimation; multivariate; rate-adaptive; Dirichlet mixture.

1. INTRODUCTION

Kernel methods for density estimation has been well studied in the past fifty years (Wand & Jones, 1995). In the nonparametric Bayesian literature, the study of asymptotic properties of

49 posterior distributions received a lot of interest since the development of efficient Markov chain
50 Monte Carlo (MCMC) methods (Escobar & West, 1995). A general result on posterior consis-
51 tency was established in Ghosal et al. (1999) and then applied on the univariate Dirichlet mix-
52 ture of normal prior; see Tokdar (2006) for an improved result. General posterior convergence
53 rate theorems were obtained in Ghosal et al. (2000) and Shen & Wasserman (2001). Ghosal &
54 van der Vaart (2001) considered univariate Bayesian density estimation problem using Dirichlet
55 mixture of normal kernel and studied the case when the true density is a location-scale mixture
56 type while its standard deviation is bounded away from zero and infinity. Although the poste-
57 rior rate is nearly the parametric rate $n^{-1/2}$, the assumption of “super smooth” true density with
58 the bounded range of standard deviation is quite restrictive. Using a new general rate theorem,
59 Ghosal & van der Vaart (2007) obtained posterior convergence rate of univariate Dirichlet mix-
60 ture of normal kernel when the true density is only twice continuously differentiable. Though the
61 number of mixture components increases, the minimax rate is still obtained. These results need
62 a prior on the bandwidth parameter that scales appropriately with increasing sample size.

63 In recent studies, rate-adaptive estimators based on posterior distributions have been con-
64 structed to accommodate different levels of smoothness of the underlying true function of in-
65 terest. Belitser & Ghosal (2003) considered the problem of estimating a signal with Gaussian
66 white noise and showed that the posterior rate automatically adapts to the unknown smoothness
67 condition if the “smoothness parameter” only takes values in a discrete set. Huang (2004) and
68 Ghosal et al. (2008) showed that appropriate mixture of priors based on spline expansions or
69 wavelets yield optimal posterior rates for a finite or countable range of smoothness parameters
70 for density estimation and nonparametric regression problems. Alternatively, van der Vaart &
71 van Zanten (2009) constructed a prior based on a randomly rescaled smooth Gaussian process,

72

73

74

75

76

77

97 which automatically adapts for a continuous range of smoothness parameters. They treated the
98 multidimensional case as well.

99 A technical challenge in proving adaptation of the posterior distribution based on mixture pri-
100 ors is to find an approximation of the true function within the model, whose accuracy increases
101 appropriately with increasing smoothness level of the true density. An interesting approximation
102 idea proposed by Rousseau (2010) in the context of beta mixtures prior turns out to be very
103 helpful for constructing the required approximation and subsequent adaptive posterior distribu-
104 tions. A similar idea for normal mixtures was proposed by Kruijer et al. (2010). An analogous
105 approximation in the multi-dimensional situation was constructed recently in de Jonge & van
106 Zanten (2010). They used a special type of Gaussian process to construct an adaptive procedure.
107 However, their constructions apply only to compactly supported densities. The issue of unbound-
108 edness of the support was resolved in Kruijer et al. (2010) for univariate Gaussian mixtures by
109 imposing appropriate tail conditions on the true density.

110 The adaptation results in Kruijer et al. (2010) used a prior based on finite mixture of the
111 normal kernel in a univariate setting. In practice, Dirichlet mixture priors are popularly used
112 in the univariate density estimation problems; see Ferguson (1983) and Lo (1984), as well as
113 in the multivariate situations (Müller et al., 1996). Posterior consistency results in terms of the
114 L_1 -distance were studied in Wu & Ghosal (2010) under a multivariate setting. An extension to
115 multivariate mixed-scale density estimation has been recently discussed in Canale & Dunson
116 (2011).

117 In this paper, we study the posterior convergence rates for Bayesian multivariate density es-
118 timation. We extend the approximation result in Kruijer et al. (2010) to the multi-dimensional
119 setting assuming local β -Hölder smoothness and exponential tail conditions. Using the stick-
120 breaking representation (Muliere & Tardella, 1998), we approximate a Dirichlet process by a

121

122

123

124

125

145 finite sum of mixtures while the error is controlled within a pre-determined level, which helps
 146 us construct appropriate sieves for the problem. Similar technique has been used in Pati et al.
 147 (2011) to prove posterior consistency for conditional density estimation. We calculate the en-
 148 tropy and prior concentration rate around the true density. The posterior rate is shown to be
 149 $n^{-\beta/(2\beta+d)}(\log n)^\kappa$, where κ is determined by the smoothness level, the dimension of the sam-
 150 ple space and the tail behavior of the true density. The rate coincides with the minimax rate up
 151 to a logarithmic factor.

152 To the best of our knowledge, most frequentist approaches for adaptive estimation are focused
 153 on using wavelets under a regression model setting; see Donoho & Johnstone (1995) and Rigollet
 154 (2006). The performance of adaptive multivariate kernel density estimation depends heavily on
 155 the choice of the bandwidth matrix and the smoothing kernel (Scott, 1992). Our model considers
 156 kernel based Bayesian adaptive estimation procedure that achieves optimal rates using product
 157 kernel.

158 The paper is organized as follows. In Section 2, some notations and assumptions on the true
 159 density are introduced. The main results on posterior convergence rates are presented in Sec-
 160 tion 3. Approximation results are given in Section 4. Section 5 gives the proof of the main rate
 161 theorem. A few auxiliary lemmas and their technical proofs are presented in the Appendix.

162

163

2. NOTATIONS AND ASSUMPTIONS

164

2.1. Notations

165

166

167

168

169

170

171

172

173

Throughout the paper, we consider estimating a density f on \mathbb{R}^d based on n independent and
 identically distributed (i.i.d) samples $\mathbf{X}_1, \dots, \mathbf{X}_n$ taking values in \mathbb{R}^d . Let $\mathbf{X} = (X_1, \dots, X_d)$
 stand for a generic observation from density f . We define marginal density functions of f for X_i
 as $f_i(x_i)$, $i = 1, \dots, d$. Let $\mathbb{N} = \{0, 1, 2, \dots\}$ and let Δ_k be a k -dimensional unit simplex. For

193 $\mathbf{k} \in \mathbb{N}^d$, $\mathbf{x} \in \mathbb{R}^d$, let $\mathbf{k} \cdot = k_1 + k_2 + \dots + k_d$, $\mathbf{k}! = k_1! \dots k_d!$ and $\mathbf{x}^{\mathbf{k}} = x_1^{k_1} \dots x_d^{k_d}$. Similarly,
 194 for a real-valued function f on \mathbb{R}^d , let $f(\mathbf{x})^{\mathbf{k}} = f(x_1)^{k_1} \dots f(x_d)^{k_d}$. We define partial order for
 195 \mathbf{j} and \mathbf{k} as $\mathbf{j} \geq \mathbf{k}$ if $j_i \geq k_i$ for $i = 1, \dots, d$. Let $\|\mathbf{x}\|_p = \{\sum_{i=1}^d |x_i|^p\}^{1/p}$ stand for the ℓ_p -norm
 196 of a vector $\mathbf{x} \in \mathbb{R}^d$; $1 \leq p < \infty$ and $\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq d} |x_i|$. Moreover, for $p = 2$, we simply
 197 write $\|\mathbf{x}\|_2$ as $\|\mathbf{x}\|$. For $b > 0$, let r_b stand for the largest integer strictly smaller than b .

198 We use $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_d)' \in \mathbb{R}_+^d$ as the scale parameter and define a $d \times d$ diagonal matrix
 199 $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\sigma})$. Let $\phi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ be the standard normal density and $\phi_\sigma(x) =$
 200 $\sigma^{-1} \phi(x/\sigma)$. The corresponding multivariate normal density with independent components is
 201 denoted by $\phi_{\boldsymbol{\sigma}}(\mathbf{x}) = \prod_{i=1}^d \phi_{\sigma_i}(x_i)$.

202 We use \lesssim for inequality up to a constant multiple, where the underlying constant of propor-
 203 tionality is universal or not important for our purposes. We define a linear operator K_{σ_i} as

$$204 (K_{\sigma_i} f)(\mathbf{x}) = \int_{-\infty}^{\infty} f(x_1, \dots, x_{i-1}, x_i - y_i, x_{i+1}, \dots, x_d) \phi_{\sigma_i}(y_i) dy_i. \quad (1)$$

205
 206
 207 Then a composition operator is defined as $K_{\sigma_i}^{m_i} = K_{\sigma_i}(K_{\sigma_i}^{m_i-1} f)$. Note that these convolution
 208 operators commute with each other. We extend this notation to the multivariate case as $K_{\boldsymbol{\sigma}}^m f =$
 209 $(K_{\sigma_1}^{m_1} \dots K_{\sigma_d}^{m_d}) f$. For simplicity, we define $K_{\boldsymbol{\sigma}} = K_{\boldsymbol{\sigma}}^{(1, \dots, 1)}$.

210 We use $D(\epsilon, T, d)$ to denote the packing number, which is defined as the maximum cardinality
 211 of an ϵ -dispersed subset of T with respect to distance d . Similarly, we write $N(\epsilon, T, d)$ for the
 212 covering number, the minimal cardinality of an ϵ -net for T in terms of the distance d . We define
 213 $\log_+(x) = \max(\log x, 0)$.

214
 215
 216

2.2. Assumptions on the true density

217 Let f_0 stand for the true density. We assume the following conditions on f_0 .

218
 219
 220
 221

241 • (C1) Smoothness: The function $\log f_0$ is assumed to be locally β -Hölder with derivatives

242 $l_j(\mathbf{x}) = \frac{\partial \log f(\mathbf{x})}{\partial x_1^{j_1} \cdots \partial x_d^{j_d}}$. We assume the existence of a polynomial L and a constant $\gamma > 0$,

243 such that for $r = r_\beta$,

244

$$|l_k(\mathbf{x}) - l_k(\mathbf{y})| \leq r!L(\mathbf{x})\|\mathbf{x} - \mathbf{y}\|^{\beta-r}$$

245

246 for all $k. = r$ and \mathbf{x}, \mathbf{y} satisfying $\|\mathbf{x} - \mathbf{y}\| \leq \gamma$. Moreover, there exists a constant $\xi_0 > 0$ such

247 that for all $j. \leq r$,

$$248 \int f_0(\mathbf{x})|l_j(\mathbf{x})|^{(2\beta+\xi_0)/j} d\mathbf{x} < \infty, \int f_0(\mathbf{x})|L(\mathbf{x})|^{2+\xi_0/\beta} d\mathbf{x} < \infty. \quad (2)$$

249

250 • (C2) Marginal-joint relationship: There exist a constant C_0 and density functions g_1, \dots, g_d

251 such that $f_0(x_1, \dots, x_d) \geq C_0 \prod_{i=1}^d g_i(x_i)$, and $\int f_0(\mathbf{x})(1/g(\mathbf{x}))^\xi \max(1, \|\mathbf{x}\|^2) d\mathbf{x} < \infty$

252

for some $\xi > 0$, where $g(\mathbf{x}) = \prod_{i=1}^d g_i(x_i)$.

252

253 • (C3) Tail monotonicity: On a region $D = [-a, a]^d$, where $a > 0$, we have that $\inf_{\mathbf{x} \in D} g(\mathbf{x}) =$

253

$c_0 > 0$, g_i is nondecreasing on $x_i < -a$ and nonincreasing on $x_i > a$ for $i = 1, \dots, d$.

254

255 • (C4) Tail decay: The true density f_0 has exponential tails on D^c , i.e., there exist constants

255

$C > 0$ and $\tau_1, \tau_2 > 0$, which only depend on f_0 , such that

256

$$257 f_0(\mathbf{x}) \leq C e^{-\tau_1 \|\mathbf{x}\|^{\tau_2}}, \mathbf{x} \in D^c. \quad (3)$$

258

259 *Remark 1.* Conditions (C2) and (C4) imply $\int f_0(\log_+(f_0/g))^p < \infty$ for any $p > 0$. Condi-

259

260 tions (C1), (C3) and (C4) imply $\int f_0(\log_+ f_0)^p < \infty$ for any $p > 0$.

260

261 A wide range of multivariate density functions satisfy Condition (C2), e.g., nonsingular mul-

261

262 tivariate normal distribution and their finite mixtures. To see this, consider k multivariate normal

262

263 densities $f_j, j = 1, \dots, k$, with mean $\mathbf{0}$ and covariance matrix Σ_j . For any convex combination

263

264 of f_j 's $f^* = \sum_{j=1}^k \omega_j f_j$, there exists $\lambda > 0$ such that $f^*(\mathbf{x}) \gtrsim \exp\{-\lambda \|\mathbf{x}\|^2/2\}$. Define density

264

265 $g^* = (\lambda/2\pi)^{d/2} \exp\{-\lambda \|\mathbf{x}\|^2/2\}$, then $f^* \gtrsim g^*$. To see this, choose λ to be the smallest eigen-

265

266

267

268

269

289 value of all Σ_j^{-1} s. Then for any $0 < \xi < 1$, $\int f^*(1/g^*)^\xi \max(1, \|\mathbf{x}\|^2) < \infty$. Hence Condition
 290 (C2) holds for f^* .

291 Condition (C2) also holds for product type densities $f_0(\mathbf{x}) = \prod_{j=1}^d f_j(x_j)$ with $g = f_0$, if
 292 $\int f_0^{\xi_1} \max(1, \|\mathbf{x}\|^2) d\mathbf{x} < \infty$ for some $0 < \xi_1 < 1$.

293
 294 *Remark 2.* Condition (C2) is used to lower bound $K_\sigma f_0$ as in Lemma 2. Condition (C3) gen-
 295 eralizes the monotone tail condition in Kruijer et al. (2010) to the multivariate case.

296 297 3. MAIN RESULTS

298 We construct a prior for f as follows:

- 299
- 300 • $p_{F,\sigma} = \int_{\mathbb{R}^d} \phi_\sigma(\mathbf{x} - \boldsymbol{\mu}) dF(\boldsymbol{\mu})$;
- 301 • F follows a Dirichlet process D_α with base measure α . Denote $\bar{\alpha} = \alpha/\alpha(\mathbb{R}^d)$. We assume that
 302 there exist constants $a_1, a_2 > 0$ such that $1 - \bar{\alpha}([-x, x]^d) \leq \exp\{-a_1 z^{a_2}\}$ for sufficiently
 303 large $x > 0$.
- 304 • $\sigma_i \stackrel{\text{iid}}{\sim} G$ for $i = 1, \dots, d$, where G is a fixed probability distribution satisfying $G(x) \lesssim$
 305 $\exp\{-C_1 x^{-a_3}\}$ as $x \rightarrow 0$ and $1 - G(x) \lesssim x^{a_3}$ as $x \rightarrow \infty$, where $C_1 > 0$ and $a_3 \geq 1$ are
 306 fixed constants. This condition allows a wide class of distributions, e.g., an inverse gamma
 307 distribution for σ^2 when $a_3 = 2$ or an inverse gamma distribution for σ when $a_3 = 1$.

308
 309 We have the following result for posterior convergence rates:

310
 311 **THEOREM 1.** *Let f_0 satisfy Conditions (C1)–(C4). Then the posterior rate of convergence*
 312 *with respect to Hellinger or L_1 -distance is given by $\epsilon_n = n^{-\beta/(d+2\beta)}(\log n)^t$, where $t > \left(\frac{d}{\tau_2} +$
 313 $d + 1\right) \frac{\beta}{2\beta+d}$.*

314
 315
 316
 317

337 The assumption on the base measure $\bar{\alpha}$ is analogous to (11) of Kruijer et al. (2010). Our tail
 338 conditions on the prior of σ is weaker than the one in Ghosal & van der Vaart (2007). Both sets
 339 of conditions are needed to control the prior probability of the model.

340 The results also hold if $\sigma_1 = \dots = \sigma_d = \sigma$ and $\sigma \sim G$. It will be interesting to consider other
 341 types of location-scale kernels instead of normal kernel in the future study. The techniques de-
 342 veloped in our paper might be helpful.

343 Our result also applies for finite-mixture priors. We consider the prior for f as follows:

344 • $m(\mathbf{x}; k, \boldsymbol{\mu}, \boldsymbol{\omega}, \boldsymbol{\sigma}) = \sum_{j=1}^k \omega_j \phi_{\boldsymbol{\sigma}}(\mathbf{x} - \boldsymbol{\mu}_j);$

345 • There exists constants $c_1 > c_2 > 0$ and $c_3 > 0$ such that

346
$$\exp\{-c_1 k(\log k)^{c_3}\} \lesssim \Pi(k) \lesssim \exp\{-c_2 k(\log k)^{c_3}\}.$$

348 • Given $k, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k$ are i.i.d realizations from a distribution, which satisfies $\Pi(\boldsymbol{\mu} \notin$
 349 $[-z, z]^d) < \exp\{-c_4 z^{c_5}\}$ for sufficiently large $z > 0$ and constants $c_4, c_5 > 0$.

350 • Given k , the prior on weights $\boldsymbol{\omega} = (\omega_1, \dots, \omega_k)'$ satisfies

351
$$\Pi(\|\boldsymbol{\omega} - \boldsymbol{\omega}_0\|_1 \leq \epsilon) \gtrsim \exp\{-c_6 k(\log k)^{a_4} \log_- \epsilon\}$$

352 for any $\boldsymbol{\omega}_0 \in \Delta_k$ and constants $a_4, c_6 > 0$ and $0 < \epsilon < 1/k$.

353 • Bandwidth $\sigma_1, \dots, \sigma_d$ (i.i.d) follow an inverse gamma distribution.

354 Then we have the following rate theorem, which is a generalization of Theorem 2 of Kruijer
 355 et al. (2010).

358 THEOREM 2. *Let f_0 satisfy Conditions (C1)–(C4). Then the posterior rate of convergence*
 359 *with respect to Hellinger or L_1 -distance is given by $\epsilon_n = n^{-\beta/(d+2\beta)}(\log n)^t$, where $t >$*

360
$$\frac{\beta}{2\beta+d} \left(\frac{d}{\tau_2} + d + \max\{c_3, 1 + a_4, \frac{c_5}{\tau_2}\} \right) + \max\{0, (1 - c_3)/2\}.$$

361

362

363

364

365

4. APPROXIMATION RESULTS

The following proposition helps prove the main theorem on posterior convergence rates. It is also of interest on its own as it bounds the Kullback-Leibler (KL) divergence between f_0 and its approximation. The proof is given in Appendix.

PROPOSITION 1. *Let f_0 be the true density satisfying Conditions (C1)–(C4). Then there exists a density h_β such that for all sufficiently small σ ,*

$$\int f_0(\mathbf{x}) \log \frac{f_0(\mathbf{x})}{K_\sigma h_\beta(\mathbf{x})} d\mathbf{x} = O(\sigma^{2\beta}), \quad (4)$$

$$\int f_0(\mathbf{x}) \left(\log \frac{f_0(\mathbf{x})}{K_\sigma h_\beta(\mathbf{x})} \right)^2 d\mathbf{x} = O(\sigma^{2\beta}), \quad (5)$$

where $\sigma = (\sigma, \dots, \sigma)$.

In order to prove approximation result, we use the expansion technique in Kruijer et al. (2010) and its multivariate modification described by de Jonge & van Zanten (2010).

Let r and β be defined as in Condition (C1). For $\mathbf{k} \in \mathbb{N}^d$, we define moments $m_{\mathbf{k}} = \int y_1^{k_1} \dots y_d^{k_d} \phi(\mathbf{y}) d\mathbf{y}$. Then we recursively define two collections of numbers $c_{\mathbf{n}}$ and $d_{\mathbf{n}}$ as follows:

For $\mathbf{n} \in \mathbb{N}^d$, if $\mathbf{n} = 1$, then $c_{\mathbf{n}} = d_{\mathbf{n}} = 0$. For $\mathbf{n} \geq 2$, define

$$c_{\mathbf{n}} = \sum_{\mathbf{n}=\mathbf{l}+\mathbf{k}, \mathbf{l} \geq 1, \mathbf{k} \geq 1} \frac{(-1)^{\mathbf{k} \cdot +1}}{\mathbf{k}!} m_{\mathbf{k}} d_{\mathbf{l}}, \quad d_{\mathbf{n}} = \frac{m_{\mathbf{n}}}{\mathbf{n}!} + c_{\mathbf{n}}. \quad (6)$$

Since the Gaussian kernel is symmetric about 0, all odd moments are 0. Hence $c_{\mathbf{n}}$ can be simplified as $c_{\mathbf{n}} = - \sum_{\mathbf{n}=\mathbf{l}+2\mathbf{k}, \mathbf{l} \geq 1, \mathbf{k} \geq 1} m_{2\mathbf{k}} d_{\mathbf{l}} / (2\mathbf{k}!)$

Define $f_\beta = f - \sum_{j=1}^r \sum_{\mathbf{k}=j} d_{\mathbf{k}} \sigma^{\mathbf{k}} (D_{\mathbf{k}}^j f)$, where $D_{\mathbf{k}}^j = \frac{\partial^j}{\partial x_1^{k_1} \dots \partial x_d^{k_d}}$. Lemma 3.4 in de Jonge & van Zanten (2010) shows that the supremum distance between f_0 and f_β is $O(\sigma^\beta)$.

However, this type of construction does not guarantee that f_β is a density function because it may take negative values. To overcome the problem, we define a truncated version of f_β and

then standardize it to obtain a density function:

$$h_\beta^*(\mathbf{x}) = f_\beta \mathbb{1}\{f_\beta > \frac{1}{2}f_0(\mathbf{x})\} + \frac{1}{2}f_0(\mathbf{x}) \mathbb{1}\{f_\beta \leq \frac{1}{2}f_0(\mathbf{x})\}$$

$$h_\beta(\mathbf{x}) = h_\beta^*(\mathbf{x}) / \int h_\beta^*(\mathbf{u}) d\mathbf{u} \quad (7)$$

Remark 3. From (7), we get $h_\beta \lesssim h_\beta^* \lesssim f_\beta + f_0$. Using the same arguments in Kruijer et al. (2010), we can show $f_\beta \lesssim f_0$. Then combining these two facts, we conclude that h_β is upper bounded by a multiple of f_0 .

Remark 4. From the definition, f_β can be expressed as a linear combination of $K_\sigma^j f_0$'s:

$$f_\beta = C_\beta f_0 - \sum_{j \geq 0} c_j K_\sigma^j f_0, \quad (8)$$

where C_β and c_j are constants determined by f_0 and β . The coefficients c_j satisfy $\sum_{j \geq 0} c_j = C_\beta - 1$. Hence $K_\sigma f_\beta$ is also a linear combination of $K_\sigma^j f_0$'s.

The approximation mixture in Proposition 1 can be discretized without changing the order of the approximation error. The following lemma is a multivariate generalization of Lemma 4 in Kruijer et al. (2010). This will be used to lower bound the prior probability on the KL-ball around f_0 . Its proof is given in Appendix.

LEMMA 1. *Let f_0 be a density satisfying Conditions (C1)–(C4). Then there exists a finitely supported probability measure F with at most $C_4 \sigma^{-d} |\log \sigma|^{d/\tau_2 + d}$ support points from the set $\{\mathbf{x} : f_0(\mathbf{x}) \geq c\sigma^{H_1 + 2\beta}\}$, where $C_4 > 0$ is a constant such that*

$$\int f_0 \log \frac{f_0}{p_{F,\sigma}} = O(\sigma^{2\beta}), \quad \int f_0 \left(\log \frac{f_0}{p_{F,\sigma}}\right)^2 = O(\sigma^{2\beta}). \quad (9)$$

5. PROOF OF THEOREMS

5.1. Some useful results

We first state a few results that are helpful for proving Theorem 1.

481 Since a Dirichlet process $F \sim D_\alpha$ can be represented by a Sethuraman's stick-breaking pro-
 482 cess as $\sum_{i=1}^{\infty} V_i \delta_{\theta_i}$, where $V_i = \prod_{j=1}^{i-1} (1 - Y_j) Y_i$, $\theta_i \stackrel{\text{iid}}{\sim} G$, $Y_i \stackrel{\text{iid}}{\sim} \text{Be}(1, M)$, $i = 1, 2, \dots$, G is the
 483 cumulative distribution function of $\bar{\alpha}$ and $M = \alpha(\mathbb{R})$. We truncate the stick-breaking proce-
 484 dure after a certain level such that the error is within a predetermined level. Define the num-
 485 ber of terms needed in the finite mixture as $N_\epsilon = \inf\{m \geq 1 : \sum_{i=1}^m V_i > 1 - \epsilon\}$. Let $F_\epsilon =$
 486 $\sum_{i=1}^{N_\epsilon} V_i \delta_{\theta_i} + \bar{V}_\epsilon \delta_{\theta_0}$, where $\bar{V}_\epsilon = \prod_{i=1}^{N_\epsilon} (1 - Y_i)$ and $\theta_0 \sim G$ independently of everything else.
 487 By Lemma 3 of Muliere & Tardella (1998), it follows that

$$488 \quad d_{\text{TV}}(F, F_\epsilon) \leq \epsilon, \quad (10)$$

$$489 \quad N_\epsilon - 1 \sim \text{Poi}(M \log_{-} \epsilon), \quad (11)$$

491 where d_{TV} stands for the total variation distance. It is easy to see from (10) that $\|p_{F, \sigma} -$
 492 $p_{F_\epsilon, \sigma}\|_1 \leq \epsilon$. The following lemma lower bounds $K_\sigma f_0$.

494 **LEMMA 2.** *Assume f_0 satisfy Conditions (C2) and (C3). Then given σ sufficiently small,*
 495 *$K_\sigma f_0 \geq C_5 g$ for some constant $C_5 > 0$ and density function g defined in (C2).*

496 We need the following inequalities to help lower bound the prior probability in the KL-ball
 497 around f_0 .

499 **LEMMA 3.** *Let $\mathbb{R}^d = \bigcup_{j=1}^N U_j$ be a partition of \mathbb{R}^d and $F' = \sum_{j=1}^N p_j \delta_{z_j}$ be a probabil-*
 500 *ity measure with $z_j \in U_j$ and $\|z_j - z_k\|_1 > 2\epsilon$ for $j, k = 1, \dots, N$, $j \neq k$ and $\epsilon > 0$. Define*
 501 *$V(z_j, \epsilon) = [z_{j,1} - \epsilon, z_{j,1} + \epsilon] \times \dots \times [z_{j,d} - \epsilon, z_{j,d} + \epsilon]$ and $x_{(1)} = \min_{1 \leq i \leq d} x_i$ for $\mathbf{x} \in \mathbb{R}^d$.*
 502 *Then for any probability measure F on \mathbb{R}^d , $\sigma, \sigma' \in \mathbb{R}_+^d$, we have that*

$$503 \quad \|p_{F, \sigma} - p_{F', \sigma'}\|_1 \lesssim \max_{i=1, \dots, d} \frac{|\sigma_i - \sigma'_i|}{\sigma_i \wedge \sigma'_i} + \frac{\epsilon}{(\sigma_{(1)} \wedge \sigma'_{(1)})^d} + \sum_{j=1}^N |F(V(z_j, \epsilon)) - p_j|, \quad (12)$$

504
505
506
507
508
509

529 and

$$\begin{aligned}
 530 \\
 531 \quad \|p_{F,\sigma} - p_{F',\sigma'}\|_\infty &\lesssim \frac{\epsilon}{(\sigma_{(1)} \wedge \sigma'_{(1)})^{2d}} + \frac{|\prod_{i=1}^d \sigma_i - \prod_{i=1}^d \sigma'_i|}{(\sigma_{(1)} \wedge \sigma'_{(1)})^{2d}} \\
 532 &+ \frac{1}{(\sigma_{(1)} \wedge \sigma'_{(1)})^d} \sum_{j=1}^N |F(V(\mathbf{z}_j, \epsilon)) - p_j|. \quad (13) \\
 533
 \end{aligned}$$

534 The following discretization result gives multidimensional extensions of Lemmas 3.1 and 3.3
 535 of Ghosal & van der Vaart (2001). Their proofs are technical and included in Shen & Ghosal
 536 (2011).

537
 538 LEMMA 4. (1) Let $0 \leq \epsilon \leq 1/2$ be given. Fix $\sigma, \sigma' \in [\underline{\sigma}_0, \bar{\sigma}_0]^d$ satisfying $|\sigma_0^d - \sigma_0'^d| < \bar{\sigma}_0^d \epsilon$,
 539 then for any probability measure F on a region $D' = [-a_1, a_1] \times \cdots \times [-a_d, a_d]$, where
 540 $\max_i a_i \leq L(\log_- \epsilon)^{\gamma_0}$, $\gamma_0 \geq 1/2$ and $L > 0$ are constants, there exists a discrete probabil-
 541 ity measure F' on D with at most $N \lesssim \underline{\sigma}_0^{-2d} (\log_- \epsilon)^{2\gamma_0 d}$ support points such that $\|p_{F,\sigma} -$
 542 $p_{F',\sigma'}\|_\infty \lesssim \epsilon \bar{\sigma}_0^d / \underline{\sigma}_0^{2d}$.

543 (2) Define $\sigma = \sigma \mathbf{1}$. If $\sigma \rightarrow 0$, then for any probability measure F on $[-a_\epsilon, a_\epsilon]^d$ with $a_\epsilon =$
 544 $L(\log_- \epsilon)^{\gamma_1}$, where $\gamma_1 \geq 0$, $0 \leq \epsilon \leq 1/2$ and $L > 0$ are constants, there exists a discrete proba-
 545 bility measure F' on $[-a_\epsilon, a_\epsilon]^d$ with at most $N \lesssim \sigma^{-d} (\log_- \epsilon)^{\gamma_1 d + d}$ support points such that

$$547 \quad \|p_{F,\sigma} - p_{F',\sigma}\|_\infty \lesssim \sigma^{-d} \epsilon, \quad (14)$$

$$548 \quad \|p_{F,\sigma} - p_{F',\sigma}\|_1 \lesssim \sigma^d (\sigma (\log_- \epsilon)^{1/2} \vee (\log_- \epsilon)^{\gamma_1})^d \epsilon. \quad (15)$$

549

550

5.2. Proof of Theorem 1 (Part I)

551 We apply Theorem 5 of Ghosal & van der Vaart (2007) for $\tilde{\epsilon}_n = n^{-\beta/(2\beta+d)} (\log n)^{t_1}$, $\bar{\epsilon}_n =$
 552 $n^{-\beta/(2\beta+d)} (\log n)^{t_2}$ for $t_2 > t_1$. We construct appropriate sieves $\mathcal{F}_{n,j}$ and verify the following

553

554

555

556

557

577 three conditions:

$$578 \sum_{j=0}^{\infty} \sqrt{N(\bar{\epsilon}_n, \mathcal{F}_{n,j}, d)} \sqrt{\Pi_n(\mathcal{F}_{n,j})} e^{-n\bar{\epsilon}_n^2} \rightarrow 0 \quad (16)$$

$$579 \Pi_n(\mathcal{K}(f_0, \tilde{\epsilon}_n)) \geq e^{-n\tilde{\epsilon}_n^2} \quad (17)$$

$$580 \Pi_n(\mathcal{F}_n^c) \leq e^{-4n\tilde{\epsilon}_n^2}, \quad (18)$$

582 where $\mathcal{K}(f_0, \epsilon) = \{f : \int f_0(\log f_0/f) < \epsilon^2, \int f_0(\log f_0/f)^2 < \epsilon^2\}$ is the KL ball around f_0 of
583 size ϵ . Choose $\sigma_{n,1} = \dots = \sigma_{n,d} = \tilde{\epsilon}_n^{1/\beta}$. Define

$$584 \underline{\sigma}_n = n^{-A}, \bar{\sigma}_n = \exp\{n\tilde{\epsilon}_n^2(\log n)^\delta\}, r_n = \lfloor n^{d/(2\beta+d)}(\log n)^{t_r} \rfloor + 1 \quad (19)$$

586 and $b_n > n^{d/a_2(d+2\beta)}$ for $A > 1, a_2, t_r, \delta > 0$. First we consider the collection of finite mixtures:

$$587 \mathcal{F}_n^* = \left\{ \sum_{i=1}^k \omega_i \phi_{\boldsymbol{\sigma}}(\mathbf{x} - \boldsymbol{\mu}_i) : k \leq r_n, \boldsymbol{\omega} \in \Delta_k, \boldsymbol{\mu}_i \in [-b_n, b_n]^d, \boldsymbol{\sigma} \in S_n, i = 1, \dots, k \right\}$$

589 as in Kruijer et al. (2010), where $S_n = [\underline{\sigma}_n, \bar{\sigma}_n]^d$.

590 Define the sieve

$$591 \mathcal{F}_n = \{p_{F,\boldsymbol{\sigma}} : \text{there exists } p_{F',\boldsymbol{\sigma}} \in \mathcal{F}_n^* \text{ such that } d_{\text{TV}}(F, F') \leq \bar{\epsilon}_n\}. \quad (20)$$

593 Notice that $\mathcal{F}_n^* \subset \mathcal{F}_n$.

594 We first verify equation (18). From the construction of priors of σ_i as in Section 3,

$$595 \Pi_n(\sigma_i \in (\underline{\sigma}_n, \bar{\sigma}_n)^c) \lesssim \exp\{-C_1 n^{A a_3}\} + \exp\{a_3 n \tilde{\epsilon}_n^2 (\log n)^\delta\}$$

$$596 \lesssim \exp\{-C_6 n \tilde{\epsilon}_n^2\}$$

598 for $i = 1, \dots, d$ and some constant $C_6 > 0$ when n is sufficiently large.

599 Given the number of mixtures $N_{\bar{\epsilon}_n}$ fixed, from the assumption, we have

$$600 \Pi_n(\boldsymbol{\mu} \notin [-b_n, b_n]^d | N_{\bar{\epsilon}_n} = k) \leq k \Pi_n(\boldsymbol{\mu}_1 \notin [-b_n, b_n]^d) \lesssim k e^{-a_1 b_n^{a_2}}. \quad (21)$$

601

602

603

604

605

Therefore, using $E(N_{\bar{\epsilon}_n}) = O(\log n)$, we have

$$\begin{aligned} \Pi_n(\boldsymbol{\mu} \notin [-b_n, b_n]^d) &= \sum_{k=1}^{\infty} \Pi(N_{\bar{\epsilon}_n} = k) \Pi_n(\boldsymbol{\mu} \notin [-b_n, b_n]^d | N_{\bar{\epsilon}_n} = k) \\ &\lesssim e^{-a_1 b_n^{a_2}} \log n. \end{aligned} \quad (22)$$

Using (11) and tail estimates of Poisson distribution $P(X > r) \lesssim \exp\{-r \log r\}$ if $X \sim \text{Poi}(\lambda)$ and $r > \lambda e$, we have the following results for $X = N_{\bar{\epsilon}_n}$ and $r = r_n$

$$\Pi_n(N_{\bar{\epsilon}_n} > r_n) \lesssim \exp\{-r_n \log r_n\} \lesssim \exp\{-n^{d/(d+2\beta)} (\log n)^{t_r+1}\}. \quad (23)$$

All three bounds together give

$$\begin{aligned} \Pi_n(\mathcal{F}_n^c) &\leq \Pi_n(\mathcal{F}_n^{*c}) \leq \Pi_n(S_n^c) + \Pi_n(N_{\bar{\epsilon}_n} > r_n) + \Pi_n(\boldsymbol{\mu} \notin [-b_n, b_n]^d) \\ &\lesssim \exp\{-C_7 n^{d/(d+2\beta)} (\log n)^{t_r+1}\} \end{aligned} \quad (24)$$

for some constant $C_7 > 0$, which decreases faster than $e^{-4n\bar{\epsilon}_n^2}$ if $t_r + 1 > 2t_1$.

5.3. Proof of Theorem 1 (Part II)

In order to verify (16), we split $(\underline{\sigma}_n, \bar{\sigma}_n)$ into $J_n + 1$ disjoint subsets

$$(\underline{\sigma}_n, \bar{\sigma}_n) = \bigcup_{i=1}^{J_n} (\underline{\sigma}_n(1 + \tilde{\epsilon}_n)^{i-1}, \underline{\sigma}_n(1 + \tilde{\epsilon}_n)^i) \cup (\underline{\sigma}_n(1 + \tilde{\epsilon}_n)^{J_n}, \bar{\sigma}_n) \quad (25)$$

for $J_n = \lfloor (\log \bar{\sigma}_n / \underline{\sigma}_n) / \log(1 + \tilde{\epsilon}_n) \rfloor$. Hence we obtain a partition of S_n with $(J_n + 1)^d$ subsets.

Denote $S_{n,j} = \bigotimes_{i=1}^d [\underline{\sigma}_n(1 + \tilde{\epsilon}_n)^{j_i-1}, \underline{\sigma}_n(1 + \tilde{\epsilon}_n)^{j_i} \vee \bar{\sigma}_n]$, where $j_i = 1, \dots, J_n + 1$.

Then define

$$\mathcal{F}_{n,j}^* = \left\{ \sum_{i=1}^k \omega_i \phi_{\boldsymbol{\sigma}}(\mathbf{x} - \boldsymbol{\mu}_i) : k \leq r_n, \boldsymbol{\omega} \in \Delta_k, \boldsymbol{\mu}_i \in [-b_n, b_n]^d, \boldsymbol{\sigma} \in S_{n,j} \right\},$$

$$\mathcal{F}_{n,j} = \{p_{F,\boldsymbol{\sigma}} : \text{there exist } p_{F',\boldsymbol{\sigma}} \in \mathcal{F}_{n,j}^* \text{ such that } d_{\text{TV}}(F, F') \leq \bar{\epsilon}_n\}.$$

We can bound the prior probability on $\mathcal{F}_{n,j}$ by

$$\Pi_n(\mathcal{F}_{n,j}) \leq \Pi_n(S_{n,j}) \lesssim (1 + \tilde{\epsilon}_n)^{j \cdot -1} \underline{\sigma}_n^d \bar{\epsilon}_n^d. \quad (26)$$

In order to calculate the entropy, we further decompose $\mathcal{F}_{n,j}^*$ into

$$\begin{aligned} \mathcal{F}_{n,j}^* &= \bigcup_{k=1}^{r_n} \mathcal{F}_{n,j,k}^* \\ &= \bigcup_{k=1}^{r_n} \left\{ \sum_{i=1}^k \omega_i \phi_{\sigma}(\mathbf{x} - \boldsymbol{\mu}_i) : \boldsymbol{\mu}_j \in [-b_n, b_n]^d, \sigma_i \in S_{n,j} \right\}. \end{aligned} \quad (27)$$

Using the following general results on bracketing numbers taken from Ghosal & van der Vaart (2001) and Kruijer et al. (2010),

$$D(\epsilon, \Delta_k, \|\cdot\|_1) \leq \left(\frac{5}{k}\right)^{k-1}, \quad (28)$$

$$D(\epsilon, \bigotimes_{i=1}^k [b_i, d_i], \|\cdot\|_1) \leq \frac{k! \prod_{i=1}^k (d_i - b_i + 2\epsilon)}{(2\epsilon)^k}, \quad (29)$$

we obtain the following estimates of packing numbers

$$D(\bar{\epsilon}_n, \Delta_{r_n}, \|\cdot\|_1) \leq \left(\frac{5}{\bar{\epsilon}_n}\right)^{r_n-1}, \quad (30)$$

$$D(\bar{\epsilon}_n, [-b_n, b_n]^{r_n d}, \|\cdot\|_1) \leq (r_n d)! \left(2\bar{\epsilon}_n\right)^{-r_n d} (2b_n + 2\bar{\epsilon}_n)^{r_n d}, \quad (31)$$

$$D(\bar{\epsilon}_n, S_{n,j}, \|\cdot\|_1) \leq d! (2\bar{\epsilon}_n)^{-d} \prod_{i=1}^d (\underline{\sigma}_n (1 + \tilde{\epsilon}_n)^{j_i} - \underline{\sigma}_n (1 + \tilde{\epsilon}_n)^{j_i-1} + 2\bar{\epsilon}_n), \quad (32)$$

$$D(\bar{\epsilon}_n, \mathcal{F}_{n,j,k}^*, \|\cdot\|_1) \leq D(\bar{\epsilon}_n, \Delta_k, \|\cdot\|_1) D(\bar{\epsilon}_n, [-b_n, b_n]^{r_n d}, \|\cdot\|_1) D(\bar{\epsilon}_n, S_{n,j}, \|\cdot\|_1). \quad (33)$$

Combining (30), (31), (32), and using the relationship between covering and packing numbers,

we have

$$\begin{aligned} N(3\bar{\epsilon}_n, \mathcal{F}_{n,j}, \|\cdot\|_1) &\leq r_n D(\bar{\epsilon}_n, \mathcal{F}_{n,j,r_n}^*, \|\cdot\|_1) \\ &\lesssim r_n (r_n d)! (\bar{\epsilon}_n)^{1-r_n-r_n d} b_n^{r_n d} C_8^{r_n} (n^{-A} (1 + \tilde{\epsilon}_n)^{j \cdot -1} + 2)^d \end{aligned} \quad (34)$$

for some constant $C_8 > 0$. Combining (26) and (34) and applying Stirling's formula on $(r_n d)!$,

we find that $\sqrt{N(\bar{\epsilon}_n, \mathcal{F}_{n,j}, d)} \sqrt{\Pi_n(\mathcal{F}_{n,j})}$ is bounded by a multiple of

$$\begin{aligned} &n^{-Ad/2} (1 + \tilde{\epsilon}_n)^{j \cdot /2} (r_n)^{r_n/2+3/4} (\bar{\epsilon}_n)^{(1+d)(1-r_n)/2} b_n^{r_n d/2} C_8^{r_n/2} (n^{-A} (1 + \tilde{\epsilon}_n)^{j \cdot -1} + 2)^{d/2} \\ &\lesssim n^{-Ad/2} (1 + \tilde{\epsilon}_n)^{j \cdot /2} (r_n)^{r_n/2+3/4} (\bar{\epsilon}_n)^{(1+d)(1-r_n)/2} b_n^{r_n d/2} C_8^{r_n/2} (n^{-A} (1 + \tilde{\epsilon}_n)^{j \cdot -1} \vee 2)^{d/2} \\ &\lesssim \exp\{C_9 r_n (\log n)\} (1 + \tilde{\epsilon}_n)^{j \cdot /2} (n^{-A} (1 + \tilde{\epsilon}_n)^{j \cdot -1} \vee 2)^{d/2}, \end{aligned} \quad (35)$$

for some constant $C_9 > 0$. Observe that $n^{-A}(1 + \tilde{\epsilon}_n)^{j-1} \leq 2$ implies $(1 + \tilde{\epsilon}_n)^{j/2} \lesssim n^{A/2}$.

Therefore from equation (35), we have the following:

$$\begin{aligned}
& \sum_{i=1}^d \sum_{j_i=1}^{J_n} \sqrt{N(\tilde{\epsilon}_n, \mathcal{F}_{n,j}, d)} \sqrt{\Pi_n(\mathcal{F}_{n,j})} \\
& \lesssim \sum_{i=1}^d \sum_{j_i=1}^{J_n} \exp\{C_{10}r_n(\log n)\} n^{A/2} 2^{d/2} \\
& \quad + \sum_{i=1}^d \sum_{j_i=1}^{J_n} C_9 \exp\{C_{11}r_n(\log n)\} n^{-Ad/2} (1 + \tilde{\epsilon}_n)^{(1+d)j/2} \\
& \lesssim \exp\{C_{12}r_n(\log n)\} (n^{A/2} J_n^d + n^{-Ad/2} (1 + \tilde{\epsilon}_n)^{(1+d)dJ_n/2} J_n^d)
\end{aligned} \tag{36}$$

for some new constants C_{10}, C_{11}, C_{12} . Since J_n is defined that $n^{-A}(1 + \tilde{\epsilon}_n)^{J_n} \leq \exp\{n\tilde{\epsilon}_n^2(\log n)^\delta\}$, the r.h.s of (36) is bounded by a multiple of

$$\exp\{C_{13}r_n(\log n) + n\tilde{\epsilon}_n^2(\log n)^\delta(1+d)d/2\}. \tag{37}$$

In order to let (37) increase slower than $\exp\{n\tilde{\epsilon}_n^2\}$, we need $2t_2 > \max(t_r + 1, 2t_1 + \delta)$.

Finally, we verify (17) using similar arguments as in Ghosal & van der Vaart (2007). For sufficiently large $b > 0$.

$$\begin{aligned}
& \bigcap_{i=1}^d \bigcap_{j=1}^{N_n} \{(F, \sigma) : \sum_{j=1}^{N_n} |F(U_j) - p_j| \leq \tilde{\epsilon}_n^b, F(U_j) \geq \tilde{\epsilon}_n^{2b}, |\sigma_i - \sigma_{n,i}| \leq \tilde{\epsilon}_n^{b/d} \sigma_{n,i}\} \\
& \subset \{(F, \sigma) : P_0(\log \frac{p_0}{p_{F,\sigma}})^k \lesssim \sigma_n^{2\beta}, k = 1, 2\},
\end{aligned}$$

where $N_n \lesssim \sigma^{-d} |\log \sigma|^{d/\tau_2+d}$ is obtained using Lemma 1. Applying Lemma 10 of Ghosal & van der Vaart (2007) with $N = N_n$ and $\epsilon = \tilde{\epsilon}_n^b$, the prior probability is lower bounded by a multiple of

$$\exp\{-C_{14}N_n \log_- \tilde{\epsilon}_n\} \gtrsim \exp\{-C_{14}n^{d/(2\beta+d)} (\log n)^{d/\tau_2+d+1-t_1d/\beta}\}, \tag{38}$$

which decreases more slowly than $e^{-n\tilde{\epsilon}_n^2}$ if $\frac{d}{\tau_2} + d + 1 - \frac{t_1d}{\beta} < 2t_1$. Combining with $t_2 > t_1$, $t_r + 1 > 2t_1$ and $2t_2 > \max(t_r + 1, 2t_1 + \delta)$, we obtain $t_2 > \frac{\delta}{2} + \left(\frac{d}{\tau_2} + d + 1\right) \frac{\beta}{2\beta+d}$, where δ is an arbitrary positive number and hence can be absorbed in the remaining terms.

769 The proof of Theorem 2 uses a multivariate modification to the proof in Kruijer et al. (2010)
 770 except the number of finite mixture terms changes into $k_n = O(n^{d/(d+2\beta)}(\log n)^{d/\tau+d-t_1d/\beta})$.
 771 Details are discussed in Shen & Ghosal (2011).

772

773

APPENDIX

774 The following three lemmas are helpful in controlling the KL divergence between f_0 and $K_\sigma h_\beta$ with
 775 $\sigma = (\sigma, \dots, \sigma)$. Lemmas 5 and 7 are multivariate generalizations of Lemmas 1 and 2 in Kruijer et al.
 776 (2010). The proofs follow those in the univariate case with some modification. To save space, we skip the
 777 proofs here. Details proofs are given in Shen & Ghosal (2011), the full version of the paper.

778 LEMMA 5. Given $\beta > 0$, let f_0 satisfy Condition (C1). Then for all sufficiently small σ and all \mathbf{x}
 779 contained in the set

780

$$A_\sigma = \{\mathbf{x} \in \mathbb{R}^d : |l_j(\mathbf{x})| \leq B\sigma^{-j} |\log \sigma|^{-j/2}, j = 1, 2, \dots, r,$$

781

$$|L(\mathbf{x})| \leq B\sigma^{-\beta} |\log \sigma|^{-\beta/2} d^{-\beta/2}\},$$

782

we have

783

$$K_\sigma f_\beta(\mathbf{x}) = f_0(\mathbf{x})(1 + O(\sigma^\beta)R(\mathbf{x})) + O(\sigma^H)(1 + R(\mathbf{x})), \quad (\text{A1})$$

784

785 where $R(\mathbf{x}) = s_{r+1}|L(\mathbf{x})| + \sum_{j=1}^r s_j |l_j(\mathbf{x})|^{\beta/j}$, H is a positive number that can be chosen arbitrarily
 786 large, and s_{r+1} and s_j are nonnegative constants.

787 LEMMA 6. Define $E_\sigma = \{\mathbf{x} : g(\mathbf{x}) \geq \sigma^{H_1}\}$. Assume that f_0 satisfies Conditions (C1)–(C4). Then for
 788 all all $\mathbf{i} \in \mathbb{N}^d$, sufficiently small σ and $\epsilon > 0$:

789

$$\int_{A_\sigma^c} K_\sigma^{\mathbf{i}} f_0(\mathbf{x}) d\mathbf{x} = O(\sigma^{2\beta+\epsilon}), \quad \int_{E_\sigma^c} K_\sigma^{\mathbf{i}} f_0(\mathbf{x}) d\mathbf{x} = O(\sigma^{2\beta+\epsilon}) \quad (\text{A2})$$

790

provided that H_1 is sufficiently large.

791

792 Comparing with Lemma 2 in Kruijer et al. (2010), We obtain an extra σ^ϵ at the right hand side (r.h.s)
 793 of (A2), which is needed in Proposition 1.

794

795

796

797

798

817 *Proof.* Observe $\mathbf{i} = \sum_{k=1}^{\bar{i}} \mathbf{j}_k$, $\mathbf{j}_k \in \mathbb{N}^d$, $\bar{i} = \max\{i_1, \dots, i_d\}$, where each component of \mathbf{j}_k only
 818 takes two values 0 and 1. If some components of \mathbf{j}_k are 0, then we can remove these components
 819 from our problem and consider a corresponding convolution operator in a low-dimension case. There-
 820 fore it is good enough to prove (A2) when $i_1 = \dots = i_d = m$ for $m \in \mathbb{N}$. The proof for other cases
 821 can proceed in a similar way. In order to bound the first integral in (A2), we consider sets $A_{\sigma, \delta} =$
 822 $\{\mathbf{x} : |l_j(\mathbf{x})| \leq \delta B \sigma^{-j} |\log \sigma|^{-j/2}, j = 1, \dots, r, |L(\mathbf{x})| \leq \delta B \sigma^{-\beta} |\log \sigma|^{-\beta/2} d^{-\beta/2}\}$ indexed by $\delta \leq$
 823 1. Using Markov's inequality and Condition (C3),

$$\begin{aligned}
 824 & \int_{A_{\bar{c}}} (K_{\sigma}^0 f_0)(\mathbf{x}) d\mathbf{x} \\
 825 & \leq \mathbb{P}\{|L(\mathbf{x})| \geq (\delta B)^{(2\beta+2\epsilon)/\beta} \sigma^{-2\beta-2\epsilon} |\log \sigma|^{-(2\beta+2\epsilon)/2}\} \\
 826 & \quad + \sum_{j=1}^r \mathbb{P}\{|l_j(\mathbf{x})|^{(2\beta+2\epsilon)/j} \geq (\delta B)^{(2\beta+2\epsilon)/j} \sigma^{-(2\beta+2\epsilon)} |\log \sigma|^{-(\beta+\epsilon)}\} \\
 827 & = O(\sigma^{2\beta+\epsilon}), \tag{A3}
 \end{aligned}$$

828
 829 provided that $\sigma^{-\epsilon} |\log \sigma|^{-\beta-\epsilon} > 1$ and $\epsilon > 0$, which is the case if σ is sufficiently small. This completes
 830 the proof for $m = 0$.

831 If $m = 1$, consider independent random vectors \mathbf{X} and \mathbf{U} with densities f_0 and standard normal
 832 ϕ respectively. Then $\mathbf{X} + \Sigma \mathbf{U}$ has density $K_{\sigma} f_0$. We want to prove $\mathbf{X} \in A_{\sigma, \delta}$ together with $\|\mathbf{U}\| \leq$
 833 $k' |\log \sigma|^{1/2}$ are in contradiction with $\mathbf{X} + \Sigma \mathbf{U} \in A_{\sigma}^c$ when δ is sufficiently small.

834 We observe that $\mathbf{X} + \Sigma \mathbf{U} \in A_{\sigma, 1}^c$ implies

$$835 |L(\mathbf{X} + \Sigma \mathbf{U})| \geq B \sigma^{-\beta} |\log \sigma|^{-\beta/2} d^{-\beta/2} \text{ or } |l_i(\mathbf{X} + \Sigma \mathbf{U})| \leq B \sigma^{-i} |\log \sigma|^{-i/2}$$

836
 837 for some i satisfying $i \leq r$.

838 From Condition (C1), if δ is sufficiently small, then for all $i = 1, \dots, r$,

$$\begin{aligned}
 839 & |l_i(\mathbf{X} + \Sigma \mathbf{U})| \leq \left| \sum_{j \geq i} \frac{l_j(\mathbf{X})}{(j-1)!} (\Sigma \mathbf{U})^{j-1} \right| + \sum_{j=i}^r \sum_{j \geq i} \frac{j!}{(j-i)!} |L(\mathbf{X})| \|\Sigma \mathbf{U}\|^{j-i} \\
 840 & \leq B \sigma^{-i} |\log \sigma|^{-i/2}. \tag{A4}
 \end{aligned}$$

841

842

843

844

845

865 Therefore it has to be a large value of $|L(\mathbf{X} + \Sigma\mathbf{U})|$ that forces $\mathbf{X} + \Sigma\mathbf{U}$ to be in A_σ^c . Hence it
 866 suffices to show $|L(\mathbf{X})| \leq \delta B \sigma^{-\beta} |\log \sigma|^{-\beta/2} d^{-\beta/2}$ and $\|\mathbf{U}\| \leq k' |\log \sigma|^{1/2}$ are in contradiction with
 867 $|L(\mathbf{X} + \Sigma\mathbf{U})| \geq B \sigma^{-\beta} |\log \sigma|^{-\beta/2} d^{-\beta/2}$.

868 From Condition (C1), we assume L is a polynomial of degree q and has roots z_1, \dots, z_q . Let $\boldsymbol{\eta} =$
 869 $(\max_i |z_{i1}|, \dots, \max_i |z_{id}|)$. If $|X_i| \leq \eta_i + 1$ for $i = 1, \dots, d$, then each component of $\|\mathbf{X} + \Sigma\mathbf{U}\|$ is
 870 bounded by corresponding component of $\boldsymbol{\eta} + 2$ when σ is sufficiently small. As a result, $|L(\mathbf{X} + \Sigma\mathbf{U})| \leq$
 871 $B \sigma^{-\beta} |\log \sigma|^{-\beta/2} d^{-\beta/2}$. Alternatively, if there exists a $1 \leq i^* \leq d$ such that $|X_{i^*}| > \eta_{i^*} + 1$, then we
 consider the Taylor expansion of $L(\mathbf{X} + \Sigma\mathbf{U})$:

$$872 \quad |L(\mathbf{X} + \Sigma\mathbf{U})| \leq |L(\mathbf{X})| + \left| \sum_{j=1}^q \frac{\sigma^j \mathbf{U}^j L^{(j)}(\mathbf{X})}{j!} \right| + \frac{\sigma^q \|\mathbf{U}\|^q}{q!} |L^{(q)}(\boldsymbol{\eta}) - L^{(q)}(\mathbf{X})|$$

$$873 \quad \leq \delta B (d \sigma^2 |\log \sigma|)^{-\beta/2} + \sum_{j=1}^q O(\sigma^{j-\beta} |\log \sigma|^{(j-\beta)/2}), \quad (\text{A5})$$

874 which is less than $B \sigma^{-\beta} |\log \sigma|^{-\beta/2} d^{-\beta/2}$ when $\sigma < 1$ and $\delta < 1$ are small enough.

875 Because $\mathbb{P}(\|\mathbf{U}\| \geq k' |\log \sigma|^{1/2}) = O(\sigma^{2\beta+\epsilon})$ for $\epsilon > 0$ if k' is sufficiently large, we have

$$876 \quad \mathbb{P}(\mathbf{X} + \Sigma\mathbf{U} \in A_\sigma^c)$$

$$877 \quad \leq \mathbb{P}(\mathbf{X} + \Sigma\mathbf{U} \in A_\sigma^c, \|\mathbf{U}\| \leq k' |\log \sigma|^{1/2}) + \mathbb{P}(\|\mathbf{U}\| \geq k' |\log \sigma|^{1/2})$$

$$878 \quad = \mathbb{P}(\mathbf{X} + \Sigma\mathbf{U} \in A_\sigma^c, \mathbf{X} \in A_{\sigma,\delta}, \|\mathbf{U}\| \leq k' |\log \sigma|^{1/2}) + O(\sigma^{2\beta+\epsilon})$$

$$879 \quad \quad \quad + \mathbb{P}(\mathbf{X} + \Sigma\mathbf{U} \in A_\sigma^c, \mathbf{X} \in A_{\sigma,\delta}^c, \|\mathbf{U}\| \leq k' |\log \sigma|^{1/2})$$

$$880 \quad \leq 0 + O(\sigma^{2\beta+\epsilon}) + \mathbb{P}(\mathbf{X} \in A_{\sigma,\delta}^c)$$

$$881 \quad \leq O(\sigma^{2\beta+\epsilon}). \quad (\text{A6})$$

882 This completes the proof of first equation in (A2) for $m = 1$. For $m > 1$, we can redefine the density of
 883 X as $K_\sigma^{(m-1, \dots, m-1)} f_0$ and apply the same arguments above with a decreasing sequence of δ 's.

884 Now we bound the second integral in (A2). If $m = 0$, using Condition (C2), we have

$$885 \quad \int_{E_\sigma^c} f_0(\mathbf{x}) d\mathbf{x} = \int_{E_\sigma^c} f_0(\mathbf{x}) \frac{1}{(g(\mathbf{x}))^\xi} (g(\mathbf{x}))^\xi d\mathbf{x} \lesssim \sigma^{\xi H_1} = O(\sigma^{2\beta+\epsilon}) \quad (\text{A7})$$

886 when $H_1 \geq (2\beta + \epsilon)/\xi$.

887
888
889
890
891
892
893

913 Consider $m = 1$, we define sets $E_{\sigma,\delta} = \{x : f_0(x) \geq \sigma^{\delta H_1}\}$ indexed by $\delta \leq 1$, random vectors \mathbf{X}
 914 having density f_0 and \mathbf{U} following standard normal distribution. Observe $\mathbf{X} + \Sigma\mathbf{U} \in E_{\sigma}^c \cap A_{\sigma}$ con-
 915 tradicts with $\mathbf{X} \in E_{\sigma,\delta} \cap A_{\sigma}$: on the one hand, $\mathbf{X} + \Sigma\mathbf{U} \in E_{\sigma}^c$ and $\mathbf{X} \in E_{\sigma,\delta}$ imply $|l(\mathbf{X} + \Sigma\mathbf{U}) -$
 916 $l(\mathbf{X})| \geq (1 - \delta)H_1 \log \sigma$. On the other hand, $\mathbf{X}, \mathbf{X} + \Sigma\mathbf{U} \in A_{\sigma}$ implies that $|l(\mathbf{X} + \Sigma\mathbf{U}) - l(\mathbf{X})| \leq$
 917 $B\sigma^{-d} |\log \sigma|^{-1/2} \sigma^d k' |\log \sigma|^{1/2} = O(1)$.

918 Similarly with the previous treatment, for a sufficiently large constant k' and $H_1 \geq (4\beta + 2\epsilon)/\delta$, we
 919 have

$$\begin{aligned}
 920 & \int_{E_{\sigma}^c} K_{\sigma} f_0(\mathbf{x}) d\mathbf{x} \\
 921 & \leq \int_{A_{\sigma}^c} K_{\sigma} f_0(\mathbf{x}) d\mathbf{x} + \int_{E_{\sigma}^c \cap A_{\sigma}} K_{\sigma} f_0(\mathbf{x}) d\mathbf{x} \\
 922 & \leq O(\sigma^{2\beta+\epsilon}) + \mathbf{P}(\mathbf{X} + \Sigma\mathbf{U} \in E_{\sigma}^c \cap A_{\sigma}) \\
 923 & \leq O(\sigma^{2\beta+\epsilon}) + \mathbf{P}(\mathbf{X} + \Sigma\mathbf{U} \in E_{\sigma}^c \cap A_{\sigma}, \mathbf{X} \in E_{\sigma,\delta} \cap A_{\sigma}, \|\mathbf{U}\| \leq k' |\log \sigma|^{1/2}) \\
 924 & \quad + \mathbf{P}(\mathbf{X} + \Sigma\mathbf{U} \in E_{\sigma}^c \cap A_{\sigma}, \mathbf{X} \in E_{\sigma,\delta} \cap A_{\sigma}, \|\mathbf{U}\| \leq k' |\log \sigma|^{1/2}) \\
 925 & \leq O(\sigma^{2\beta+\epsilon}) + P(\mathbf{X} \in E_{\sigma,\delta}^c) \\
 926 & = O(\sigma^{2\beta+\epsilon}). \tag{A8}
 \end{aligned}$$

928 This completes the proof for $m = 1$. The above procedure can be repeated in the same way when H_1 is
 929 chosen sufficiently large for $m > 1$. Hence we obtain (A2). \square

930
 931 LEMMA 7. Assume that f_0 satisfies Conditions (C1)–(C4). If $\beta > 2$, $\mathbf{x} \in A_{\sigma} \cap E_{\sigma}$ and σ is suffi-
 932 ciently small, then

$$933 K_{\sigma} h_{\beta}(\mathbf{x}) = f_0(\mathbf{x})(1 + O(\sigma^{\beta})R(\mathbf{x})) + O(\sigma^H)(1 + R(\mathbf{x})), \tag{A9}$$

935 where $R(\mathbf{x})$ is defined in Lemma 5.
 936

937

938

939

940

941

Remark 5. The density function $h_\beta(\mathbf{x})$ is lower bounded by a multiple of $g(\mathbf{x})$ because

$$\begin{aligned} \int h(\mathbf{x})d\mathbf{x} &= 1 + \int_{J^c} \left(\frac{1}{2}f_0 - f_\beta\right)d\mathbf{x} \\ &= 1 + \int_{J^c} \left(\frac{1}{2}f_0 - C_\beta f_0 + \sum_{j \geq 0} c_j K_{\sigma_j} f_0\right)d\mathbf{x} \\ &= 1 + O(\sigma^{2\beta}) \end{aligned} \tag{A10}$$

Therefore $K_\sigma h_\beta$ is also lower bounded by a multiple of $g(\mathbf{x})$.

Proof of Proposition 1. Using inequality $\log x \leq x - 1$ for all $x > 0$, we have

$$\int_S p \log \frac{p}{q} \leq \int_S p \frac{p-q}{q} = \int_S \frac{(p-q)^2}{q} + \int_S (p-q). \tag{A11}$$

for any densities p and q , and any set S . We apply this result for $p = f_0(\mathbf{x})$, $q = K_\sigma h_\beta(\mathbf{x})$ and $S = A_\sigma \cap E_\sigma$:

$$\begin{aligned} \int f_0(\mathbf{x}) \log \frac{f_0(\mathbf{x})}{K_\sigma h_\beta(\mathbf{x})} d\mathbf{x} &\leq \int_{A_\sigma^c \cup E_\sigma^c} f_0(\mathbf{x}) \log \frac{f_0(\mathbf{x})}{K_\sigma h_\beta(\mathbf{x})} d\mathbf{x} + \int_{A_\sigma \cap E_\sigma} (K_\sigma h_\beta(\mathbf{x}) - f_0(\mathbf{x})) d\mathbf{x} \\ &\quad + \int_{A_\sigma \cap E_\sigma} \frac{(f_0(\mathbf{x}) - K_\sigma h_\beta(\mathbf{x}))^2}{K_\sigma h_\beta(\mathbf{x})} d\mathbf{x}. \end{aligned} \tag{A12}$$

Using Remark 5, $K_\sigma h_\beta(\mathbf{x})$ is lower bounded by a multiple of $g(\mathbf{x})$. Using Hölder's inequality and Remark 1, the first term of (A12) is bounded by

$$\begin{aligned} &\int_{A_\sigma^c \cup E_\sigma^c} f_0(\mathbf{x}) \left| \log_+ \frac{f_0(\mathbf{x})}{g(\mathbf{x})} \right| d\mathbf{x} \\ &\leq \left\{ \int f_0(\mathbf{x}) \left(\log_+ \frac{f_0(\mathbf{x})}{g(\mathbf{x})} \right)^p d\mathbf{x} \right\}^{1/p} \left\{ \int_{A_\sigma^c \cup E_\sigma^c} f_0(\mathbf{x}) \right\}^{1/q} \\ &\leq C_1 \left\{ \int f_0 \left(\log_+ f_0 \right)^p + \int C_2 f_0 \left(\frac{1}{g^\xi} \right) d\mathbf{x} \right\}^{1/p} \left\{ \int_{A_\sigma^c \cup E_\sigma^c} f_0(\mathbf{x}) \right\}^{1/q} \end{aligned}$$

for constants $C_1, C_2 > 0$, $q = (2\beta + \epsilon)/2\beta$, $p = (2\beta + \epsilon)/\epsilon$ and ξ as defined in Condition (C2). By choosing H_1 such that equation (A2) in Lemma 6 holds for $\mathbf{i} = \mathbf{0}$, the first integral of the r.h.s of (A12) is $O(\sigma^{2\beta})$.

Since h_β is a linear combination of $K_\sigma^i f_0$'s, so is $K_\sigma h_\beta(\mathbf{x})$. Therefore by another application of (A2), we obtain the second integral is a finite sum of $O(\sigma^{2\beta+\epsilon})$, which is still $O(\sigma^{2\beta+\epsilon})$.

1009 For the last integral of the r.h.s of (A12), we apply Lemma 7. Observe that when $x \in A_\sigma \cap E_\sigma$, $K_\sigma(x)$
 1010 is bounded by a multiple of f_0 given $H \geq H_1$.

$$1011 \quad \left(\int_{A_\sigma \cap E_\sigma} f_0(\mathbf{x}) R^2(\mathbf{x}) d\mathbf{x} \right) O(\sigma^{2\beta}) + \left(\int_{A_\sigma \cap E_\sigma} (1 + R(\mathbf{x}))^2 / f(\mathbf{x}) d\mathbf{x} \right) O(\sigma^{2H})$$

$$1012 \quad + 2 \left(\int_{A_\sigma \cap E_\sigma} R(\mathbf{x})(1 + R(\mathbf{x})) d\mathbf{x} \right) O(\sigma^{\beta+H}) \quad (\text{A13})$$

1013 Condition (C1) implies $\int_{A_\sigma \cap E_\sigma} f_0(\mathbf{x}) R^k(\mathbf{x}) d\mathbf{x} = O(1)$ for $k = 1, 2$. By choosing H satisfying $H \geq$
 1014 $H_1 + \beta$ and using $f_0(\mathbf{x}) \geq \sigma^{H_1}$ on E_σ , these three integrals in (A13) are $O(\sigma^{2\beta})$, hence (4) follows.

1015 The integral in (5) can be treated in a similar way:

$$1016 \quad \int f_0 \left(\log \frac{f_0}{K_\sigma h_\beta} \right)^2 \leq \int_{A_\sigma^c \cup E_\sigma^c} f_0(\mathbf{x}) \left(\log \frac{f_0(\mathbf{x})}{K_\sigma h_\beta(\mathbf{x})} \right)^2 d\mathbf{x} + \int_{A_\sigma \cap E_\sigma} \frac{(f_0(\mathbf{x}) - K_\sigma h_\beta(\mathbf{x}))^2}{K_\sigma h_\beta(\mathbf{x})} d\mathbf{x}$$

$$1017 \quad + \int_{A_\sigma \cap E_\sigma} \frac{(f_0(\mathbf{x}) - K_\sigma h_\beta(\mathbf{x}))^3}{(K_\sigma h_\beta(\mathbf{x}))^2} d\mathbf{x}, \quad (\text{A14})$$

1018 where the first two terms on r.h.s are shown to be $O(\sigma^{2\beta})$, and the last integral can be bounded by a
 1019 multiple of

$$1020 \quad \int_{A_\sigma \cap E_\sigma} f_0(\mathbf{x}) R^3(\mathbf{x}) O(\sigma^{3\beta}) d\mathbf{x} + 3 \int_{A_\sigma \cap E_\sigma} R^3(\mathbf{x}) O(\sigma^{2\beta+H})$$

$$1021 \quad + 3 \int_{A_\sigma \cap E_\sigma} R^3(\mathbf{x}) O(\sigma^{\beta+2H}) / f_0(\mathbf{x}) + \int_{A_\sigma \cap E_\sigma} R^3(\mathbf{x}) O(\sigma^{3H}) / f_0^2(\mathbf{x}) d\mathbf{x}$$

$$1022 \quad = O(\sigma^{3\beta})$$

1023 by choosing $H \geq H_1 + \beta$. □

1024 *Proof of Lemma 1.* Define set $E'_\sigma = \{x : h_\beta(\mathbf{x}) \geq \sigma^{H_2}\}$ with $H_2 \geq H_1$ and $\tilde{h}_\beta(\mathbf{x}) =$
 1025 $h_\beta \mathbb{1}_{E'_\sigma}(\mathbf{x}) / \int_{E'_\sigma} h_\beta(\mathbf{x}) d\mathbf{x}$. Remark 5 implies $E'_\sigma \supset E_\sigma$. Using Lemma 6 and Remark 4, we have

$$1026 \quad \int_{E'_\sigma} h_\beta(\mathbf{x}) d\mathbf{x} = 1 - \int_{E_\sigma^c} h_\beta(\mathbf{x}) d\mathbf{x} = 1 + O(\sigma^{2\beta}). \quad (\text{A15})$$

1027 and

$$1028 \quad \int f_0 \log \frac{f_0}{p_{F,\sigma}} = \int f_0 \log \frac{f_0}{K_\sigma h_\beta} + \int_{E_\sigma} f_0 \left(\log \frac{K_\sigma h_\beta}{K_\sigma \tilde{h}_\beta} + \log \frac{K_\sigma \tilde{h}_\beta}{p_{F,\sigma}} \right) + \int_{E_\sigma^c} f_0 \log \frac{K_\sigma h_\beta}{p_{F,\sigma}} \quad (\text{A16})$$

1029 From Theorem 1, the first term is $O(\sigma^{2\beta})$. Using (A15), we observe

$$1030 \quad \frac{K_\sigma h_\beta}{K_\sigma \tilde{h}_\beta} = \frac{K_\sigma h_\beta(\mathbf{x})}{K_\sigma h_\beta \mathbb{1}_{E'_\sigma}(\mathbf{x})} (1 + O(\sigma^{2\beta})) = (1 + O(\sigma^{2\beta})) \left(1 + \frac{\int_{E'_\sigma} \phi_\sigma(\mathbf{x} - \mathbf{y}) h_\beta(\mathbf{y}) d\mathbf{y}}{\int_{E'_\sigma} \phi_\sigma(\mathbf{x} - \mathbf{y}) h_\beta(\mathbf{y}) d\mathbf{y}} \right) \quad (\text{A17})$$

1031

1032

1033

1034

1035

1036

1037

1057 For $\mathbf{x} \in E_\sigma$ and $\mathbf{y} \in E'_\sigma$, we have $g(\mathbf{x}) \geq \sigma^{H_1}$ and $h_\beta(\mathbf{x}) \leq \sigma^{H_2}$, hence $\int_{E'_\sigma} \phi_\sigma(\mathbf{x} - \mathbf{y})h_\beta(\mathbf{y})d\mathbf{y} \leq$
 1058 $\sigma^{H_2} \leq \sigma^{H_2-H_1}g(\mathbf{x})$. Using Remark 5, $\int_{E'_\sigma} \phi_\sigma(\mathbf{x} - \mathbf{y})h_\beta(\mathbf{y})d\mathbf{y} \geq C_0g(\mathbf{x})$ for constant $C_0 > 0$. There-
 1059 fore (A17) is upper bounded by $(1 + O(\sigma^{2\beta}))(1 + C_0^{-1}\sigma^{H_2-H_1}) = 1 + O(\sigma^{2\beta})$ when we choose $H_2 \geq$
 1060 $H_1 + 2\beta$. On the other hand, (A17) is lower bounded by $1 + O(\sigma^{2\beta})$. Hence $\frac{K_\sigma h_\beta}{K_\sigma \tilde{h}_\beta} = 1 + O(\sigma^{2\beta})$ and
 1061 therefore $\int_{E_\sigma} f_0 \log \frac{K_\sigma h_\beta}{K_\sigma \tilde{h}_\beta} = O(\sigma^{2\beta})$.

1062 Now we bound $\int_{E_\sigma} f_0 \log \frac{K_\sigma \tilde{h}_\beta}{p_{F,\sigma}}$. Apply Lemma 4 part (2) for $\epsilon = e^{-C_1|\log \sigma|}$ for some constant C_1 and
 1063 $\gamma_1 = 1/\tau_2$, let $p_{F,\sigma}$ be the finitely supported mixture approximating \tilde{h}_β such that $\|K_\sigma \tilde{h}_\beta - p_{F,\sigma}\|_\infty \leq$
 1064 $\sigma^{-d}e^{-C_1|\log \sigma|}$ and F has at most a multiple of $\sigma^{-d}|\log \sigma|^{d/\tau_2+d}$ many support points, which are all
 1065 contained in E'_σ because of Condition (C3). Notice that these support points are also contained in $\{\mathbf{x} :$
 1066 $f_0(\mathbf{x}) \geq c\sigma^{H_2}\}$ for sufficiently small $c > 0$ by Remark 3. Applying

$$1067 \log \frac{p(\mathbf{x})}{q(\mathbf{x})} \leq \frac{|p(\mathbf{x}) - q(\mathbf{x})|}{\min\{p(\mathbf{x}), q(\mathbf{x})\}} \leq \frac{\|p - q\|_\infty}{(\min_{\mathbf{y}} p(\mathbf{y}) - \|p - q\|_\infty)} \quad (\text{A18})$$

1068 if $\min_{\mathbf{y}} p(\mathbf{y}) - \|p - q\|_\infty > 0$ for $p = K_\sigma \tilde{h}_\beta$ and $q = p_{F,\sigma}$, we have

$$1069 \int_{E_\sigma} f_0 \log \frac{K_\sigma \tilde{h}_\beta}{p_{F,\sigma}} \leq \int_{E_\sigma} \frac{\|K_\sigma \tilde{h}_\beta - p_{F,\sigma}\|_\infty}{C_2 \sigma^{H_2} - \|K_\sigma \tilde{h}_\beta - p_{F,\sigma}\|_\infty} \lesssim \sigma^{-H_2-d} e^{-C_1|\log \sigma|} \quad (\text{A19})$$

1071 for some constant $C_2 > 0$. When σ is small enough and C_1 is large enough, the above estimate is $O(\sigma^{2\beta})$.

1072 Finally, we bound the last term in (A16). Using Lemma 3, we can add a mixture component with mean
 1073 0 and weight $\sigma^{2\beta}$ without influencing approximation results. Combine this result with the fact that $K_\sigma h_\beta$
 1074 is upper bounded by a constant $C_3 > 0$, we have

$$1075 \int_{E_\sigma^c} f_0(\mathbf{x}) \log \frac{K_\sigma h_\beta(\mathbf{x})}{p_{F,\sigma}} d\mathbf{x} \leq \sigma^{H_1\xi} \int_{E_\sigma^c} f_0(\mathbf{x}) \frac{1}{g^\xi(\mathbf{x})} \log \frac{C_3}{\sigma^{2\beta}\phi_\sigma(\mathbf{x})} d\mathbf{x}$$

$$1076 \lesssim \sigma^{H_1\xi} \int_{E_\sigma^c} f_0(\mathbf{x}) \frac{1}{g^\xi(\mathbf{x})} \|\mathbf{x}\|^2 \sigma^{-2d} d\mathbf{x}$$

$$1077 = O(\sigma^{2\beta}) \quad (\text{A20})$$

1078 when H_1 is chosen to be large enough. Hence the proof of the first equation in (9) is complete. The proof
 1079 for the second equation proceeds in the same way as in Appendix E of Kruijer et al. (2010). \square

1081
 1082
 1083
 1084
 1085

1105 *Proof of Lemma 2.* Choose $\sigma_0 = 2a/\Phi^{-1}(5/6)$ such that $N(0, \sigma_0)$ gives probability $1/3$ to $(0, 2a)$.

1106 Let $\sigma < \sigma_0$ and $\boldsymbol{\sigma} = (\sigma, \dots, \sigma)$. Then if $\mathbf{x} \in D$

$$\begin{aligned}
1107 \quad K_{\boldsymbol{\sigma}} f_0(\mathbf{x}) &\geq \int_D f_0(\boldsymbol{\theta}) \phi_{\boldsymbol{\sigma}}(\mathbf{x} - \boldsymbol{\theta}) d\boldsymbol{\theta} \\
1108 &\geq c_0 \int_D \phi_{\boldsymbol{\sigma}}(\mathbf{x} - \boldsymbol{\theta}) d\boldsymbol{\theta} \\
1109 &= c_0 \prod_{i=1}^d \left[\Phi\left(\frac{a - x_i}{\sigma_0}\right) + \Phi\left(\frac{x_i + a}{\sigma_0}\right) - 1 \right] \\
1110 &\geq \frac{c_0}{3^d}. \tag{A21}
\end{aligned}$$

1111 If $x \notin D$, then at least one of x_i 's are not in $[-a, a]$. We only consider the case $x_1 > a$, $x_2 < -a$ and
1112 $x_3, \dots, x_d \in [-a, a]$. The calculation can be done for other cases in a similar way.

$$\begin{aligned}
1113 \quad K_{\boldsymbol{\sigma}} f_0(\mathbf{x}) &\geq c \int_{-a}^{x_1} \int_{x_2}^a \int \cdots \int \prod_{i=1}^d g_i(\theta_i) \phi_{\boldsymbol{\sigma}}(\mathbf{x} - \boldsymbol{\theta}) d\boldsymbol{\theta} \\
1114 &\geq c \int_{-a}^{x_1} g_1(x_1) \phi_{\sigma_0}(x_1 - \theta_1) d\theta_1 \int_{x_2}^a g_2(x_2) \phi_{\sigma_0}(x_2 - \theta_2) d\theta_2 \frac{c_0}{3^{d-2}} \\
1115 &\geq \frac{c_0 c}{3^{d-2}} g_1(x_1) g_2(x_2) \left(\Phi\left(\frac{2a}{\sigma_0}\right) - \frac{1}{2} \right) \left(\frac{1}{2} - \Phi\left(\frac{-2a}{\sigma_0}\right) \right) \\
1116 &\geq C_0 g_1(x_1) g_2(x_2) \\
1117 &\geq C_1 g(\mathbf{x}) \tag{A22}
\end{aligned}$$

1119 for some positive constants C_0 and C_1 . □

1120 *Proof of Lemma 3.* By an easy multidimensional extension of Lemma 5 in Ghosal & van der Vaart
1121 (2007), we have

$$1122 \quad \|p_{F, \boldsymbol{\sigma}} - p_{F', \boldsymbol{\sigma}}\|_1 \lesssim \frac{\epsilon}{(\sigma_{(1)} \wedge \sigma'_{(1)})^d} + \sum_{j=1}^N |F(V(\mathbf{z}_j, \epsilon)) - p_j| \tag{A23}$$

$$1124 \quad \|p_{F, \boldsymbol{\sigma}} - p_{F', \boldsymbol{\sigma}'}\|_{\infty} \lesssim \frac{\epsilon}{(\sigma_{(1)} \wedge \sigma'_{(1)})^{2d}} + \frac{1}{(\sigma_{(1)} \wedge \sigma'_{(1)})^d} \sum_{j=1}^N |F(V(\mathbf{z}_j, \epsilon)) - p_j| \tag{A24}$$

1125 Similarly, by a multidimensional extension of Lemma 3 in Kruijer et al. (2010)

$$1126 \quad \|p_{F', \boldsymbol{\sigma}} - p_{F', \boldsymbol{\sigma}'}\|_1 \leq \|\phi_{\boldsymbol{\sigma}} - \phi_{\boldsymbol{\sigma}'}\|_1 \leq \sum_{i=1}^d \|\phi_{\sigma_i} - \phi_{\sigma'_i}\|_1 \lesssim \max_{i=1, \dots, d} \frac{|\sigma_i - \sigma'_i|}{\sigma_i \wedge \sigma'_i} \tag{A25}$$

$$1128 \quad \|p_{F', \boldsymbol{\sigma}} - p_{F', \boldsymbol{\sigma}'}\|_{\infty} \lesssim \left| \prod_{i=1}^d \frac{1}{\sigma_i} - \prod_{i=1}^d \frac{1}{\sigma'_i} \right| \leq \frac{1}{(\sigma_{(1)} \wedge \sigma'_{(1)})^{2d}} \left| \prod_{i=1}^d \sigma_i - \prod_{i=1}^d \sigma'_i \right| \tag{A26}$$

1129

1130

1131

1132

1133

1153 Using triangle inequality on (A23) and (A25) gives (12). Similarly, combining (A24) and (A26) gives
1154 (13). □

1155

1156

REFERENCES

1157

BELITSER, E. & GHOSAL, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal
1158 distribution. *Ann. Statist.* **31**, 536–559.

1159

CANALE, A. & DUNSON, D. (2011). Bayesian multivariate mixed-scale density estimation. Tech. rep.
1160 [arXiv:1110.1265v2](https://arxiv.org/abs/1110.1265v2).

1160

DE JONGE, R. & VAN ZANTEN, H. (2010). Adaptive nonparametric Bayesian inference using location-scale mixture
1161 priors. *Ann. Statist.* **38**, 3300–3320.

1162

DONOHO, D. L. & JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer.*
1163 *Statist. Assoc.* **90**, 1200–1224.

1163

ESCOBAR, M. D. & WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist.*
1164 *Assoc.* **90**, 577–588.

1165

FERGUSON, T. S. (1983). *Bayesian density estimation by mixtures of normal distributions*. Academic Press, pp.
1166 287–302.

1166

GHOSAL, S., GHOSH, J. K. & RAMAMOORTHI, R. V. (1999). Posterior consistency of Dirichlet mixtures in density
1167 estimation. *Ann. Statist.* **27**, 143–158.

1168

GHOSAL, S., GHOSH, J. K. & VAN DER VAART, A. (2000). Convergence rates of posterior distributions. *Ann.*
1169 *Statist.* **28**, 500–531.

1169

GHOSAL, S., LEMBER, J. & VAN DER VAART, A. (2008). Nonparametric Bayesian model selection and averaging.
1170 *Electron. J. Stat.* **2**, 63–89.

1171

GHOSAL, S. & VAN DER VAART, A. (2001). Entropies and rates of convergence for maximum likelihood and Bayes
1172 estimation for mixtures of normal densities. *Ann. Statist.* **29**, 1233–1263.

1172

GHOSAL, S. & VAN DER VAART, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities.
1173 *Ann. Statist.* **35**, 697–723.

1173

1174

HUANG, T.-M. (2004). Convergence rates for posterior distributions and adaptive estimation. *Ann. Statist.* **32**,
1175 1556–1593.

1175

1176

KRUIJER, W., ROUSSEAU, J. & VAN DER VAART, A. (2010). Adaptive Bayesian density estimation with location-
1177 scale mixtures. *Electron. J. Stat.* **4**, 1225–1257.

1177

1178

1179

1180

1181

- 1201 LO, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.* **12**, 351–357.
- 1202 MULIERE, P. & TARDELLA, L. (1998). Approximating distributions of random functionals of Ferguson-Dirichlet
1203 priors. *Canad. J. Statist* **26**, 283–297.
- 1204 MÜLLER, P., ERKANLI, A. & WEST, M. (1996). Bayesian curve fitting using multivariate normal mixtures.
Biometrika **83**, 67–79.
- 1205 PATI, D., DUNSON, D. & TOKDAR, S. T. (2011). Posterior consistency in conditional distribution estimation.
Unpublished manuscript, Duke University, Department of Statistical Science.
- 1206 RIGOLLET, P. (2006). Adaptive density estimation using the blockwise Stein method. *Bernoulli* **12**, 351–370.
- 1207 ROUSSEAU, J. (2010). Rates of convergence for the posterior distributions of mixtures of Betas and adaptive non-
1208 parametric estimation of the density. *Ann. Statist.* **38**, 146–180.
- 1209 SCOTT, D. W. (1992). *Multivariate Density Estimation. Thoery, Practice and Visualization*. Wiley, New York.
- 1210 SHEN, W. & GHOSAL, S. (2011). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. Tech.
rep. [arXiv:1109.6406v2](https://arxiv.org/abs/1109.6406v2) .
- 1211 SHEN, X. & WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29**, 687–714.
- 1212 TOKDAR, S. T. (2006). Posterior consistency of Dirichlet location-scale mixture of normals in density estimation
and regression. *Sankhyā* **68**, 90–110.
- 1213 VAN DER VAART, A. & VAN ZANTEN, H. (2009). Adaptive Bayesian estimation using a Gaussian random field with
1214 inverse gamma bandwidth. *Ann. Statist.* **37**, 2655–2675.
- 1215 WAND, M. P. & JONES, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.
- 1216 WU, Y. & GHOSAL, S. (2010). The L_1 -consistency of Dirichlet mixtures in multivariate Bayesian density estimation.
J. Multivar. Anal. **101**, 2411–2419.
- 1217
- 1218
- 1219
- 1220
- 1221
- 1222
- 1223
- 1224
- 1225
- 1226
- 1227
- 1228
- 1229