Submitted to the Annals of Statistics

SUPPLEMENT TO "NONPARAMETRIC MODAL REGRESSION"

By Yen-Chi Chen, Christopher R. Genovese,
Ryan J. Tibshirani, Larry Wasserman

Carnegie Mellon University

This following contains proofs of key results in the paper "Nonparametric Modal Regression". For clarity, we keep the numbering for the lemmas and theorems consistent with those in the original paper. Before we proceed, we first recall a useful theorem.

Theorem 12. *Assume (A1,K1-2). Then*

$$\|\widehat{p}_n - p\|^*_{\infty,k} = O(h^2) + O_{\mathbb{P}}\left(\sqrt{\frac{\log n}{nh^{d+1+2k}}}\right).$$

Moreover, when n is sufficiently large and $\frac{\log n}{nh^{d+1+2k}} \to 0$,

$$\mathbb{P}\left(\|\widehat{p}_n - p\|^*_{\infty,k} > \epsilon\right) \leq (k+1)e^{-Anh^{d+1+2k}\epsilon^2} \tag{1}$$

for some constant $A > 0$.

The first assertion can be proved by the same method as used in Einmahl and Mason (2000); Einmahl et al. (2005); Giné and Guillou (2002) and the second assertion is an application of Talagrand's inequality (Talagrand, 1996; Massart, 2000; Giné and Guillou, 2002). Thus we omit the proof. Similar results for the kernel density estimator can be found in Chen et al. (2014b).

Proof of Theorem 3. In this proof we write elements of $M(x)$ as y_j. It can be shown that when $\|\widehat{p}_n - p\|^*_{\infty,2}$ is sufficiently small, for every x, each the local mode $y_j \in M(x)$ corresponds to a unique closest estimated local mode $\widehat{y}_j$ by assumption (A3). See the proof to Theorem 4 in Chen et al. (2014a).

Part 1: Empirical approximation. Let x be a fixed point in D. Let y_j be a local mode and $\widehat{y}_j$ be the estimator corresponding to y_j. By definition,

$$p_y(x, y_j) = 0, \quad \widehat{p}_{y,n}(x, \widehat{y}_j) = 0.$$

By Taylor's Theorem,

$$\begin{aligned}\widehat{p}_{y,n}(x,y_j) &= \widehat{p}_{y,n}(x,y_j) - \widehat{p}_{y,n}(x,\widehat{y}_j)\\ &= (y_j-\widehat{y}_j)\widehat{p}_{yy,n}(x,y_j^*),\end{aligned} \tag{2}$$

where y_j^* is a point between y_j and $\widehat{y}_j$.

Thus, after dividing $\widehat{p}_{yy,n}(x,y_j^*)$ on both sides,

$$\begin{aligned}\widehat{y}_j - y_j &= -\widehat{p}_{yy,n}(x,y_j^*)^{-1}\widehat{p}_{y,n}(x,y_j)\\ &= -p_{yy}(x,y_j)^{-1}\widehat{p}_{y,n}(x,y_j) + O_{\mathbb{P}}(\|\widehat{p}-p\|_{\infty,2}^*)\widehat{p}_{y,n}(x,y_j).\end{aligned} \tag{3}$$

Note that we use

$$\left|\widehat{p}_{yy,n}(x,y_j^*)^{-1} - p_{yy}(x,y_j)^{-1}\right| = O_{\mathbb{P}}(\|\widehat{p}-p\|_{\infty,2}^*).$$

This is valid since both $p_{yy}, \widehat{p}_{yy,n}$ are bounded away from 0 when x, y are sufficiently close to $\mathcal{S}$ (the modal manifold collection) by assumption (A3). Thus, the inverse is bounded above by (A1) and (K1).

Therefore, by taking absolute values we obtain

$$|\widehat{y}_j - y_j| - |p_{yy}(x,y_j)^{-1}||\widehat{p}_{y,n}(x,y_j)| = O_{\mathbb{P}}(\|\widehat{p}-p\|_{\infty,2}^* \cdot |\widehat{p}_{y,n}(x,y_j)|).$$

Taking a maximum over all local modes, and using $\Delta_n(x) = \max|\widehat{y}_j - y_j|$, we have

$$\begin{aligned}\Delta_n(x) - \max_j\left\{|p_{yy}(x,y_j)^{-1}||\widehat{p}_{y,n}(x,y_j)|\right\}&\\ = O_{\mathbb{P}}\left(\|\widehat{p}-p\|_{\infty,2}^* \cdot \max_j\{|\widehat{p}_{y,n}(x,y_j)|\}\right),&\end{aligned} \tag{4}$$

which implies

$$\begin{aligned}\max_j\{|\widehat{p}_{y,n}(x,y_j)|\}^{-1}\left|\Delta_n(x) - \max_j\left\{|p_{yy}(x,y_j)^{-1}||\widehat{p}_{y,n}(x,y_j)|\right\}\right|&\\ = O_{\mathbb{P}}\left(\|\widehat{p}-p\|_{\infty,2}^*\right).&\end{aligned}$$

Thus, $\Delta_n(x)$ can be approximated by $\max_j\left\{|p_{yy}(x,y_j)^{-1}||\widehat{p}_{y,n}(x,y_j)\right\}$. We note that $|p_{yy}(x,y_j)^{-1}|$ is bounded from above and below by (A1-3), so $\max_j\{|\widehat{p}_{y,n}(x,y_j)|\}$ shares the rate of $\max_j\left\{|p_{yy}(x,y_j)^{-1}||\widehat{p}_{y,n}(x,y_j)|\right\}$, and equation (4) implies

$$\frac{1}{\Delta_n(x)}\left|\Delta_n(x) - \max_j\left\{|p_{yy}(x,y_j)^{-1}||\widehat{p}_{y,n}(x,y_j)|\right\}\right| = O_{\mathbb{P}}\left(\|\widehat{p}-p\|_{\infty,2}^*\right). \tag{5}$$

According to its definition, $A_n(x)$ is the same as left-hand side of (5) whenever $\Delta_n(x) > 0$, and as the right-hand side of this expression does not depend on x, we can take the supremum over $x \in D$ to establish the first assertion in the theorem.

Part 2: Rate of convergence. For each j, we focus on $\widehat{p}_{y,n}(x, y_j)$ since $p_{yy}(x, y_j)^{-1}$ is bounded:

$$\begin{aligned}|\widehat{p}_{y,n}(x, y_j)| &= |\widehat{p}_{y,n}(x, y_j) - p_y(x, y_j)| \\ &\leq |\widehat{p}_{y,n}(x, y_j) - \mathbb{E}\left(\widehat{p}_{y,n}(x, y_j)\right)| + |\mathbb{E}\left(\widehat{p}_{y,n}(x, y_j)\right) - p_y(x, y_j)| \\ &= O_{\mathbb{P}}\left(\sqrt{\frac{1}{nh^{d+3}}}\right) + O(h^2).\end{aligned}$$

The last step follows from the usual bias-variance tradeoff for the kernel density estimator. By repeating the above argument, the rate holds for every local mode. Since there are at most $K(x) < \infty$ local modes for fixed x, the rate is the same as we take the maximum over all local modes. Thus, we have proved the second assertion. □

Proof of Theorem 4. By Theorem 3,

$$\begin{aligned}\Delta_n(x) &= \max_j \left\{|p_{yy}(x, y_j)^{-1}||\widehat{p}_{y,n}(x, y_j)|\right\} + o_{\mathbb{P}}(1) \\ &= \max_j \left\{|p_{yy}(x, y_j)^{-1}|\left(|\widehat{p}_{y,n}(x, y_j) - \mathbb{E}\left(\widehat{p}_{y,n}(x, y_j)\right)| + B(x, y_j)\right)\right\} + o_{\mathbb{P}}(1),\end{aligned}$$

where $B(x, y_j) = |\mathbb{E}\left(\widehat{p}_{y,n}(x, y_j)\right) - p_y(x, y_j)| = O(h^2)$ denotes the bias, and the $o_{\mathbb{P}}(1)$ term is from $O_{\mathbb{P}}(\|\widehat{p} - p\|^*_{\infty,2}\Delta_n(x))$.

Since $|p_{yy}(x, y_j)^{-1}|$ is bounded, the above implies

$$\Delta_n(x) = \max_j \left\{|p_{yy}(x, y_j)^{-1}|\,|\widehat{p}_{y,n}(x, y_j) - \mathbb{E}\left(\widehat{p}_{y,n}(x, y_j)\right)|\right\} + O(h^2) + o_{\mathbb{P}}(1).$$

Note the big O term involves the bias and is independent of x. Thus, taking supremum over $x \in D$ yields

$$\Delta_n = \mathbf{Z} + O(h^2) + o_{\mathbb{P}}(1), \tag{6}$$

where

$$\mathbf{Z} = \sup_{x \in D} \max_j \left\{|p_{yy}(x, y_j)^{-1}|\,|\widehat{p}_{y,n}(x, y_j) - \mathbb{E}\left(\widehat{p}_{y,n}(x, y_j)\right)|\right\},$$

the maximum over a stochastic process. Now let

$$\mathcal{F}_0 = \left\{ (u,v) \mapsto f_{x,y}(u,v) : f_{x,y}(u,v) = p_{yy}^{-1}(x,y) \cdot K\left(\frac{\|x-u\|}{h}\right) K^{(1)}\left(\frac{y-v}{h}\right), y \in M(x), x \in \mathbb{R} \right\}$$

be a function space similar to the function space defined in (15) of the original paper (the original function space also appears in (8)). Recall that $K^{(\alpha)}$ denotes the α-th order derivative of K. We define the empirical process $\mathbb{G}_n$ to be

$$\mathbb{G}_n(f) = \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} f(Z_i) - \mathbb{E}(f(Z_i)) \right), \quad f \in \mathcal{F}_0, \tag{7}$$

where $Z_i = (X_i, Y_i)$ is the observed data. Thus,

$$\begin{aligned} \mathbf{Z} &= \sup_{x \in D} \max_j \left\{ |p_{yy}(x,y_j)^{-1}| \, |\widehat{p}_{y,n}(x,y_j) - \mathbb{E}\left(\widehat{p}_{y,n}(x,y_j)\right)| \right\} \\ &= \frac{1}{\sqrt{n}h^{d+3}} \sup_{f \in \mathcal{F}_0} |\mathbb{G}_n(f)|. \end{aligned}$$

By assumption (A1) and (K1–2), $\mathcal{F}_0$ is a VC-type class with constant envelope C_K^2/λ_2. Thus, applying Theorem 2.3 in Giné and Guillou (2002) gives

$$\mathbf{Z} = \sup_{x \in D} \max_j \left\{ |p_{yy}(x,y_j)^{-1}| \, |\widehat{p}_{y,n}(x,y_j) - \mathbb{E}\left(\widehat{p}_{y,n}(x,y_j)\right)| \right\} | = O_{\mathbb{P}}\left(\sqrt{\frac{\log n}{nh^{d+3}}} \right).$$

Now by equation (6), the result follows. □

Proof of Theorem 5. By applying to Theorem 3, the expected square of the local error can be written as

$$\mathbb{E}\left(\Delta_n^2(x)\right) = O(h^4) + O\left(\frac{1}{nh^{d+3}}\right) = \mathsf{Bias}^2(x) + \mathsf{Variance}(x).$$

Using arguments in Chacón et al. (2011); Chacón and Duong (2013), the integrated bias and variance over x yields the same rate of convergence. □

Proof of Theorem 7. We follow a strategy similar to that used in the proof of Theorem 6 of Chen et al. (2014b). Let $\mathcal{F}$ be the function space

defined in (15) of the original paper:

$$\mathcal{F} = \Bigg\{ (u,v) \mapsto f_{x,y}(u,v) : f_{x,y}(u,v) = \widetilde{p}_{yy}^{-1}(x,y) \cdot K\left(\frac{\|x-u\|}{h}\right) K^{(1)}\left(\frac{y-v}{h}\right), x \in D, y \in \widetilde{M}(x) \Bigg\}. \tag{8}$$

Recall the definition of $\mathbb{G}_n$ in (7), and the definition of $\mathbb{B}$ in (17) of the original paper. Denote

$$\mathbf{G}_n = \frac{1}{\sqrt{h^{d+3}}} \sup_{f \in \mathcal{F}} |\mathbb{G}_n(f)|, \quad \mathbf{B} = \frac{1}{\sqrt{h^{d+3}}} \sup_{f \in \mathcal{F}} |\mathbb{B}(f)|.$$

Our proof consists of three steps:

1. establish a coupling between $\sqrt{nh^{d+3}}\widetilde{\Delta}_n$ and $\mathbf{G}_n$;
2. establish a coupling between $\mathbf{G}_n$ and $\mathbf{B}$;
3. apply Gaussian anti-concentration Chernozhukov et al. (2014a, 2012) to obtain a Berry-Esseen bound between $\sqrt{nh^{d+3}}\widetilde{\Delta}_n$ and $\mathbf{B}$.

Step 1. Our goal is to show that

$$\mathbb{P}\left(\left|\sqrt{nh^{d+3}}\widetilde{\Delta}_n - \mathbf{G}_n\right| > \epsilon\right) \leq D_1 e^{-D_2 nh^{d+5}\epsilon^2}, \tag{9}$$

for some constants D_1, D_2. Recall Corollary 6, which shows that

$$\left|\sqrt{nh^{d+3}}\widetilde{\Delta}_n - \mathbf{G}_n\right| = O(\epsilon_{n,2}) = O\left(\sup_{x,y} |\widehat{p}_{yy,n}(x,y) - \mathbb{E}\left(\widehat{p}_{yy,n}(x,y)\right)|\right).$$

Thus, there exists a constant $D_0 > 0$ such that

$$\left|\sqrt{nh^{d+3}}\widetilde{\Delta}_n - \mathbf{G}_n\right| \leq D_0 \sup_{x,y} |\widehat{p}_{yy,n}(x,y) - \mathbb{E}\left(\widehat{p}_{yy,n}(x,y)\right)|.$$

By Talagrand's inequality (equation (1) in Theorem 12),

$$\begin{aligned} &\mathbb{P}\left(\left|\sqrt{nh^{d+3}}\widetilde{\Delta}_n - \mathbf{G}_n\right| > \epsilon\right) \\ &\qquad \leq \mathbb{P}\left(\sup_{x,y} |\widehat{p}_{yy,n}(x,y) - \mathbb{E}\left(\widehat{p}_{yy,n}(x,y)\right)| > \epsilon/D_0\right) \\ &\qquad \leq D_1 e^{-D_2 nh^{d+5}\epsilon^2}, \end{aligned} \tag{10}$$

for some constants $D_1, D_2 > 0$. This gives the desired result.

Step 2. We will show that

$$\mathbb{P}\left(|\mathbf{G}_n - \mathbf{B}| > A_1 \frac{b_0 \log^{2/3} n}{\gamma^{1/3}(nh^{d+3})^{1/6}}\right) \leq A_2\gamma, \tag{11}$$

for some constants A_1, A_2. We first recall a useful Gaussian approximation result from Chernozhukov et al. (2014a) and Chernozhukov et al. (2014b).

THEOREM 13 (Corollary 2.2 in Chernozhukov et al. (2014b); Theorem A.1 in Chernozhukov et al. (2014a)). *Let $\mathcal{G}$ be a collection of functions that is a VC-type class (see condition (K2)) with a constant envelope function b. Let σ^2 be a constant such that $\sup_{g\in\mathcal{G}} \mathbb{E}[g(X_i)^2] \leq \sigma^2 \leq b^2$. Let $\mathbb{B}$ be a centered, tight Gaussian process defined on $\mathcal{G}$ with covariance function*

$$\mathrm{Cov}(\mathbb{B}(g_1), \mathbb{B}(g_2)) = \mathbb{E}[g_1(X_i)g_2(X_i)] - \mathbb{E}[g_1(X_i)]\mathbb{E}[g_2(X_i)],$$

where $g_1, g_2 \in \mathcal{G}$. Then for any $\gamma \in (0,1)$ as n is sufficiently large, there exists a random variable $\mathbf{B}' \stackrel{d}{=} \sup_{f\in\mathcal{G}} |\mathbb{B}(g)|$ such that

$$\mathbb{P}\left(\left|\sup_{f\in\mathcal{G}} |\mathbb{G}_n(g)| - \mathbf{B}'\right| > A_1 \frac{b^{1/3}\sigma^{2/3}\log^{2/3} n}{\gamma^{1/3} n^{1/6}}\right) \leq A_2\gamma,$$

where A_1, A_2 are two universal constants. Note that $A \stackrel{d}{=} B$ for random variables A, B means that A and B have the same distribution.

To apply Theorem 13, we need to verify the conditions. By assumptions (K2) and (A2)$'$, $\mathcal{F}$ is a VC-type class with constant envelope $b_0 = C_K^2\widetilde{\lambda}_2 < \infty$. Note that $1/\lambda_2$ is the bound on the inverse second derivative of $\widetilde{p}_{yy}(x,y)$, for y close to a local mode. As for σ^2, by definition,

$$\sup_{f\in\mathcal{F}} \mathbb{E}[f(X_i)^2] \leq h^{d+3} b_0^2.$$

Thus, we can pick $\sigma^2 = h^{d+3}b_0^2 \leq b_0^2$ if $h \leq 1$. Hence, applying Theorem 13 gives

$$\mathbb{P}\left(\left|\sup_{f\in\mathcal{F}} |\mathbb{G}_n(f)| - \mathbf{B}'\right| > A_1 \frac{b_0 h^{2/3} h^2 \log^{2/3} n}{\gamma^{1/3} n^{1/6}}\right) \leq A_2\gamma,$$

for some constants A_1, A_2 and $\gamma < 1$ and $\mathbf{B}' \stackrel{d}{=} \sup_{f\in\mathcal{F}} |\mathbb{B}(f)|$, where $\mathbb{B}$ is a tight Gaussian process defined on $\mathcal{F}$.

Now multiplying both sides of the above expression by $\sqrt{h^{-d-3}}$, and using the definition of $\mathbf{G}_n$ and the fact that $\frac{1}{\sqrt{h^{d+3}}}\mathbf{B}' = \mathbf{B}$, we have

$$\mathbb{P}\left(|\mathbf{G}_n - \mathbf{B}| > A_1 \frac{b_0 \log^{2/3} n}{\gamma^{1/3}(nh^{d+3})^{1/6}}\right) \leq A_2\gamma, \tag{12}$$

which gives the desired result. Note that if we instead consider the function space $\mathcal{F}_1 = \{\frac{1}{\sqrt{h^{d+3}}} f : f \in \mathcal{F}\}$, then

$$\mathbf{B} = \frac{1}{\sqrt{h^{d+3}}}\mathbf{B}' \stackrel{d}{=} \frac{1}{\sqrt{h^{d+3}}} \sup_{f\in\mathcal{F}} |\mathbb{B}(f)| \stackrel{d}{=} \sup_{f\in\mathcal{F}_1} |\mathbb{B}(f)|,$$

so $\mathbf{B}$ is the maximum of a Gaussian process. We have $\mathbb{E}(\mathbf{B}) = O(\sqrt{\log n})$ by Dudley's inequality for Gaussian processes (Van Der Vaart and Wellner, 1996) so that $\mathbf{B}$ is not tight; however, this is not a problem since (12) bounds the difference between $\mathbf{G}_n$ and $\mathbf{B}$.

Step 3. We first show a coupling between $\sqrt{nh^{d+3}}\widetilde{\Delta}_n$ and $\mathbf{B}$. We pick $\epsilon = (nh^{d+5})^{-1/4}$ in (9) so that

$$\mathbb{P}\left(\left|\sqrt{nh^{d+3}}\widetilde{\Delta}_n - \mathbf{G}_n\right| > (nh^{d+5})^{-1/4}\right) \leq D_1 e^{-D_2\sqrt{nh^{d+5}}}.$$

For n sufficiently large, by the triangle inequality along with (11),

$$\mathbb{P}\left(\left|\sqrt{nh^{d+3}}\widetilde{\Delta}_n - \mathbf{B}\right| > A_3\frac{\log^{2/3} n}{\gamma^{1/3}(nh^{d+3})^{1/6}}\right) \leq A_4\gamma,$$

for some constants $A_3, A_4 > 0$. Note that we absorbed the rate $(nh^{d+5})^{-1/4}$ into $A_3 \log^{2/3} n/(\gamma^{1/3}(nh^{d+3})^{1/6})$. This is valid since $(nh^{d+5})^{-1/4}$ converges faster. Also, we absorbed $D_1 e^{-D_2\sqrt{nh^{d+5}}}$ into $A_4\gamma$. We allow $\gamma \to 0$ as long as γ converges at rate slower than $(nh^{d+5})^{-1/4}$.

Lastly, to obtain the desired Berry-Esseen bound, we apply a Gaussian approximation result in Kolmogorov distance, which is given in Lemma 2.3 Chernozhukov et al. (2014b) (this is an application of the anti-concentration inequality in Chernozhukov et al. (2014a)).

LEMMA 14 (Modification of Lemma 2.3 in Chernozhukov et al. (2014b)). *Let* $\mathbf{B}$ *be defined as the above. Assume (K1-2) and that there exists a random variable* Y *such that* $\mathbb{P}(|Y - \mathbf{B}| > \eta) < \delta(\eta)$. *Then*

$$\sup_t |\mathbb{P}(Y < t) - \mathbb{P}(\mathbf{B} < t)| \leq A_5\mathbb{E}(\mathbf{B})\eta + \delta(\eta),$$

for some constant A_5.

This modification follows from Remark 2.5 and Remark 3.2 of Chernozhukov et al. (2014b); in these remarks, they discuss how to obtain the desired bound for kernel density estimators under similar assumptions to

(K1–2). Note that in Lemma 2.3 of Chernozhukov et al. (2014b), $\mathbb{E}(\mathbf{B})$ should be replaced by $\mathbb{E}(\mathbf{B}) + \log \eta$. We can ignore $\log \eta$ since it is small compared to $\mathbb{E}(\mathbf{B})$.

By Lemma 14, we conclude that

$$\sup_t \left| \mathbb{P}\left(\sqrt{nh^{d+3}}\widetilde{\Delta}_n < t\right) - \mathbb{P}\left(\mathbf{B} < t\right) \right| \leq A_5 \mathbb{E}(\mathbf{B}) \left(A_3 \frac{\log^{2/3} n}{\gamma^{1/3} (nh^{d+3})^{1/6}} \right) + A_4 \gamma$$
$$= A_6 \left(A_3 \frac{\log^{7/6} n}{\gamma^{1/3} (nh^{d+3})^{1/6}} \right) + A_4 \gamma,$$

for some $A_6 > 0$. Taking $\gamma = \left(\frac{\log^7 n}{nh^{d+3}}\right)^{1/8}$, we have established the theorem. □

Proof of Theorem 8. This proof is essentially the same as proof to Theorem 7 in Chen et al. (2014b), following from Theorem 7 of the current paper. We state the basic ideas and omit the details. Note that the function space (8) depends on the probability measure $\mathbb{P}$ and bandwidth h,

$$\mathcal{F} = \mathcal{F}(\mathbb{P}, h) = \Bigg\{ (u,v) \mapsto f_{x,y}(u,v) : f_{x,y}(u,v) = \widetilde{p}_{yy}^{-1}(x,y) \cdot$$
$$K\left(\frac{\|x-u\|}{h}\right) K^{(1)}\left(\frac{y-v}{h}\right), \, x \in D, \, y \in \widetilde{M}(x) \Bigg\},$$

since the index y is defined over the smoothed local mode $\widetilde{M}(x)$, and both $\widetilde{M}(x)$ and $\widetilde{p}(x,y)$ depend on $\mathbb{P}$ and h. For the bootstrap estimate, Theorem 7 implies that $\widehat{\Delta}_n^*$ can be approximated by the maximum of a certain Gaussian process

$$\sup_{f \in \mathcal{F}(\mathbb{P}_n, h)} |\mathbb{B}(f)|,$$

where the function space above now depends on $\mathbb{P}_n$ and h, i.e., the role of $\mathbb{P}$ is completely replaced by $\mathbb{P}_n$. For the function space, the index y takes values at the estimated local modes $\widehat{M}_n(x)$, and $\widetilde{p}_{yy}(x,y)$ will be replaced by the second derivative of KDE $\widehat{p}_n(x,y)$. Both quantities now are determined by the empirical measure $\mathbb{P}_n$ and the smoothing parameter h.

By Lemmas 17, 19, and 20 in Chen et al. (2014b), the maxima of the Gaussian processes defined over the function spaces $\mathcal{F}(\mathbb{P}, h)$ and $\mathcal{F}(\mathbb{P}_n, h)$ will agree asymptotically. Putting this together, the result follows from the approximation

$$\widehat{\Delta}_n^* \approx \sup_{f \in \mathcal{F}(\mathbb{P}_n, h)} |\mathbb{B}(f)| \approx \sup_{f \in \mathcal{F}(\mathbb{P}, h)} |\mathbb{B}(f)| \approx \widetilde{\Delta}_n.$$

□

Before we prove Theorem 10, we first prove the following useful lemma on Gaussian mixtures and corresponding local modes.

LEMMA 15 (Gaussian mixture and local modes). *Consider a Gaussian mixture density* $p(y) = \sum_{j=1}^{K} \pi_j \phi(y; \mu_j, \sigma_j^2)$, *with* $\mu_1 < \ldots < \mu_K$ *and* $y \in \mathbb{R}$. *Let* $W = \Delta_{\min}/\sigma_{\max}$ *and* $\Delta_{\min} = \min\{|\mu_j - \mu_i| : i \neq j\}$ *and* $\sigma_{\max} = \max_j \sigma_j$. *If*

$$W \geq \sqrt{2 \log \left(4(K \vee 3 - 1) \frac{\pi_{\max}}{\pi_{\min}} \right)},$$

then

$$\max_{j=1,\ldots K} |\mu_j - m_j| \leq \sigma_{\max} \times 4 \frac{\pi_{\max}}{\pi_{\min}} \frac{1}{W} e^{-\frac{W^2}{2}}.$$

PROOF. Given any set of parameters $\{\pi_j, \mu_j, \sigma_j^2 : j = 1, \ldots, K\}$, we consider another mixture (but not necessarily a density)

$$h(y) = \pi_{\min} \phi(y; \mu_1, \sigma_{\max}^2) + \sum_{j=2}^{K} \phi(y; \mu_1 + (j-1)\Delta_{\min}, \sigma_{\max}^2).$$

We assume

(MK) $h(y)$ has K distinct local modes.

Note that this implies $p(y)$ has K distinct local modes. Later we will derive a sufficient condition for this assumption. Let the ordered local modes of $h(y)$ be $m_1' < \ldots < m_K'$. Then

$$|m_1' - \mu_1| \geq \max_{j=1,\ldots K} |m_j - \mu_j|.$$

We define s_1 such that

$$h(\mu_1 + s_1) = h(\mu_1), \quad h(s) \geq h(\mu_1), \quad \forall s \in [\mu_1, \mu_1 + s_1].$$

It is easy to see that $m_1' \leq s_1 + \mu_1$ since m_1' is the smallest (in terms of location) local mode of h. Thus, if we can bound s_1, we bound the difference $|m_1' - \mu_1|$. Note that s_1 must be very small (at least smaller than $\sigma_{\max}$) otherwise we will not obtain K local modes.

Now by the definition of h, we can find s_1 such that

$$\begin{aligned} h(\mu_1) &= \pi_{\min}\phi(\mu_1;\mu_1,\sigma^2_{\max}) + \pi_{\max}\sum_{j=2}^{K}\phi(\mu_1;\mu_1+(j-1)\Delta_{\min},\sigma^2_{\max}) \\ &= \frac{1}{\sqrt{2\pi\sigma^2_{\max}}}\pi_{\min} + \frac{1}{\sqrt{2\pi\sigma^2_{\max}}}\pi_{\max}\sum_{j=2}^{K} e^{-\frac{1}{2}(\frac{(j-1)\Delta_{\min}}{\sigma_{\max}})^2} \\ &= h(\mu_1+s_1) \\ &= \frac{1}{\sqrt{2\pi\sigma^2_{\max}}}\pi_{\min}e^{-\frac{1}{2}(\frac{s_1}{\sigma_{\max}})^2} + \frac{1}{\sqrt{2\pi\sigma^2_{\max}}}\pi_{\max}\sum_{j=2}^{K} e^{-\frac{1}{2}(\frac{(j-1)\Delta_{\min}-s_1}{\sigma_{\max}})^2}. \end{aligned}$$

Therefore, s_1 can be obtained by solving

(13)
$$\pi_{\min}\left(1-e^{-\frac{1}{2}(\frac{s_1}{\sigma_{\max}})^2}\right) = \pi_{\max}\sum_{j=2}^{K} e^{-\frac{1}{2}(\frac{(j-1)\Delta_{\min}}{\sigma_{\max}})^2}(e^{\frac{(j-1)\Delta_{\min}}{\sigma_{\max}}\frac{s_1}{\sigma_{\max}}-\frac{s_1^2}{2\sigma^2_{\max}}}-1)$$

Note that $e^x < 1+2x$ if $x<1$. Thus, when

$$\frac{(j-1)\Delta_{\min}}{\sigma_{\max}}\frac{s_1}{\sigma_{\max}} < 1, \tag{14}$$

we have

$$e^{\frac{(j-1)\Delta_{\min}}{\sigma_{\max}}\frac{s_1}{\sigma_{\max}}-\frac{s_1^2}{2\sigma^2_{\max}}} - 1 < 2\frac{(j-1)\Delta_{\min}}{\sigma_{\max}}\frac{s_1}{\sigma_{\max}} = 2(j-1)W\frac{s_1}{\sigma_{\max}}, \tag{15}$$

where $W = \frac{\Delta_{\min}}{\sigma_{\max}}$. Also note that

$$1-e^{-\frac{1}{2}(\frac{s_1}{\sigma_{\max}})^2} > \frac{1}{2}\left(\frac{s_1}{\sigma_{\max}}\right)^2, \tag{16}$$

since $s_1 < \sigma_{\max}$. Let s_2 be a small number satisfying

$$\begin{aligned} \frac{1}{2}(\frac{s_2}{\sigma_{\max}})^2 &= 2\frac{\pi_{\max}}{\pi_{\min}}W\frac{s_2}{\sigma_{\max}}\int_1^{\infty} xe^{-\frac{W^2}{2}x^2}dx \\ &= \frac{s_2}{\sigma_{\max}}\frac{\pi_{\max}}{\pi_{\min}}\frac{2}{W}e^{-\frac{W^2}{2}} \\ &\geq W\frac{\pi_{\max}}{\pi_{\min}}\frac{s_2}{\sigma_{\max}}\sum_{j=1}^{K} e^{-\frac{1}{2}j^2W^2}j^2 \\ &= \pi_{\max}\sum_{j=2}^{K} e^{-\frac{1}{2}(\frac{(j-1)\Delta_{\min}}{\sigma_{\max}})^2}2(j-1)Ws_2\frac{\pi_{\max}}{\pi_{\min}}, \end{aligned}$$

where we have used (13), (15) and (16). The above result gives

$$s_2 = \sigma_{\max} \times \frac{\pi_{\max}}{\pi_{\min}} \frac{4}{W} e^{-W^2/2} > s_1 \geq \max_j |m_j - \mu_j|, \tag{17}$$

which is the desired result.

Note that the above method requires (14), which requires

$$\frac{1}{K-1}\frac{1}{W} > \frac{s_2}{\sigma_{\max}} = \frac{\pi_{\max}}{\pi_{\min}} \frac{4}{W} e^{-W^2/2}.$$

This is true whenever

$$W > \sqrt{2 \log \left(4(K-1) \frac{\pi_{\max}}{\pi_{\min}} \right)}, \tag{18}$$

which gives one part of the condition in this Lemma.

Finally, recall that we assume (MK) at the beginning. We now prove that when W is sufficiently large, (MK) holds. It is easy to see that

$$\begin{aligned} &|\mu_i - \mu_j| > \Delta_{\min} \\ \Rightarrow\ &|m_i - m_j| > \Delta_{\min} - 2 \max_i |m_i - \mu_i|. \end{aligned}$$

Thus, as long as $\Delta_{\min} - 2\max_i |m_i - \mu_i| > 0$, there exists K distinct local modes for $p(y)$.

By equation (17), a sufficient condition to $\Delta_{\min} - 2\max_i |m_i - \mu_i| > 0$ is

$$\Delta_{\min} > 2\sigma_{\max} \times \frac{\pi_{\max}}{\pi_{\min}} \frac{4}{W} e^{-W^2/2}, \tag{19}$$

which is equivalent to

$$W^2 e^{W^2/2} > 8 \frac{\pi_{\max}}{\pi_{\min}}.$$

When $W > 1$ (which is satisfied by (20)), we see that

$$e^{W^2/2} > 8 \frac{\pi_{\max}}{\pi_{\min}}$$

implies (19), so that a sufficient condition for $p(y)$ having K distinct local modes is

$$W > \sqrt{2 \log \left(8 \frac{\pi_{\max}}{\pi_{\min}} \right)}. \tag{20}$$

Combining this condition and equation (18) completes the proof. □

PROOF OF THEOREM 10. The proof consists of four steps. The first three steps consider pointwise prediction sets, and the last uniform prediction sets. In summary, the four steps are as follows:

1. we prove that
$$\epsilon_{1-\alpha}(x) \leq z_{1-\alpha/2}\sigma_{\max}(x) + \max_i |u_i(x) - m_i(x)|,$$
where $m_1(x) < m_2(x) < \ldots < m_{K(x)}(x)$ are the ordered local modes;
2. we prove that $\eta_{1-\alpha}(x) \geq \frac{1}{2}K(x)\Delta_{\min}(x)$;
3. we apply Lemma 15 to bound $\max_i |u_i(x) - m_i(x)|$ by $\Delta_{\min}(x)$ and use the first two steps to conclude the desired pointwise result;
4. we extend the first three steps to the uniform case.

Step 1. By assumption (GP), the set
$$A = \bigcup_{j=1}^{K(x)} \mu_j(x) \oplus \big(z_{1-\alpha/2}\sigma_j(x)\big)$$
is a level $(1-\alpha)$ prediction set. Let $m_1(x) < m_2(x) < \ldots < m_{K(x)}(x)$ be the ordered local modes of $p(y|x)$. Then we have

(21)
$$\begin{aligned}\mu_j(x) \oplus \big(z_{1-\alpha/2}\sigma_j(x)\big) &\subseteq m_j(x) \oplus \big(z_{1-\alpha/2}\sigma_j(x) + |\mu_j(x) - m_j(x)|\big) \\ &\subseteq m_j(x) \oplus \left(z_{1-\alpha/2}\sigma_{\max}(x) + \max_j |\mu_j(x) - m_j(x)|\right).\end{aligned}$$

This holds for all j. Tthe regression mode set is $M(x) = \{m_1(x), \ldots, m_{K(x)}(x)\}$, so that
$$A \subseteq M(x) \oplus \left(z_{1-\alpha/2}\sigma_{\max}(x) + \max_j |\mu_j(x) - m_j(x)|\right),$$
which implies
$$\epsilon_{1-\alpha}(x) \leq z_{1-\alpha/2}\sigma_{\max}(x) + \max_j |\mu_j(x) - m_j(x)|, \tag{22}$$
since $\epsilon_{1-\alpha}(x)$ is the smallest size to construct a pointwise prediction set with with $1-\alpha$ prediction accuracy.

Step 2. We pick α such that $\alpha < \pi_1(x), \pi_{K(x)}(x)$. The prediction set from the regression function must contain all the mixture centers. Thus,
$$2\eta_{1-\alpha}(x) \geq \mu_{K(x)}(x) - \mu_1(x) \geq (K(x)-1)\Delta_{\min}(x).$$

Step 3. The length of prediction set $\mathcal{P}_{1-\alpha} = M(x) \oplus \epsilon_{1-\alpha}(x)$ is $2K(x)\epsilon_{1-\alpha}(x)$ and the length of prediction set $\mathcal{R}_{1-\alpha} = m(x) \oplus \eta_{1-\alpha}(x)$ is $2\eta_{1-\alpha}(x)$. Thus, we need to show that

$$\eta_{1-\alpha}(x) > K(x)\epsilon_{1-\alpha}(x). \tag{23}$$

By (22) and Step 2, a sufficient condition for (23) is

$$(K(x)-1)\Delta_{\min}(x) > K(x)\left(z_{1-\alpha/2}\sigma_{\max}(x) + \max_j |\mu_j(x) - m_j(x)|\right).$$

The last term can be bounded by Lemma 15, which shows that

$$\max_j |\mu_j(x) - m_j(x)| \le \sigma_{\max}(x) \times 4\frac{\pi_{\max}(x)}{\pi_{\min}(x)}\frac{1}{W(x)}e^{-\frac{W(x)^2}{2}}, \tag{24}$$

whenever

$$W(x) \ge \sqrt{2\log\left(4(K(x)\vee 3 - 1)\frac{\pi_{\max}(x)}{\pi_{\min}(x)}\right)}, \tag{25}$$

where $W(x) = \Delta_{\min}(x)/\sigma_{\max}(x)$. For convenience, we assume that $\alpha < 0.1$. This implies that $z_{1-\alpha/2} > 1.64$. To simplify matters, we wish to bound $\max_j |\mu_j(x) - m_j(x)|$ by $0.1 \times z_{1-\alpha/2}\sigma_{\max}(x)$ so that we can have a reference rule that only depends on $z_{1-\alpha/2}$. To attain this, we use (24) so that what we need is

$$\begin{aligned} 4\frac{\pi_{\max}(x)}{\pi_{\min}(x)}\frac{1}{W(x)}e^{-\frac{W(x)^2}{2}} &\le 4\frac{\pi_{\max}(x)}{\pi_{\min}(x)}e^{-\frac{W(x)^2}{2}} \\ &\le 0.1 \times z_{1-\alpha/2} \\ &< 0.1 \times 1.64. \end{aligned} \tag{26}$$

Thus, a sufficient condition is

$$\begin{aligned} W(x) &> \sqrt{2\log\left(\frac{40}{1.64}\frac{\pi_{\max}(x)}{\pi_{\min}(x)}\right)} \\ &= \sqrt{6.4 + 2\log\left(\frac{\pi_{\max}(x)}{\pi_{\min}(x)}\right)}. \end{aligned} \tag{27}$$

Hence, when $W(x) > \sqrt{6.4 + 2\log\left(\frac{\pi_{\max}(x)}{\pi_{\min}(x)}\right)}$, the condition

$$(K(x)-1)\Delta_{\min}(x) > 1.1 \times K(x) z_{1-\alpha/2}\sigma_{\max}(x) \tag{28}$$

implies that modal regression has a smaller prediction set. Condition (28) is equivalent to

$$W(x) > 1.1 \times \frac{K(x)}{K(x)-1} z_{1-\alpha/2}, \tag{29}$$

after rearrangement. Now the conditions on $W(x) = \Delta_{\min}(x)/\sigma_{\max}(x)$ involve equations (25),(27) and (29). We conclude that whenever

$$\begin{aligned} W(x) &= \frac{\Delta_{\min}(x)}{\sigma_{\max}(x)} \\ &> \max\left\{1.1 \times \frac{K(x)}{K(x)-1} z_{1-\alpha/2}, \right. \\ &\qquad \left. \sqrt{6.4 \vee 2\log\left(4(K(x) \vee 3 - 1)\right) + 2\log\left(\frac{\pi_{\max}(x)}{\pi_{\min}(x)}\right)}\right\}, \end{aligned}$$

the prediction set $\mathcal{P}_{1-\alpha}(x)$ is smaller than $\mathcal{R}_{1-\alpha}(x)$.

Step 4. Finally, we consider the uniform case. Note that

$$\begin{aligned} \epsilon_{1-\alpha} &\le \sup_x \epsilon_{1-\alpha}(x), \\ \eta_{1-\alpha} &\ge \inf_x \eta_{1-\alpha}(x). \end{aligned}$$

Therefore,

$$\begin{aligned} \epsilon_{1-\alpha} &\le \sup_x \epsilon_{1-\alpha}(x) \\ &\le \sup_x \left(z_{1-\alpha/2}\sigma_{\max}(x) + \max_j |\mu_j(x) - m_j(x)| \right) \\ &\le z_{1-\alpha/2}\sigma_{\max} + \sup_x \max_j |\mu_j(x) - m_j(x)|, \end{aligned} \tag{30}$$

and similarly

$$\begin{aligned} \eta_{1-\alpha} &\ge \inf_x \eta_{1-\alpha}(x) \\ &\ge \inf_x (K(x)-1)\Delta_{\min}(x) \\ &\ge (K_{\min}-1)\Delta_{\min}. \end{aligned} \tag{31}$$

Note that the second term in the last inequality of (30) can be bounded by

$$\begin{aligned} \sup_x \max_j |\mu_j(x) - m_j(x)| &\le \sup_x \sigma_{\max}(x) \times 4\frac{\pi_{\max}(x)}{\pi_{\min}(x)} \frac{1}{W(x)} e^{-W(x)^2/2} \\ &\le \sigma_{\max} \times 4\frac{\pi_{\max}}{\pi_{\min}} \frac{1}{W} e^{-W^2/2}, \end{aligned} \tag{32}$$

where $W = \Delta_{\min}/\sigma_{\max} \leq W(x)$.

Now using (26), and combining (30) and (32), we see ethat

$$\epsilon_{1-\alpha} \leq 1.1 \times z_{1-\alpha/2}\sigma_{\max} \tag{33}$$

whenever

$$W = \frac{\Delta_{\min}}{\sigma_{\max}} > \max\left\{\sqrt{6.4 \vee 2\log\left(4(K_{\max} \vee 3 - 1)\right) + 2\log\left(\frac{\pi_{\max}}{\pi_{\min}}\right)}\right\}. \tag{34}$$

The volume of the prediction sets $\mathcal{P}_{1-\alpha}$ and $\mathcal{R}_{1-\alpha}$ is

$$\mathsf{Vol}(\mathcal{P}_{1-\alpha}) = 2\epsilon_{1-\alpha}\int_D K(x)dx, \qquad \mathsf{Vol}(\mathcal{R}_{1-\alpha}) = 2\eta_{1-\alpha}\int_D dx.$$

Thus, $\mathsf{Vol}(\mathcal{P}_{1-\alpha}) < \mathsf{Vol}(\mathcal{R}_{1-\alpha})$ if and only if

$$\epsilon_{1-\alpha}\overline{K} < \eta_{1-\alpha}. \tag{35}$$

Applying equation (33) and (32) to (35), we require that

$$\eta_{1-\alpha} \geq (K_{\min} - 1)\Delta_{\min} > \overline{K} \times 1.1 \times z_{1-\alpha/2}\sigma_{\max} \geq \epsilon_{1-\alpha}$$

which leads to

$$\frac{\Delta_{\min}}{\sigma_{\max}} > 1.1 \times \frac{\overline{K}}{K_{\min} - 1} z_{1-\alpha/2}.$$

Combining this condition and (34) completes the proof. □

Proof for Lemma 11. Let the Hessian matrix of $p(x, y)$ be $H \equiv H(x, y)$. The eigenvalues of H are

$$\begin{aligned} \lambda_1(x, y) &= \mathsf{tr}(H)/2 + \sqrt{\mathsf{tr}(H)^2/2 - \mathsf{det}(H)} \\ \lambda_2(x, y) &= \mathsf{tr}(H)/2 - \sqrt{\mathsf{tr}(H)^2/2 - \mathsf{det}(H)}, \end{aligned} \tag{36}$$

and the corresponding eigenvectors are

$$v_1(x, y) = \begin{bmatrix} \lambda_1(x, y) - H_{22} \\ H_{21} \end{bmatrix} \quad v_2(x, y) = \begin{bmatrix} \lambda_2(x, y) - H_{22} \\ H_{21} \end{bmatrix},$$

where H_{ij} is the (i, j) element of H and $\mathsf{tr}(H)$ is the trace of H and $\mathsf{det}(H)$ is the determinant of H.

Thus, $\lambda_2(x, y) < 0$ if and only if ($\mathsf{tr}(H) < 0$ or $\mathsf{det}(H) < 0$). Namely,

$$\lambda_2(x, y) < 0 \iff \left(H_{11} + H_{22} < 0 \text{ or } H_{11}H_{22} < H_{12}^2\right).$$

However, since $y \in M(x)$, $H_{22} < 0$. This implies $\lambda_2(x,y) < 0$. (since whatever the sign of H_{11} is, one of the above conditions must hold)

Thus, all we need is to show $v_2^T(x,y)\nabla p(x,y) = 0$. By the formula for eigenvectors,

$$\begin{aligned} v_2^T(x,y)\nabla p(x,y) &= (\lambda_2(x,y) - H_{22})p_x(x,y) + H_{21}p_y(x,y) \\ &= (\lambda_2(x,y) - H_{22})p_x(x,y) \end{aligned}$$

since $p_y(x,y) = 0$ for $y \in M(x)$. Therefore, $v_2^T(x,y)\nabla p(x,y) = 0$ if and only if $p_x(x,y) = 0$ or $\lambda_2(x,y) - H_{22} = 0$. The former case corresponds to the first condition and by (36), $\lambda_2(x,y) = H_{22}$ if and only if $H_{12} = 0$. This completes the proof.

□

References.

J. Chacón and T. Duong. Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. *Electronic Journal of Statistics*, 7: 499–532, 2013.

J. E. Chacón, T. Duong, and M. Wand. Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, 21:807–840, 2011.

Y.-C. Chen, C. R. Genovese, and L. Wasserman. Generalized mode and ridge estimation. arXiv: 1406.1803, 2014a.

Y.-C. Chen, C. R. Genovese, and L. Wasserman. Asymptotic theory for density ridges. arXiv: 1406.5663, 2014b.

V. Chernozhukov, D. Chetverikov, and K. Kato. Comparison and anti-concentration bounds for maxima of gaussian random vectors. *Probability Theory and Related Fields*, pages 1–24, 2012.

V. Chernozhukov, D. Chetverikov, K. Kato, et al. Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5):1787–1818, 2014a.

V. Chernozhukov, D. Chetverikov, K. Kato, et al. Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597, 2014b.

U. Einmahl and D. M. Mason. An empirical process approach to the uniform consistency of kernel-type function estimators. *Journal of Theoretical Probability*, 13(1):1–37, 2000.

U. Einmahl, D. M. Mason, et al. Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*, 33(3):1380–1403, 2005.

E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier, 2002.

P. Massart. About the constants in talagrand's concentration inequalities for empirical processes. *Annals of Probability*, pages 863–884, 2000.

M. Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, 1996.

A. W. Van Der Vaart and J. A. Wellner. *Weak Convergence*. Springer, New York, 1996.

Department of Statistics
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
E-mail: yenchic@andrew.cmu.edu