Submitted to the Annals of Statistics
arXiv: arXiv:1101.5783

SUPPLEMENT TO "OPTIMAL WEIGHTED NEAREST NEIGHBOUR CLASSIFIERS"

BY RICHARD J. SAMWORTH,

Statistical Laboratory, University of Cambridge

This is the supplementary material to Samworth (2012).

COMPLETION OF THE PROOF OF THEOREM 1. **Step 6**: *To show (6.5), which gives the Radon–Nikodym derivative of the restriction of the distribution of $X_{(i)} - x$ to a small ball about the origin*

Recall that $B_\delta(u) = \{y \in \mathbb{R}^d : \|y - u\| \leq \delta\}$, and that ν_d denotes d-dimensional Lebesgue measure. For a Borel subset A of $\mathbb{R}^d$, let $N(A) = \sum_{i=1}^n \mathbb{1}_{\{X_i \in A\}}$. It follows from the hypothesis **(A.2)** that for $x \in \mathcal{S}^{\epsilon_n}$ with n sufficiently large, and for $i \leq k_2$, the restriction of the distribution of $X_{(i)} - x$ to a small ball about the origin is absolutely continuous with respect to ν_d. Thus for $x \in \mathcal{S}^{\epsilon_n}$ with n sufficiently large, for $i \leq k_2$, for $u \neq 0$ with $\|u\|$ sufficiently small, and for $\delta < \|u\|$,

$$
\begin{aligned}
&\frac{\mathbb{P}\{X_{(i)} - x \in B_\delta(u)\}}{\nu_d(B_\delta(u))} \\
&\geq \frac{1}{a_d \delta^d} \mathbb{P}\bigl\{N(B_\delta(x+u)) = 1, N(B_{\|u\|-\delta}(x)) = i-1, N(B_{\|u\|+\delta}(x)) = n-i\bigr\} \\
&= \frac{n}{a_d \delta^d} \biggl\{\int_{B_\delta(x+u)} \bar{f}(v)\,dv\biggr\} \binom{n-1}{i-1} p_{\|u\|-\delta}^{i-1} (1 - p_{\|u\|+\delta})^{n-i} \\
&\to n\bar{f}(x+u) \binom{n-1}{i-1} p_{\|u\|}^{i-1} (1 - p_{\|u\|})^{n-i}
\end{aligned}
$$

as $\delta \to 0$, by the Lebesgue differentiation theorem. For the other bound, write $A = B_{\|u\|+\delta}(x) \setminus B_{\|u\|-\delta}(x)$ and observe that

$$
\begin{aligned}
\frac{\mathbb{P}\{X_{(i)} - x \in B_\delta(u)\}}{\nu_d(B_\delta(u))} &\leq \frac{1}{a_d \delta^d} \Bigl[\mathbb{P}\{X_{(i)} \in B_\delta(x+u) \cap N(A) = 1\} \\
&\qquad\qquad + \mathbb{P}\{N(B_\delta(x+u)) \geq 1 \cap N(A) \geq 2\}\Bigr] \\
&= \frac{1}{a_d \delta^d} \biggl[n \biggl\{\int_{B_\delta(x+u)} \bar{f}(v)\,dv\biggr\} \binom{n-1}{i-1} p_{\|u\|-\delta}^{i-1} (1 - p_{\|u\|+\delta})^{n-i} + o(\delta^d)\biggr] \\
&\to n\bar{f}(x+u) \binom{n-1}{i-1} p_{\|u\|}^{i-1} (1 - p_{\|u\|})^{n-i}
\end{aligned}
$$

as $\delta \to 0$. The result therefore follows by Folland (1999, Theorem 3.22).

To show (6.4), which bounds R_1: We have

$$\begin{aligned} R_1 = \sum_{i=k_2+1}^{n} w_{ni}[\mathbb{E}\{\eta(X_{(i)})\} - \eta(x)] + \sum_{i=1}^{k_2} w_{ni}\Bigl[\mathbb{E}\{\eta(X_{(i)})\} - \eta(x) \\ - \mathbb{E}\{(X_{(i)} - x)^T \dot{\eta}(x)\} - \frac{1}{2}\mathbb{E}\{(X_{(i)} - x)^T \ddot{\eta}(x)(X_{(i)} - x)\}\Bigr] \\ \equiv R_{11} + R_{12}, \end{aligned}$$

say. Now $\eta \leq 1$, so

$$\sup_{\mathbf{w}_n \in W_{n,\beta}} \sup_{x \in \mathcal{S}^{\epsilon_n}} \frac{|R_{11}|}{t_n} \leq \sup_{\mathbf{w}_n \in W_{n,\beta}} \frac{\sum_{i=k_2+1}^{n} w_{ni}}{t_n} \leq 1/\log n.$$

To handle R_{12}, observe that by a Taylor expansion, given $\epsilon > 0$, we can find $\delta > 0$ such that for all sufficiently large n, all $x \in \mathcal{S}^{\epsilon_n}$ and all $\|y - x\| \leq \delta$, we have

$$|\eta(y) - \eta(x) - (y - x)^T \dot{\eta}(x) - \tfrac{1}{2}(y - x)^T \ddot{\eta}(x)(y - x)| \leq \epsilon \|y - x\|^2.$$

For $1 \leq i \leq k_2$, let $A_i = \{\|X_{(i)} - x\| \leq \delta\}$. Further, let $D_1 = \sup_{x \in \mathcal{S}} \|\dot{\eta}(x)\|$ and let $D_2 = \sup_{x \in \mathcal{S}} \lambda_{\max}\{\ddot{\eta}(x)\}$, where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a matrix. Then for large n and $x \in \mathcal{S}^{\epsilon_n}$,

$$\begin{aligned} |R_{12}| \leq \epsilon \sum_{i=1}^{k_2} w_{ni} \mathbb{E}(\|X_{(i)} - x\|^2 \mathbb{1}_{A_i}) + 2\mathbb{P}(A_i^c) + (1 + D_1)\mathbb{E}\{\|X_{(i)} - x\| \mathbb{1}_{A_i^c}\} \\ + (1 + D_2)\mathbb{E}\{\|X_{(i)} - x\|^2 \mathbb{1}_{A_i^c}\}. \end{aligned} \tag{0.1}$$

We can apply a very similar argument to that employed in **Step 1** to deduce that uniformly for $1 \leq i \leq k_2$,

$$\sup_{x \in \mathcal{S}^{\epsilon_n}} \mathbb{E}(\|X_{(i)} - x\|^2 \mathbb{1}_{A_i}) = O\{(i/n)^{2/d}\}. \tag{0.2}$$

Now, using notation defined in **Step 1**,

$$\begin{aligned} \mathbb{E}(\|X_{(i)} - x\|^2 \mathbb{1}_{A_i^c}) = \mathbb{P}(\|X_{(i)} - x\| > \delta) + \int_{\delta^2}^{\infty} \mathbb{P}(\|X_{(i)} - x\| > t^{1/2})\, dt \\ \leq q_\delta^n(i) + \int_{\delta^2}^{\infty} q_{t^{1/2}}^n(i)\, dt. \end{aligned} \tag{0.3}$$

For $\delta > 0$ sufficiently small, there exists $c_2 > 0$ such that for all $x \in \mathcal{S}^{\epsilon_n}$, we have $np_\delta - k_2 \geq c_2 n\delta^d$. So by Hoeffding's inequality, for any $t_0 > \delta^2$,

$$\sup_{x \in \mathcal{S}^{\epsilon_n}} \sup_{1 \leq i \leq k_2} \biggl\{ q_\delta^n(i) + \int_{\delta^2}^{t_0} q_{t^{1/2}}^n(i)\,dt \biggr\} \leq (1+t_0)e^{-2c_2^2 n\delta^{2d}} = O(n^{-M}), \tag{0.4}$$

for every $M > 0$. Moreover, using the moment bound in (**A.3**),

$$1 - p_t = \bar{P}(\{x + u : \|u\| > t\}) = O(t^{-\rho}), \tag{0.5}$$

as $t \to \infty$, uniformly for $x \in \mathcal{S}^{\epsilon_n}$. Therefore we can apply Bennett's inequality (Shorack and Wellner, 1986, p.440) to show that there exist $c_3, c_4 > 0$ such that for sufficiently large n and t_0 and all $t > t_0$,

$$\sup_{x \in \mathcal{S}^{\epsilon_n}} \sup_{1 \leq i \leq k_2} q_{t^{1/2}}^n(i) \leq \exp\biggl[-c_3 n \biggl\{\log\Bigl(1 + \frac{c_3 n}{1 - p_{t^{1/2}}}\Bigr) - 1\biggr\}\biggr] \leq (1 + c_4 t^{\frac{\rho}{2}})^{-\frac{c_3 n}{2}}. \tag{0.6}$$

We deduce from (6.26), (6.27) and (6.29) that

$$\sup_{x \in \mathcal{S}^{\epsilon_n}} \sup_{1 \leq i \leq k_2} \mathbb{E}\{\|X_{(i)} - x\|^2 \mathbb{1}_{A_i^c}\} = O(n^{-M}) \tag{0.7}$$

for all $M > 0$. This result, combined with (6.25) and Markov's inequality applied to the two central terms in (6.24) proves (6.4) as required.

To show (6.21), which bounds R_2: Observe that by **Step 1** and **Step 2**, there exist constants $c_5, C_2 > 0$ such that

$$\inf_{x_0 \in \mathcal{S}} \inf_{C_2 t_n \leq |t| \leq \epsilon_n} \frac{|1/2 - \mu_n(x_0^t)|}{\sigma_n(x_0^t)} \geq \frac{c_5|t|}{s_n},$$

uniformly for $\mathbf{w}_n \in W_{n,\beta}$. Hence,

$$\begin{aligned} |R_2| \leq{}& \frac{C_1 \sum_{i=1}^n w_{ni}^3}{s_n^3} \int_{\mathcal{S}} \int_{|t| \leq C_2 t_n} |t| \|\dot{\psi}(x_0)\|\,dt\,d\mathrm{Vol}^{d-1}(x_0) \\ &+ \frac{C_1 \sum_{i=1}^n w_{ni}^3}{s_n^3} \int_{\mathcal{S}} \int_{C_2 t_n < |t| \leq \epsilon_n} \frac{|t| \|\dot{\psi}(x_0)\|}{1 + c_5^3 |t|^3 / s_n^3}\,dt\,d\mathrm{Vol}^{d-1}(x_0) \\ ={}& o(t_n^2 + s_n^2), \end{aligned}$$

uniformly for $\mathbf{w}_n \in W_{n,\beta}$, as required.

To show (6.22), which bounds R_3. Let

$$r_{x_0} = \frac{-a(x_0)}{\|\dot{\eta}(x_0)\|} \frac{t_n}{s_n}.$$

Using the results of **Step 1** and **Step 2**, given $\epsilon \in (0, \inf_{x_0 \in \mathcal{S}} \|\dot{\eta}(x_0)\|)$ sufficiently small, for large n we have that for all $\mathbf{w}_n \in W_{n,\beta}$, all $x_0 \in \mathcal{S}$ and all $r \in [-\epsilon_n/s_n, \epsilon_n/s_n]$ that

$$\left| \frac{1/2 - \mu_n(x_0^{rs_n})}{\sigma_n(x_0^{rs_n})} - \bigl\{-2\|\dot{\eta}(x_0)\|(r - r_{x_0})\bigr\} \right| \leq \epsilon^2(|r| + t_n/s_n).$$

It follows that for large n,

$$\begin{aligned} &\left|\Phi\biggl\{\frac{1/2 - \mu_n(x_0^{rs_n})}{\sigma_n(x_0^{rs_n})}\biggr\} - \Phi\biggl\{-2\|\dot{\eta}(x_0)\|r - \frac{2t_n}{s_n}a(x_0)\biggr\}\right| \\ &\qquad \leq \begin{cases} 1 & \text{if } |r - r_{x_0}| \leq \epsilon t_n/s_n \\ \epsilon^2(|r| + t_n/s_n)\phi(\|\dot{\eta}(x_0)\||r - r_{x_0}|) & \text{if } \epsilon t_n/s_n < |r| < \epsilon_n/s_n. \end{cases} \end{aligned}$$

We deduce that for large n,

$$\begin{aligned} &\int_{-\epsilon_n}^{\epsilon_n} |t|\|\dot{\psi}(x_0)\| \left|\Phi\biggl\{\frac{1/2 - \mu_n(x_0^t)}{\sigma_n(x_0^t)}\biggr\} - \Phi\biggl\{\frac{2}{s_n}\bigl(-\|\dot{\eta}(x_0)\|t - a(x_0)t_n\bigr)\biggr\}\right| dt \\ &\leq \epsilon s_n^2 \int_{|r - r_{x_0}| \leq \epsilon t_n/s_n} |r|\,dr + s_n^2 \int_{-\infty}^{\infty} \epsilon^2(|r| + t_n/s_n)\phi(\|\dot{\eta}(x_0)\||r - r_{x_0}|)\,dr \\ &\leq \epsilon(s_n^2 + t_n^2). \end{aligned}$$

This allows us to conclude (6.22).

To show (6.23), which bounds R_4. We have

$$\begin{aligned} |R_4| &= s_n^2 \int_{\mathcal{S}} \int_{|r| > \epsilon_n/s_n} |r| \bigl[\Phi\{-2\|\dot{\eta}(x_0)\|(r - r_{x_0})\} - \mathbb{1}_{\{r<0\}}\bigr]\,dr\,d\mathrm{Vol}^{d-1}(x_0) \\ &\leq 2s_n^2 \int_{\mathcal{S}} \int_{r > \epsilon_n/s_n} |r|\Phi(-\|\dot{\eta}(x_0)\|r)\,dr\,d\mathrm{Vol}^{d-1}(x_0) = o(s_n^2), \end{aligned}$$

uniformly for $\mathbf{w}_n \in W_{n,\beta}$, as required. □

Proof that (**A.1**)–(**A.4**) imply the margin condition (2.1). For the upper bound, recall from (6.16) that by the mean value theorem, for sufficiently small $\epsilon > 0$,

$$\inf_{x \in \mathcal{R} \setminus \mathcal{S}^\epsilon} |\eta(x) - 1/2| \geq c_*\epsilon,$$

where we may take $c_* = \inf_{x \in U_0} \|\dot{\eta}(x)\|$, which is positive. By shrinking U_0 if necessary, we may assume that $D^* \equiv \sup_{x \in U_0} \bar{f}(x) < \infty$, and it follows that for small $\epsilon > 0$,

$$\mathbb{P}(|\eta(X) - 1/2| \leq \epsilon \cap X \in \mathcal{R}) \leq \bar{P}(\mathcal{S}^{\epsilon/c_*}) \leq D^*\nu_d(\mathcal{S}^{\epsilon/c_*}) \leq C\epsilon,$$

where $C < \infty$, using Weyl's tube formula (Gray, 2004).

For the lower bound, we construct a tube similar to $\mathcal{S}^\epsilon$, but contained in $\mathcal{R}$. To do this, let $\mathcal{S}_{\epsilon\epsilon} = \{x \in \mathcal{S} : \mathrm{dist}(x, \partial\mathcal{S}) > \epsilon\}$ and let

$$\mathcal{S}_\epsilon = \biggl\{x_0 + t\frac{\dot{\eta}(x_0)}{\|\dot{\eta}(x_0)\|} \, : \, x_0 \in \mathcal{S}_{\epsilon\epsilon}, |t| < \epsilon\biggr\}.$$

Further, let $C^* = \sup_{x \in U_0} \|\dot{\eta}(x)\|$, which is finite. Again by the mean value theorem, for sufficiently small $\epsilon > 0$,

$$\sup_{x \in \mathcal{S}_\epsilon} |\eta(x) - 1/2| \leq C^*\epsilon.$$

Thus, letting $d_* = \inf_{x \in U_0} \bar{f}(x) > 0$, for sufficiently small $\epsilon > 0$,

$$\mathbb{P}(|\eta(X) - 1/2| \leq \epsilon \cap X \in \mathcal{R}) \geq \bar{P}(\mathcal{S}_{\epsilon/C^*}) \geq d_*\nu_d(\mathcal{S}_{\epsilon/C^*}) \geq c\epsilon,$$

where $c > 0$, again using Weyl's tube formula. □

Proof of Theorem 2. Consider any vector $\mathbf{w}_n^{**} = (w_{ni}^{**})_{i=1}^n$ of non-negative weights that minimises the function $\gamma_n(\cdot)$ defined in the statement of Theorem 1. Since s_n^2 is symmetric in $w_{n1}, \ldots, w_{nn}$, while α_i is increasing in i, we see that $(w_{ni}^{**})_{i=1}^n$ is decreasing in i. We let $k^{**} = \max\{i : w_{ni}^{**} > 0\}$. Now form the Lagrangian

$$L(\mathbf{w}_n, \lambda) = \frac{1}{2}B_1 s_n^2 + \frac{1}{2}B_2 t_n^2 + \lambda\biggl(\sum_{i=1}^{k^{**}} w_{ni} - 1\biggr).$$

Then for some λ^{**},

$$0 = \frac{\partial L}{\partial w_{ni}}\bigg|_{(\mathbf{w}_n^{**}, \lambda^{**})} = B_1 w_{ni}^{**} + \frac{B_2}{n^{4/d}}\alpha_i \sum_{j=1}^{k^{**}} \alpha_j w_{nj}^{**} + \lambda^{**}, \tag{0.8}$$

for $i = 1, \ldots, k^{**}$. By summing (0.8) from $i = 1, \ldots, k^{**}$, and then multiplying (0.8) by α_i and again summing from $i = 1, \ldots, k^{**}$, we obtain two linear equations in $\sum_{j=1}^{k^{**}} \alpha_j w_{nj}^{**}$ and λ^{**}, which can be solved and substituted back into (0.8) to yield

$$w_{ni}^{**} = \frac{\frac{1}{k^{**}}\{1 + \frac{B_2}{B_1 n^{4/d}} \sum_{j=1}^{k^{**}} \alpha_j^2\}}{1 + \frac{B_2}{B_1 n^{4/d}}\{\sum_{j=1}^{k^{**}} \alpha_j^2 - (k^{**})^{1+4/d}\}} - \frac{\alpha_i \, \frac{B_2 (k^{**})^{2/d}}{B_1 n^{4/d}}}{1 + \frac{B_2}{B_1 n^{4/d}}\{\sum_{j=1}^{k^{**}} \alpha_j^2 - (k^{**})^{1+4/d}\}}$$

for $i = 1, \ldots, k^{**}$. In particular, $\mathbf{w}_n^{**}$ is the unique minimiser of $\gamma_n(\cdot)$. The weight vector $\mathbf{w}_n^*$ is asymptotically equivalent to $\mathbf{w}_n^{**}$ in the following sense:

elementary calculations reveal that $k^{**} = k^*\{1 + O((k^*)^{-1})\}$, and moreover that

$$\sum_{i=1}^{k^{**}} \alpha_i w_{ni}^{**} = \frac{(d+2)(k^*)^{2/d}}{d+4}\{1 + O((k^*)^{-1})\} \tag{0.9}$$

$$\sum_{i=1}^{k^{**}} (w_{ni}^{**})^2 = \frac{2(d+2)}{(d+4)k^*}\{1 + O((k^*)^{-1})\}. \tag{0.10}$$

The expressions corresponding to (6.32) and (6.33) when w_{ni}^* replaces w_{ni}^{**} are the same, except that the relative error is now $O((k^*)^{-2})$ in both cases. It follows immediately that

$$\frac{R_{\mathcal{R}}(\hat{C}^{\mathrm{wnn}}_{n,\mathbf{w}_n^{**}}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})}{R_{\mathcal{R}}(\hat{C}^{\mathrm{wnn}}_{n,\mathbf{w}_n^{*}}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})} = \frac{\gamma_n(\mathbf{w}_n^{**})}{\gamma_n(\mathbf{w}_n^{*})}\{1 + o(1)\} \to 1,$$

and therefore that (2.5) holds.

Arguing similarly to the above, we see that the pair of conditions that $\sum_{i=1}^n w_{ni}^2 / \sum_{i=1}^n (w_{ni}^*)^2 \to 1$ and $\sum_{i=1}^n \alpha_i w_{ni} / \sum_{i=1}^n \alpha_i w_{ni}^* \to 1$, or equivalently (2.6), are sufficient for (2.5) to hold. To see the necessity of these conditions, suppose for now that for some small $\beta > 0$, the weight vector $\mathbf{w}_n \in W_{n,\beta}$ satisfies

$$\frac{\sum_{i=1}^n w_{ni}^2}{\sum_{i=1}^n (w_{ni}^*)^2} \equiv \tau_n \to \tau \in [0,1). \tag{0.11}$$

Then, by almost the same Lagrangian calculation as that above, we have

$$\liminf_{n\to\infty} \frac{R_{\mathcal{R}}(\hat{C}^{\mathrm{wnn}}_{n,\mathbf{w}_n}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})}{R_{\mathcal{R}}(\hat{C}^{\mathrm{wnn}}_{n,\tilde{\mathbf{w}}_n}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})} \geq 1,$$

where $\tilde{\mathbf{w}}_n = (\tilde{w}_{ni})_{i=1}^n$ is given by

$$\tilde{w}_{ni} = \frac{1}{\tilde{k}}\biggl(1 + \frac{d}{2} - \frac{d\,\alpha_i}{2\tilde{k}^{2/d}}\biggr)\mathbb{1}_{\{1 \leq i \leq \tilde{k}\}},$$

and where $\tilde{k}/k^* \to 1/\tau$. It follows that for small $\beta > 0$, and for any $\mathbf{w}_n \in W_{n,\beta}$ satisfying $\sum_{i=1}^n w_{ni}^2 / \sum_{i=1}^n (w_{ni}^*)^2 \leq \tau_n$, we have

$$\begin{aligned}\liminf_{n\to\infty} \frac{R_{\mathcal{R}}(\hat{C}^{\mathrm{wnn}}_{n,\mathbf{w}_n}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})}{R_{\mathcal{R}}(\hat{C}^{\mathrm{wnn}}_{n,\mathbf{w}_n^*}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})} &\geq \lim_{n\to\infty} \frac{R_{\mathcal{R}}(\hat{C}^{\mathrm{wnn}}_{n,\tilde{\mathbf{w}}_n}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})}{R_{\mathcal{R}}(\hat{C}^{\mathrm{wnn}}_{n,\mathbf{w}_n^*}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})} \\ &= \frac{1}{d+4}\Bigl(\frac{d}{\tau^{4/d}} + 4\tau\Bigr) > 1. \end{aligned} \tag{0.12}$$

A very similar argument shows that if

$$\frac{\sum_{i=1}^n \alpha_i w_{ni}}{\sum_{i=1}^n \alpha_i w_{ni}^*} \to \sigma \in [0,1), \tag{0.13}$$

then the conclusion of (6.35) also holds. But if (2.5) holds and it is not the case that both $\sum_{i=1}^n w_{ni}^2 / \sum_{i=1}^n (w_{ni}^*)^2 \to 1$ and $\sum_{i=1}^n \alpha_i w_{ni} / \sum_{i=1}^n \alpha_i w_{ni}^* \to 1$, then either (6.34) or (6.36) would have to hold on a subsequence. But then we see from (6.35) that (2.5) cannot hold, and this contradiction means that the conditions (2.6) are necessary for (2.5).

The final part of the theorem, deriving (2.7), is an elementary calculation and is omitted. □

Proof of Corollary 4. Writing $a_{n,q} = \frac{1}{nq} + q^2$ and $b_{n,q} = \frac{1}{nq} + q$, this corollary follows from Theorem 1 and the following facts:

$$\sum_{i=1}^n \alpha_i w_{ni}^{\mathrm{b,with}} = \frac{\Gamma(2+\frac{2}{d})}{q^{2/d}}\{1+O(a_{n,q})\} \text{ and } \sum_{i=1}^n (w_{ni}^{\mathrm{b,with}})^2 = \frac{q}{2}\{1+O(a_{n,q})\}$$
$$\sum_{i=1}^n \alpha_i w_{ni}^{\mathrm{b,w/o}} = \frac{\Gamma(2+\frac{2}{d})}{q^{2/d}}\{1+O(b_{n,q})\} \text{ and } \sum_{i=1}^n (w_{ni}^{\mathrm{b,w/o}})^2 = \frac{q}{2}\{1+O(b_{n,q})\}$$
$$\sum_{i=1}^n \alpha_i w_{ni}^{\mathrm{Geo}} = \frac{\Gamma(2+2/d)}{q^{2/d}}\{1+O(q)\} \text{ and } \sum_{i=1}^n (w_{ni}^{\mathrm{Geo}})^2 = \frac{q}{2}\{1+O(q)\},$$

where the error terms are, in each case, uniform for $n^{-(1-\beta)} \leq q \leq n^{-\beta}$. □

Proof of Theorem 6. Let $t_n^{(r)} = n^{-2r/d} \sum_{i=1}^n \alpha_i^{(r)} w_{ni}$. We only need to show that

$$\sup_{x \in \mathcal{S}^{\epsilon_n}} \left| \mu_n(x) - \eta(x) - a^{(r)}(x) t_n^{(r)} \right| = o(t_n^{(r)}), \tag{0.14}$$

uniformly for $\mathbf{w}_n \in W_{n,\beta,r}^{\dagger}$, because the rest of the proof is virtually identical to that of Theorem 1. Analogously to (6.8), we can write

$$
\begin{aligned}
&\sum_{i=1}^{k_2} w_{ni}\biggl[\sum_{|s^1|\le 2r} \eta_{s^1}(x)\mathbb{E}\{(X_{(i)}-x)^{s^1}\} - \eta(x)\biggr] \\
&\qquad = \{1+o(1)\}\sum_{i=k_1}^{k_2} n\Delta w_{ni} \sum_{(s^1,s^2)\in\bar{S}_r} \frac{\eta_{s^1}(x)\bar{f}_{s^2}(x)}{|s^1|!|s^2|!}\int_{\|u\|\le\delta_n} q^{n-1}_{\|u\|}(i)\,du \\
&\qquad = \{1+o(1)\}\sum_{i=1}^{n} \frac{n\Delta w_{ni}}{b_n^{d+2r}} \sum_{(s^1,s^2)\in\bar{S}_r} \frac{\eta_{s^1}(x)\bar{f}_{s^2}(x)}{|s^1|!|s^2|!}\int_{\|v\|\le 1} v^{s^1+s^2}\,dv \\
&\qquad = a^{(r)}(x)t_n^{(r)} + o(t_n^{(r)}), \qquad (0.15)
\end{aligned}
$$

uniformly for $x \in \mathcal{S}^{\epsilon_n}$ and $\mathbf{w}_n \in W^{\dagger}_{n,\beta,r}$. Combining the analogues of (6.3) and (6.4) with (6.38) proves (6.37). □

Minimax properties of weighted nearest neighbour classifiers

The conditions imposed by Audibert and Tsybakov (2007) for their minimax results are closely related to the ones introduced in Section 2, and for convenience we will use their conditions in this section. For fixed $\alpha \ge 0$, fixed positive parameters C_0, γ, L, r_0, c_0, $\bar{f}_{\max} > \bar{f}_{\min} > 0$ and a fixed compact set $\mathcal{C} \subseteq \mathbb{R}^d$, let $\mathcal{P}_{\alpha,\gamma}$ denote the class of probability distributions P on $\mathbb{R}^d \times \{1,2\}$ such that:

(i) The margin condition is satisfied; that is,

$$\bar{P}(\{x \in \mathbb{R}^d : 0 < |\eta(x) - 1/2| \le \epsilon\}) \le C_0\epsilon^{\alpha}$$

for all $\epsilon > 0$, where $\bar{P}$ denotes the marginal distribution of X.

(ii) The regression function η belongs to the Hölder ball $\Sigma(\gamma, L, \mathbb{R}^d)$; that is, η is $\lfloor\gamma\rfloor$ times continuously differentiable and, writing $\eta_x^{\lfloor\gamma\rfloor}$ for the Taylor series polynomial of η of order $\lfloor\gamma\rfloor$, we have that for all $x, x' \in \mathbb{R}^d$,

$$|\eta(x') - \eta_x^{\lfloor\gamma\rfloor}(x')| \le L\|x - x'\|^{\gamma}.$$

(iii) The marginal distribution $\bar{P}$ is supported on a set $A \subseteq \mathcal{C}$ satisfying

$$\nu_d(A \cap B_r(x)) \ge c_0\nu_d(B_r(x))$$

for all $r \in [0, r_0]$ and $x \in A$; moreover, $\bar{P}$ has a Lebesgue density $\bar{f}$ satisfying $\bar{f}_{\min} \le \bar{f}(x) \le \bar{f}_{\max}$ for all $x \in A$. Finally, $\bar{f} \in \Sigma(\gamma - 1, L, A)$.

In Theorem 1 below, we let $\hat{C}_n^{\mathrm{wnn}}$ be denote either a unweighted k-nearest neighbour classifier with $k \asymp n^{2\gamma/(2\gamma+d)}$, or a weighted nearest neighbour classifier with weights of the form (1.1) with $k^* \asymp n^{2\gamma/(2\gamma+d)}$, or any of the three types of bagged nearest neighbour classifier from Section 3 with $q \asymp n^{-2\gamma/(2\gamma+d)}$.

THEOREM 1. *For any $\alpha \geq 0$ and $\gamma \in (0, 2]$, there exists $C > 0$ such that*

$$\sup_{P \in \mathcal{P}_{\alpha,\gamma}} \{\mathcal{R}(\hat{C}_n^{\mathrm{wnn}}) - \mathcal{R}(C^{\mathrm{Bayes}})\} \leq C n^{-\gamma(1+\alpha)/(2\gamma+d)}.$$

Remark: Using the ideas of Section 4, one can extend this result so that, given any $\alpha \geq 0$ and $\gamma > 0$, we can construct a weighted nearest neighbour classifier achieving the rate $n^{-\gamma(1+\alpha)/(2\gamma+d)}$ uniformly over $\mathcal{P}_{\alpha,\gamma}$.

PROOF. Write P^n for the joint distribution of $(X_1, Y_1), \ldots, (X_n, Y_n)$. By Lemma 3.1 of Audibert and Tsybakov (2007), it suffices to prove that there exists $C > 0$ such that for all $\delta > 0$, all $n \in \mathbb{N}$ and $\bar{P}$-almost all x,

$$\sup_{P \in \mathcal{P}_{\alpha,\gamma}} P^n(|S_n(x) - \eta(x)| > \delta) \leq C \exp(-n^{2\gamma/(2\gamma+\delta)}\delta^2/C). \tag{0.16}$$

Now, minor variants of the arguments in the proof of Theorem **??** show that there exist $C_1', C' > 0$ such that for all $P \in \mathcal{P}_{\alpha,\gamma}$ and $x \in A$,

$$\begin{aligned}|\mu_n(x) - \eta(x)| &\leq L \sum_{i=1}^n w_{ni} \mathbb{E}\{\|X_{(i)} - x\|^\gamma\} + \biggl|\sum_{i=1}^n w_{ni} \mathbb{E}\eta_x^{\lfloor\gamma\rfloor}(X_{(i)}) - \eta(x)\biggr| \\ &\leq C_1' \sum_{i=1}^n \Bigl(\frac{i}{n}\Bigr)^{\gamma/d} w_{ni} \leq C' n^{-\gamma/(2\gamma+d)}.\end{aligned}$$

Moreover, there exists $C'' > 0$ such that we have

$$\sum_{i=1}^n w_{ni}^2 \leq C'' n^{-2\gamma/(2\gamma+d)}.$$

It follows that for $\delta \geq 2C' n^{-\gamma/(2\gamma+d)}$ and for $\bar{P}$-almost all x,

$$\begin{aligned}\sup_{P \in \mathcal{P}_{\alpha,\gamma}} P^n(|S_n(x) - \eta(x)| > \delta) &\leq \sup_{P \in \mathcal{P}_{\alpha,\gamma}} P^n(|S_n(x) - \mu_n(x)| > \delta/2) \\ &\leq 2\exp\{-n^{2\gamma/(2\gamma+d)}\delta^2/(2C'')\},\end{aligned}$$

by Hoeffding's inequality. Therefore, if we choose $C \geq \max(2, 2C'')$ such that $C \exp\{-4(C')^2/C\} \geq 1$, then (0.16) holds for all $\delta > 0$, as required. □

For the lower bound, it is convenient to let $\bar{\mathcal{P}}_{\alpha,\gamma}$ denote the set of probability distributions on $\mathbb{R}^d \times \{1,2\}$ satisfying all the restrictions of $\mathcal{P}_{\alpha,\gamma}$ except for the condition that $\bar{f} \in \Sigma(\gamma-1, L, A)$. The following lower bound is Theorem 3.5 of Audibert and Tsybakov (2007).

Theorem 2 (Audibert and Tsybakov (2007)). *Suppose that $\alpha\gamma \leq \delta$. Then there exists $c > 0$ such that for any classifier $\hat{C}_n$, we have*

$$\sup_{P \in \bar{\mathcal{P}}_{\alpha,\gamma}} \{\mathcal{R}(\hat{C}_n) - \mathcal{R}(C^{\mathrm{Bayes}})\} \geq cn^{-\gamma(1+\alpha)/(2\gamma+d)}.$$

Remark: The condition $\alpha\gamma \leq \delta$ is not too restrictive; in particular, it allows all (α, γ, d) triples where the rate of convergence is not faster than n^{-1}.

Remark: In the proof of Theorem 3.5 of Audibert and Tsybakov (2007), the authors construct a finite subset of $\bar{\mathcal{P}}_{\alpha,\gamma}$ for which the lower bound holds. Moreover, all of the distributions in this finite subset have the same marginal density $\bar{f}$. This density is piecewise constant, so does not satisfy $\bar{f} \in \Sigma(\gamma-1, L, A)$, but it can be modified in a straightforward but cumbersome way to do so, obtaining the same lower bound while preserving the margin condition **(i)** and the other conditions in **(iii)** above. Thus the lower bound in fact holds over $\mathcal{P}_{\alpha,\gamma}$.

A plug-in approach to estimating k^*

A direct, plug-in approach to estimating the constants B_1 and B_2 in (2.4) involves estimating integrals over the manifold $\mathcal{S}$, which could be achieved using Monte Carlo integration. For instance, one might first try to assess which points x are close to $\mathcal{S}$ by testing $H_0 : x \in \mathcal{S}$ against $H_1 : x \notin \mathcal{S}$. This could be done by constructing a pilot weighted nearest neighbour estimate $\tilde{S}_n(x)$ of $\eta(x)$ and using the test statistic

$$T_n(x) = \frac{(\tilde{S}_n(x) - 1/2)^2}{\tilde{S}_n(x)\{1 - \tilde{S}_n(x)\}};$$

cf. Samworth (2011). From the expressions in (2.3), we see that one would then require estimates $\|\hat{\dot{\eta}}(\cdot)\|$ and $\hat{a}(\cdot)$ of $\|\dot{\eta}(\cdot)\|$ and $a(\cdot)$ respectively at the training points for which we do not reject the null hypothesis. The former estimates could be achieved using finite-difference approximations to the partial derivatives; for the latter estimates, while it is in principle be possible to construct further plug-in estimates of the unknown quantities in the expression for $a(\cdot)$, this might be considered unattractive owing to the need to estimate second partial derivatives. Instead, one could also base an estimate on a jackknife estimate of the bias $\mathbb{E}\{S_n(\cdot)\} - \eta(\cdot)$; cf. the discussion preceding (2.4). Finally, one would then estimate B_1 and B_2 by

$$\hat{B}_1 = \frac{1}{|\hat{S}|} \sum_{i : X_i \in \hat{S}} \frac{1}{\|\hat{\dot{\eta}}(X_i)\|} \quad \text{and} \quad \hat{B}_2 = \frac{1}{|\hat{S}|} \sum_{i : X_i \in \hat{S}} \frac{\hat{a}(X_i)^2}{\|\hat{\dot{\eta}}(X_i)\|},$$

where $\hat{S}$ is the set of training data points for which we do not reject the null hypothesis above. The potential advantages of this approach are both computational speed (we avoid the cross-validation step) and theoretical, especially if one could prove that $\hat{B}_1 \overset{p}{\to} B_1$ and $\hat{B}_2 \overset{p}{\to} B_2$. This would guarantee that the estimate of k^* was of the appropriate asymptotic order in n, and that the asymptotic results corresponding to (2.7) and (2.8) continued to hold using the plug-in estimates of the weights. Such a result would rely on the fact that the convergence of the scaled regrets to their limits is uniform in a range of k, as discussed after (2.10).

However, the plug-in approach does appear to have some drawbacks in this context. There are several choices to make, including the pilot estimate, the critical value for the test statistic, and the step size for the finite-difference approximation (which is not straightforward since the estimates of $\eta(\cdot)$ are piecewise constant, so if it is chosen too small, then the partial derivatives will be estimated as zero). Moreover, there may be very few points in $\hat{S}$, leading to highly variable estimates of the unknown quantities, and there is no guarantee that for a given sample size, the estimate of k^* will necessarily be less than n. It is for these reasons that we prefer the cross-validation approach of Section 5 in this instance.

References.

Audibert, J.-Y. and Tsybakov, A. B. (2007) Fast learning rates for plug-in classifiers *Ann. Statist.*, **35**, 608–633.

Folland, G. B. (1999) *Real Analysis.* Wiley, New York (2nd ed.).

Gray, A. (2004) *Tubes* Birkhäuser Verlag, Basel, Switzerland (2nd ed.).

Samworth, R. J. (2011) Comment on *Adaptive Confidence Intervals for the Test Error in Classification* by E. B. Laber and S. A. Murphy. *J. Amer. Statist. Assoc.*, **106**, 914–915.

Samworth, R. J. (2012) Optimal weighted nearest neighbour classifiers. *Ann. Statist.*, to appear.

Shorack, G. A. and Wellner, J. A. (1986) *Empirical Processes with Applications to Statistics*, Wiley, New York.

Statistical Laboratory
Wilberforce Road
Cambridge
CB3 0WB
United Kingdom,
E-mail: r.samworth@statslab.cam.ac.uk