The Annals of Statistics
2012, Vol. 40, No. 5, 2733–2763
DOI: 10.1214/12-AOS1049

OPTIMAL WEIGHTED NEAREST NEIGHBOUR CLASSIFIERS[1]

BY RICHARD J. SAMWORTH

University of Cambridge

We derive an asymptotic expansion for the excess risk (regret) of a weighted nearest-neighbour classifier. This allows us to find the asymptotically optimal vector of nonnegative weights, which has a rather simple form. We show that the ratio of the regret of this classifier to that of an unweighted k-nearest neighbour classifier depends asymptotically only on the dimension d of the feature vectors, and not on the underlying populations. The improvement is greatest when $d = 4$, but thereafter decreases as $d \to \infty$. The popular bagged nearest neighbour classifier can also be regarded as a weighted nearest neighbour classifier, and we show that its corresponding weights are somewhat suboptimal when d is small (in particular, worse than those of the unweighted k-nearest neighbour classifier when $d = 1$), but are close to optimal when d is large. Finally, we argue that improvements in the rate of convergence are possible under stronger smoothness assumptions, provided we allow negative weights. Our findings are supported by an empirical performance comparison on both simulated and real data sets.

1. Introduction.

Supervised classification, also known as pattern recognition, is a fundamental problem in Statistics, as it represents an abstraction of the decision-making problem faced by many applied practitioners. Examples include a doctor making a medical diagnosis, a handwriting expert performing an authorship analysis, or an email filter deciding whether or not a message is genuine.

Classifiers based on nearest neighbours are perhaps the simplest and most intuitively appealing of all nonparametric classifiers. The k-nearest neighbour classifier was originally studied in the seminal works of Fix and Hodges (1951) [later republished as Fix and Hodges (1989)] and Cover and Hart (1967), but it retains its popularity today. Surprisingly, it is only recently that detailed understanding of the nature of the error probabilities has emerged [Hall, Park and Samworth (2008)].

Arguably the most obvious defect with the k-nearest neighbour classifier is that it places equal weight on the class labels of each of the k nearest neighbours to the point x being classified. Intuitively, one would expect improvements in terms of the misclassification rate to be possible by putting decreasing weights on the class labels of the successively more distant neighbours.

Received June 2012; revised August 2012.

[1]Supported by the Leverhulme Research Fellowship and an EPSRC Early Career Fellowship.

MSC2010 subject classifications. 62G20.

Key words and phrases. Bagging, classification, nearest neighbours, weighted nearest neighbour classifiers.

The first purpose of this paper is to describe the asymptotic structure of the difference between the misclassification rate (risk) of a weighted nearest neighbour classifier and that of the optimal Bayes classifier for classification problems with feature vectors in $\mathbb{R}^d$. Theorem 1 in Section 2 below shows that, subject to certain regularity conditions on the underlying distributions of each class and the weights, this excess risk (or *regret*) asymptotically decomposes as a sum of two dominant terms, one representing bias and the other representing variance. For simplicity of exposition, we will deal initially with binary classification problems, though we also indicate the appropriate extension to general multicategory problems.

Our second contribution, following on from the first, is to derive the vector of nonnegative weights that is asymptotically optimal in the sense of minimising the misclassification rate; cf. Theorem 2. In fact this asymptotically optimal weight vector has a relatively simple form: let n denote the sample size and let w_{ni} denote the weight assigned to the ith nearest neighbour (normalised so that $\sum_{i=1}^n w_{ni} = 1$). Then the optimal choice is to set $k^* = \lfloor B^* n^{4/(d+4)} \rfloor$ [an explicit expression for B^* is given in (2.4) below] and then let

$$w_{ni}^* = \begin{cases} \dfrac{1}{k^*}\Big[1 + \dfrac{d}{2} - \dfrac{d}{2(k^*)^{2/d}}\{i^{1+2/d} - (i-1)^{1+2/d}\}\Big], \\ \qquad \text{for } i = 1, \dots, k^*, \\ 0, \qquad \text{for } i = k^* + 1, \dots, n. \end{cases} \tag{1.1}$$

Thus, in the asymptotically optimal weighting scheme, only a proportion $O(n^{-d/(d+4)})$ of the weights are positive. The maximal weight is almost $(1 + d/2)$ times the average positive weight, and the discrete distribution on $\{1, \dots, n\}$ defined by the asymptotically optimal weights decreases in a concave fashion when $d = 1$, in a linear fashion when $d = 2$ and in a convex fashion when $d \geq 3$; see Figure 1. When d is large, about $1/e$ of the weights are above the average positive weight.

Another consequence of Theorem 2 is that k^* is bigger by a factor of $\{\frac{2(d+4)}{d+2}\}^{d/(d+4)}$ than the asymptotically optimal choice of k for traditional, unweighted k-nearest neighbour classification. It is notable that this factor, which is around 1.27 when $d = 1$ and increases towards 2 for large d, does not depend on the underlying populations. This means that there is a natural correspondence between any unweighted k-nearest neighbour classifier and one of optimally weighted form, obtained by multiplying k by this dimension-dependent factor to obtain the number k' of positive weights for the weighted classifier, and then using the weights given in (1.1) with k' replacing k^*.

In Corollary 3 we describe the asymptotic improvement in the excess risk that is attainable using the procedure described in the previous paragraph. Since the rate of convergence to zero of the excess risk is $O(n^{-4/(d+4)})$ in both cases, the improvement is in the leading constant, and again it is notable that the asymptotic improvement does not depend on the underlying populations. The improvement

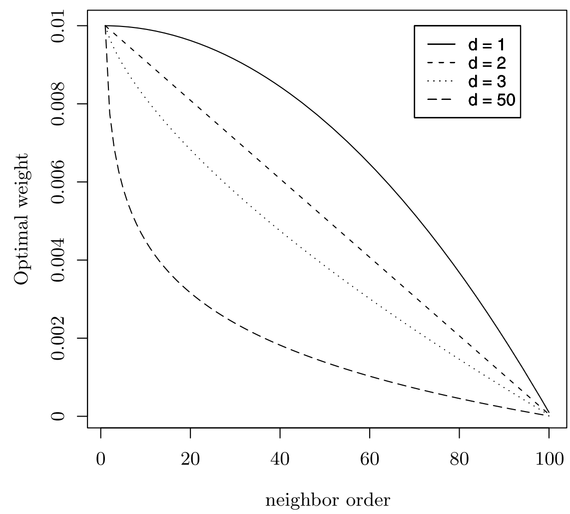

FIG. 1. *Optimal weight profiles at different dimensions. Here,* $k^* = 100$, *and the figure displays the positive weights in* (1.1), *scaled to have the same weight on the nearest neighbour at each dimension.*

is relatively modest, which goes some way to explaining the continued popularity of the (unweighted) k-nearest neighbour classifier. Nevertheless, for $d \leq 15$, the improvement in regret is at least 5%, though it is negligible as $d \to \infty$; the greatest improvement occurs when $d = 4$, and here it is just over 8%. See Figure 2.

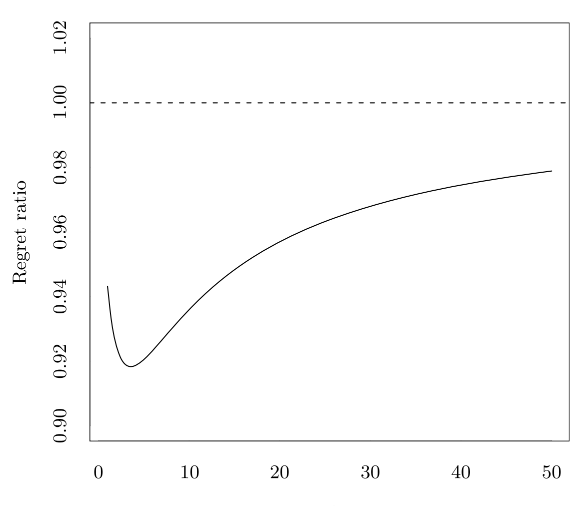

FIG. 2. *Asymptotic ratio of the regret of the optimally weighted nearest neighbour classifier to that of the optimal k-nearest neighbour classifier, as a function of the dimension d of the feature vectors.*

Another popular way of improving the performance of a classifier is by *bagging* [Breiman (1996, 1999)]. Short for "bootstrap aggregating", bagging involves combining the results of many empirically simulated predictions. Empirical analyses [e.g., Steele (2009)], have reported that bagging can result in improvements over unweighted k-nearest neighbour classification. Moreover, as explained by Biau, Cérou and Guyader (2010), understanding the properties of the bagged nearest neighbour classifier is also of interest because they provide insight into *random forests* [Breiman (2001)]. Random forest algorithms have been some of the most successful ensemble methods for regression and classification problems, but their theoretical properties remain relatively poorly understood. When bagging the nearest neighbour classifier, we can draw resamples from the data either with- or without-replacement. We treat the "infinite simulation" case, where both versions take the form of a weighted nearest neighbour classifier with weights decaying approximately exponentially on successively more distant observations from the point being classified [Hall and Samworth (2005), Biau, Cérou and Guyader (2010)]. The crucial choice is that of the resample size, or equivalently the sampling fraction, that is, the ratio of the resample size to the original sample size. In Section 3, we describe the asymptotically optimal resample fraction (showing in particular that it is the same for both with- and without-replacement sampling) and compare its regret with those of the weighted and unweighted k-nearest neighbour classifiers.

In Section 4, we consider the problem of choosing optimal weights without the restriction that they should be nonnegative. The situation here is somewhat analogous to the use of higher order kernels for classifiers based on kernel density estimates of each of the population densities. In particular, subject to additional smoothness assumptions on the population densities, we find that powers of n arbitrarily close to the "parametric rate" of $O(n^{-1})$ for the excess risk are attainable. Section 5 presents the results of an empirical performance comparison of different classifiers studied in the paper, and shows that the asymptotic theory predicts the empirical performance well. The main steps in the proof of Theorem 1 are given in the Appendix; the remaining details can be found in the supplementary material [Samworth (2012)], along with the other proofs and some ancillary material.

Classification has been the subject of several book-length treatments, including Hand (1981), Devroye, Györfi and Lugosi (1996) and Gordon (1999). In particular, classifiers based on nearest neighbours form a central theme of Devroye, Györfi and Lugosi (1996). The review paper by Boucheron, Bousquet and Lugosi (2005) contains 243 references and provides a thorough survey of the classification literature up to 2005. More recently, Audibert and Tsybakov (2007) have discussed the relative merits of *plug-in* classifiers (a family to which weighted nearest neighbour classifiers belong) and classifiers based on *empirical risk minimisation*, such as support vector machines [Cortes and Vapnik (1995), Blanchard, Bousquet and Massart (2008), Steinwart and Christmann (2008)].

Weighted nearest neighbour classifiers were first studied by Royall (1966); see also Bailey and Jain (1978). Stone (1977) proved that if $\max_{1 \le i \le n} w_{ni} \to 0$ as $n \to \infty$ and $\sum_{i=1}^{k} w_{ni} \to 1$ for some $k = k_n$ with $k/n \to 0$ as $n \to \infty$, then risk of the weighted nearest neighbour classifier converges to the risk of the Bayes classifier; see also Devroye, Györfi and Lugosi (1996), page 179. As mentioned above, this work attempts to study the difference between these risks more closely. Weighted nearest neighbour classifiers are also related to classifiers based on kernel estimates of each of the class densities; see, for example, the review by Raudys and Young (2004), as well as Hall and Kang (2005). The $O(n^{-4/(d+4)})$ rates of convergence obtained in this paper for nonnegative weights are the same as those obtained by Hall and Kang (2005) under similar twice-differentiable conditions with second-order kernel estimators of the class densities. Further related work includes the literature on *highest density region* or *level set* estimation [Polonik (1995), Rigollet and Vert (2009), Samworth and Wand (2010)].

Hall and Samworth (2005) and Biau and Devroye (2010) proved an analogous result for the bagged nearest neighbour classifier to the Stone (1977) result described in the previous paragraph. More precisely, if the resample size $m = m_n$ used for the bagging diverges to infinity, and $m/n \to 0$ as $n \to \infty$, then the risk of the bagged nearest neighbour classifier converges to the Bayes risk. Note that this result does not depend on whether the resamples are taken with or without replacement from the training data. Biau, Cérou and Guyader (2010) have recently proved a striking rate of convergence result for the bagged nearest neighbour estimate; this is described in greater detail in Section 3.

2. Main results. Let $(X, Y), (X_1, Y_1), (X_2, Y_2), \ldots$ be independent and identically distributed pairs taking values in $\mathbb{R}^d \times \{1, 2\}$. We suppose that $\mathbb{P}(Y = 1) = \pi = 1 - \mathbb{P}(Y = 2)$ for some $\pi \in (0, 1)$ and that $(X|Y = r) \sim P_r$ for $r = 1, 2$, where P_r is a probability measure on $\mathbb{R}^d$. We write $\bar{P} = \pi P_1 + (1 - \pi)P_2$ for the marginal distribution of X and let $\eta(x) = \mathbb{P}(Y = 1|X = x)$ denote the corresponding regression function.

A *classifier* C is a Borel measurable function from $\mathbb{R}^d$ to $\{1, 2\}$, with the interpretation that the point $x \in \mathbb{R}^d$ is classified as belonging to class $C(x)$. The *misclassification rate*, or *risk* of C over a Borel measurable set $\mathcal{R} \subseteq \mathbb{R}^d$ is defined to be

$$R_{\mathcal{R}}(C) = \mathbb{P}[\{C(X) \neq Y\} \cap \{X \in \mathcal{R}\}].$$

We also write $R(C)$ for this quantity when $\mathcal{R} = \mathbb{R}^d$. The classifier which minimises the risk over $\mathcal{R}$ is the Bayes classifier, given by

$$C^{\text{Bayes}}(x) = \begin{cases} 1, & \text{if } \eta(x) \ge 1/2, \\ 2, & \text{otherwise.} \end{cases}$$

Its risk is

$$R_{\mathcal{R}}(C^{\text{Bayes}}) = \int_{\mathcal{R}} \min\{\eta(x), 1 - \eta(x)\}\, d\bar{P}(x).$$

For each $n \in \mathbb{N}$, let $\mathbf{w}_n = (w_{ni})_{i=1}^n$ denote a vector of weights, normalised so that $\sum_{i=1}^n w_{ni} = 1$. Fix $x \in \mathcal{R}$ and an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$, and let $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ denote a permutation of the training sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ such that $\|X_{(1)} - x\| \leq \cdots \leq \|X_{(n)} - x\|$. We define the *weighted nearest neighbour classifier* to be

$$\hat{C}_n^{\mathrm{wnn}}(x) = \begin{cases} 1, & \text{if } \sum_{i=1}^n w_{ni} \mathbb{1}_{\{Y_{(i)}=1\}} \geq 1/2, \\ 2, & \text{otherwise.} \end{cases}$$

We also write $\hat{C}_{n,\mathbf{w}_n}^{\mathrm{wnn}}$ where it is necessary to emphasise the weight vector, for example, when comparing different weighted nearest neighbour classifiers. Our initial goal is to study the asymptotic behaviour of

$$R_{\mathcal{R}}(\hat{C}_n^{\mathrm{wnn}}) = \mathbb{P}[\{\hat{C}_n^{\mathrm{wnn}}(X) \neq Y\} \mathbb{1}_{\{X \in \mathcal{R}\}}].$$

It will be convenient to define some notation: for a smooth function $g : \mathbb{R}^d \to \mathbb{R}$, we write $\dot{g}(x)$ for its gradient vector at x, and $g_j(x)$ for its jth partial derivative at x. Analogously, we write $\ddot{g}(x)$ for the Hessian matrix of g at x, and $g_{jk}(x)$ for its (j,k)th element. We let $B_\delta(x) = \{y \in \mathbb{R}^d : \|y - x\| \leq \delta\}$ denote the closed ball of radius δ centered at x in the norm $\|\cdot\|$, and let a_d denote the d-dimensional Lebesgue measure of the unit ball $B_1(x)$. Thus, $a_d = 2^d \Gamma(1 + 1/p)^d / \Gamma(1 + d/p)$ when $\|\cdot\|$ is the ℓ_p-norm. We will make use of the following assumptions for our theoretical results:

(A.1) The set $\mathcal{R} \subseteq \mathbb{R}^d$ is a compact d-dimensional manifold with boundary $\partial\mathcal{R}$.

(A.2) The set $\mathcal{S} = \{x \in \mathcal{R} : \eta(x) = 1/2\}$ is nonempty. There exists an open subset U_0 of $\mathbb{R}^d$ that contains $\mathcal{S}$ and such that the following properties hold: first, η is continuous on $U \setminus U_0$, where U is an open set containing $\mathcal{R}$; second, the restrictions of P_1 and P_2 to U_0 are absolutely continuous with respect to Lebesgue measure, with twice continuously differentiable Radon–Nikodym derivatives f_1 and f_2, respectively.

(A.3) There exists $\rho > 0$ such that $\int_{\mathbb{R}^d} \|x\|^\rho \, d\bar{P}(x) < \infty$. Moreover, for sufficiently small $\delta > 0$, the ratio $\bar{P}(B_\delta(x))/(a_d \delta^d)$ is bounded away from zero, uniformly for $x \in \mathcal{R}$.

(A.4) For all $x \in \mathcal{S}$, we have $\dot{\eta}(x) \neq 0$, and for all $x \in \mathcal{S} \cap \partial\mathcal{R}$, we have $\dot{\partial\eta}(x) \neq 0$, where $\partial\eta$ denotes the restriction of η to $\partial\mathcal{R}$.

The introduction of the compact set $\mathcal{R}$ finesses the problem of performing classification in the tails of the feature vector distributions. See, for example, Hall and Kang (2005), Section 3, for further discussion of this point and related results, as well as Chanda and Ruymgaart (1989). Mammen and Tsybakov (1999) and Audibert and Tsybakov (2007) impose similar compactness assumptions for their results. The set $\mathcal{R}$ may be arbitrarily large, though the larger it is, the stronger are the requirements in (A.2). Although as stated, the assumptions on $\mathcal{R}$ are quite general, little is lost by thinking of $\mathcal{R}$ as a large closed Euclidean ball. Its role in the asymptotic expansion of Theorem 2 below is that it is involved in the definition of the set $\mathcal{S}$, which represents the decision boundary of the Bayes classifier. We will see that the behaviour of f_1 and f_2 on the set $\mathcal{S}$ is crucial for determining the asymptotic behaviour of weighted nearest neighbour classifiers.

The second part of (A.3) asks that the ratio of the $\bar{P}$-measure of small balls to the corresponding d-dimensional Lebesgue measure is bounded away from zero. This requirement is satisfied, for instance, if P_1 and P_2 are absolutely continuous with respect to Lebesgue measure, with Radon–Nikodym derivatives that are bounded away from zero on the open set U.

The assumption in (A.4) that $\dot{\eta}(x) \neq 0$ for $x \in \mathcal{S}$ asks that f_1 and f_2, weighted by the respective prior probabilities of each class, should cut at a nonzero angle along $\mathcal{S}$. In the language of differential topology, this means that $1/2$ is a *regular value* of the function η, and the second part of (A.4) asks for $1/2$ to be a regular value of the restriction of η to $\partial\mathcal{R}$. Together, these two requirements ensure that $\mathcal{S}$ is a $(d-1)$-dimensional submanifold with boundary of $\mathbb{R}^d$, and the boundary of $\mathcal{S}$ is $\{x \in \partial\mathcal{R} : \eta(x) = 1/2\}$ [Guillemin and Pollack (1974), page 60].

The requirement in (A.4) that $\dot{\eta}(x) \neq 0$ for $x \in \mathcal{S}$ is related to the well-known *margin condition* of, for example, Mammen and Tsybakov (1999) and Tsybakov (2004); when it holds (and in the presence of the other conditions), there exist $c, C > 0$ such that

$$c\varepsilon \leq \mathbb{P}\bigl(|\eta(X) - 1/2| \leq \varepsilon \cap X \in \mathcal{R}\bigr) \leq C\varepsilon \tag{2.1}$$

for sufficiently small $\varepsilon > 0$; see Tsybakov (2004), Proposition 1. A proof of this fact, which uses Weyl's tube formula [Gray (2004)], is given after the completion of the proof of Theorem 1 in the supplementary material [Samworth (2012)]. In this sense, we work in the setting of a margin condition with the power parameter equal to 1.

We now introduce some notation needed for Theorem 1 below. For $\beta > 0$, let $W_{n,\beta}$ denote the set of all sequences of nonnegative deterministic weight vectors $\mathbf{w}_n = (w_{ni})_{i=1}^n$ satisfying:

- $\sum_{i=1}^n w_{ni}^2 \leq n^{-\beta}$;
- $n^{-4/d}(\sum_{i=1}^n \alpha_i w_{ni})^2 \leq n^{-\beta}$, where $\alpha_i = i^{1+2/d} - (i-1)^{1+2/d}$; note that this latter expression appears in (1.1);
- $n^{2/d} \sum_{i=k_2+1}^n w_{ni} / \sum_{i=1}^n \alpha_i w_{ni} \leq 1/\log n$, where $k_2 = \lfloor n^{1-\beta} \rfloor$;
- $\sum_{i=k_2+1}^n w_{ni}^2 / \sum_{i=1}^n w_{ni}^2 \leq 1/\log n$;
- $\sum_{i=1}^n w_{ni}^3 / (\sum_{i=1}^n w_{ni}^2)^{3/2} \leq 1/\log n$.

Observe that $W_{n,\beta_1} \supset W_{n,\beta_2}$ for $\beta_1 < \beta_2$. The first and last conditions ensure that the weights are not too concentrated on a small number of points; the second amounts to a mild moment condition on the probability distribution on $\{1, \ldots, n\}$

defined by the weights. The next two conditions ensure that not too much weight (or squared weight in the case of the latter condition) is assigned to observations that are too far from the point being classified. Although there are many requirements on the weight vectors, they are rather mild conditions when β is small, as can be seen by considering the limiting case $\beta = 0$. For instance, for the unweighted k-nearest neighbour classifier with weights $\mathbf{w}_n = (w_{ni})_{i=1}^n$ given by $w_{ni} = k^{-1}\mathbb{1}_{\{1 \le i \le k\}}$, we have that $\mathbf{w}_n \in W_{n,\beta}$ for small $\beta > 0$ provided that $\max(n^\beta, \log^2 n) \le k \le \min(n^{(1-\beta d/4)}, n^{1-\beta})$. Thus for the vector of k-nearest neighbour weights to belong to $W_{n,\beta}$ for all large n, it is necessary that the usual conditions $k \to \infty$ and $k/n \to 0$ for consistency are satisfied, and these conditions are almost sufficient when $\beta > 0$ is small. The situation is similar for the bagged nearest neighbour classifier—see Section 3 below.

The fact that the weights are assumed to be deterministic means that they depend only on the ordering of the distances, not the raw distances themselves (as would be the case for a classifier based on kernel density estimates of the population densities). Such kernel-based classifiers are not necessarily straightforward to implement, however: Hall and Kang (2005) showed that even in the simple situation where $d = 1$ and πf_1 and $(1-\pi)f_2$ cross at a single point x_0, the optimal order of the bandwidth for the kernel depends on the sign of $\ddot{f}_1(x_0)\ddot{f}_2(x_0)$.

Continuing with our notational definitions, let $\bar{f} = \pi f_1 + (1-\pi)f_2$. Define

$$a(x) = \frac{\sum_{j=1}^d c_{j,d}\{\eta_j(x)\bar{f}_j(x) + (1/2)\eta_{jj}(x)\bar{f}(x)\}}{a_d^{1+2/d}\bar{f}(x)^{1+2/d}}, \tag{2.2}$$

where $c_{j,d} = \int_{v:\|v\|\le 1} v_j^2\, dv$. Finally, let

$$\begin{aligned} B_1 &= \int_{\mathcal{S}} \frac{\bar{f}(x_0)}{4\|\dot{\eta}(x_0)\|}\, d\mathrm{Vol}^{d-1}(x_0) \quad \text{and} \\ B_2 &= \int_{\mathcal{S}} \frac{\bar{f}(x_0)}{\|\dot{\eta}(x_0)\|} a(x_0)^2\, d\mathrm{Vol}^{d-1}(x_0), \end{aligned} \tag{2.3}$$

where Vol^{d-1} denotes the natural $(d-1)$-dimensional volume measure that $\mathcal{S}$ inherits as a subset of $\mathbb{R}^d$. Note that $B_1 > 0$, and $B_2 \ge 0$, with equality if and only if a is identically zero on $\mathcal{S}$. Although the definitions of B_1 and B_2 are complicated, we will see after the statement of Theorem 1 below that they are comprised of terms that have natural interpretations.

THEOREM 1. *Assume* (A.1), (A.2), (A.3) *and* (A.4). *Then for each* $\beta \in (0, 1/2)$,

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

as $n \to \infty$, uniformly for $\mathbf{w}_n \in W_{n,\beta}$, where

$$\gamma_n(\mathbf{w}_n) = B_1 \sum_{i=1}^n w_{ni}^2 + B_2 \Biggl(\sum_{i=1}^n \frac{\alpha_i w_{ni}}{n^{2/d}} \Biggr)^2.$$

Theorem 1 tells us that, asymptotically, the dominant contribution to the regret over $\mathcal{R}$ of the weighted nearest neighbour classifier can be decomposed as a sum of two terms. The two terms, constant multiples of $\sum_{i=1}^n w_{ni}^2$ and $(\sum_{i=1}^n \frac{\alpha_i w_{ni}}{n^{2/d}})^2$, respectively, represent variance and squared bias contributions to the regret. It is interesting to observe that, although the 0–1 classification loss function is quite different from the squared error loss often used in regression problems, we nevertheless obtain such an asymptotic decomposition.

The constant multiples of the dominant variance and squared bias terms depend only on the behaviour of f_1 and f_2 (and their first and second derivatives) on $\mathcal{S}$, as seen from (2.3). Moreover, we can see from the expression for B_1 in (2.3) that the contribution to the dominant variance term in the regret will tend to be large in the following three situations: first, when $\bar{f}(\cdot)$ is large on $\mathcal{S}$; second, when the Vol^{d-1} measure of $\mathcal{S}$ is large; and third, when $\|\dot{\eta}(\cdot)\|$ is small on $\mathcal{S}$. In the first two of these situations, the probability is relatively high that a point to be classified will be close to the Bayes decision boundary $\mathcal{S}$, where classification is difficult. In the latter case, the regression function η moves away from $1/2$ only slowly as we move away from $\mathcal{S}$, meaning that there is a relatively large region of points near $\mathcal{S}$ where classification is difficult. From the expression for B_2 in (2.3), we see that the dominant squared bias term is also large in these situations, and also when $a(\cdot)^2$ is large on $\mathcal{S}$. From the proof of Theorem 1, it is apparent that $a(x) \sum_{i=1}^n \frac{\alpha_i w_{ni}}{n^{2/d}}$ is the dominant bias term for $S_n(x) = \sum_{i=1}^n w_{ni} \mathbb{1}_{\{Y_{(i)}=1\}}$ as an estimator of $\eta(x)$. Indeed, by a Taylor expansion,

$$\begin{aligned}
&\mathbb{E}\{S_n(x)\} - \eta(x) \\
&\quad = \sum_{i=1}^n w_{ni} \mathbb{E}\eta(X_{(i)}) - \eta(x) \\
&\quad \approx \sum_{i=1}^n w_{ni} \mathbb{E}\{(X_{(i)} - x)^T \dot{\eta}(x)\} + \frac{1}{2} \sum_{i=1}^n w_{ni} \mathbb{E}\{(X_{(i)} - x)^T \ddot{\eta}(x)(X_{(i)} - x)\}.
\end{aligned}$$

The two summands in the definition of $a(x)$ represent asymptotic approximations to the respective summands in this approximation.

Consider now the problem of optimising the choice of weight vectors. Let

$$k^* = \Biggl\lfloor \biggl\{ \frac{d(d+4)}{2(d+2)} \biggr\}^{d/(d+4)} \biggl(\frac{B_1}{B_2} \biggr)^{d/(d+4)} n^{4/(d+4)} \Biggr\rfloor, \tag{2.4}$$

and then define the weights $\mathbf{w}_n^* = (w_{ni}^*)_{i=1}^n$ as in (1.1). The first part of Theorem 2 below can be regarded as saying that the weights $\mathbf{w}_n^*$ are asymptotically optimal.

THEOREM 2. *Assume* (A.1)–(A.4), *and assume also that* $B_2 > 0$. *For any* $\beta > 0$ *and any sequence* $\mathbf{w}_n = (w_{ni})_{i=1}^n \in W_{n,\beta}$, *we have*

$$\liminf_{n\to\infty} \frac{R_{\mathcal{R}}(\hat{C}^{\mathrm{wnn}}_{n,\mathbf{w}_n}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})}{R_{\mathcal{R}}(\hat{C}^{\mathrm{wnn}}_{n,\mathbf{w}_n^*}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})} \geq 1. \tag{2.5}$$

Moreover, the ratio in (2.5) *above converges to* 1 *if and only if we have both* $\sum_{i=1}^n w_{ni}^2 / \sum_{i=1}^n (w_{ni}^*)^2 \to 1$ *and* $\sum_{i=1}^n \alpha_i w_{ni} / \sum_{i=1}^n \alpha_i w_{ni}^* \to 1$. *Equivalently, this occurs if and only if both*

$$\begin{aligned} &n^{4/(d+4)} \sum_{i=1}^n \{w_{ni}^2 - (w_{ni}^*)^2\} \to 0 \quad \textit{and} \\ &n^{-8/(d(d+4))} \sum_{i=1}^n \alpha_i (w_{ni} - w_{ni}^*) \to 0. \end{aligned} \tag{2.6}$$

Finally,

$$\begin{aligned} &n^{4/(d+4)} \{R_{\mathcal{R}}(\hat{C}^{\mathrm{wnn}}_{n,\mathbf{w}_n^*}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})\} \\ &\quad \to \frac{(d+2)^{(2d+4)/(d+4)}}{2^{4/(d+4)}} \left(\frac{d+4}{d}\right)^{d/(d+4)} B_1^{4/(d+4)} B_2^{d/(d+4)}. \end{aligned} \tag{2.7}$$

Now write $\hat{C}^{\mathrm{nn}}_{n,k}$ for the traditional, unweighted k-nearest neighbour classifier (or equivalently, the weighted nearest neighbour classifier with $w_{ni} = 1/k$ for $i = 1, \ldots, k$ and $w_{ni} = 0$ otherwise). Another consequence of Theorem 1 is that, provided (A.1)–(A.4) hold and $B_2 > 0$, the quantity k^* defined in (2.4) is larger by a factor of $\{\frac{2(d+4)}{d+2}\}^{d/(d+4)}$ (up to an unimportant rounding error) than the asymptotically optimal choice of k^{opt} for $\hat{C}^{\mathrm{nn}}_{n,k}$; see also Hall, Park and Samworth (2008). We can therefore compare the performance of $\hat{C}^{\mathrm{nn}}_{n,k^{\mathrm{opt}}}$ with that of $\hat{C}^{\mathrm{wnn}}_{n,\mathbf{w}_n^*}$.

COROLLARY 3. *Assume* (A.1)–(A.4) *and assume also that* $B_2 > 0$. *Then*

$$\frac{R_{\mathcal{R}}(\hat{C}^{\mathrm{wnn}}_{n,\mathbf{w}_n^*}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})}{R_{\mathcal{R}}(\hat{C}^{\mathrm{nn}}_{n,k^{\mathrm{opt}}}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})} \to \frac{1}{4^{d/(d+4)}} \left(\frac{2d+4}{d+4}\right)^{(2d+4)/(d+4)} \tag{2.8}$$

as $n \to \infty$.

Since the limit in (2.8) does not depend on the underlying populations, we can plot it as a function of d; cf. Figure 2. In fact, Corollary 3 suggests a natural correspondence between any unweighted k-nearest neighbour classifier $\hat{C}^{\mathrm{nn}}_{n,k}$ and the weighted nearest neighbour classifier which we denote by $\hat{C}^{\mathrm{wnn}}_{n,\mathbf{w}_n^{\mu(k)}}$ whose weights are of the optimal form (1.1), but with k^* replaced with

$$\mu(k) = \left\lfloor \left\{\frac{2(d+4)}{d+2}\right\}^{d/(d+4)} k \right\rfloor. \tag{2.9}$$

Under the conditions of Corollary 3, we can compare $\hat{C}^{\mathrm{nn}}_{n,k}$ and $\hat{C}^{\mathrm{wnn}}_{n,\mathbf{w}_n^{\mu(k)}}$, concluding that for each $\beta \in (0, 1/2)$,

$$\frac{R_{\mathcal{R}}(\hat{C}^{\mathrm{wnn}}_{n,\mathbf{w}_n^{\mu(k)}}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})}{R_{\mathcal{R}}(\hat{C}^{\mathrm{nn}}_{n,k}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})} \to \frac{1}{4^{d/(d+4)}}\left(\frac{2d+4}{d+4}\right)^{(2d+4)/(d+4)} \tag{2.10}$$

as $n \to \infty$, uniformly for $n^{\beta} \le k \le n^{1-\beta}$. The fact that the convergence in (2.10) is uniform for k in this range means that the ratio on the left-hand side of (2.10) has the same limit if we replace k by an estimator $\hat{k}$ constructed from the training data $(X_1, Y_1), \dots, (X_n, Y_n)$, provided that $\hat{k}$ lies in this range with probability tending to 1.

In a complementary approach to that taken in most of this paper, Audibert and Tsybakov (2007) study the minimax properties of plug-in classifiers. They show in particular that a certain classifier obtained by modifying a local polynomial estimator of the regression function η attains the minimax rate over a set of distributions P of random vectors (X, Y) on $\mathbb{R}^d \times \{1, 2\}$ for which the regression function belongs to a Hölder class, P satisfies a margin condition and the marginal distribution of X satisfies a so-called *strong density assumption*. This rate is $O(n^{-4/(d+4)})$ when the Hölder smoothness parameter is 2, and the margin power parameter is 1. By adapting their arguments, we are able to show in the supplementary material [Samworth (2012)] that several weighted nearest-neighbour classifiers (including the unweighted, optimally weighted and bagged versions of Section 3) can also attain this minimax rate. Such results give reassurance about worst-case behaviour; however, they do not lead naturally to an optimal weighting scheme or a quantification of the relative performance of two weighted nearest neighbour classifiers attaining the same rate. These are the main goals of this work.

Finally in this section, we note that the theory presented above can be extended in a natural way to multicategory classification problems, where the class labels take values in the set $\{1, \dots, K\}$. Writing $\eta_r(x) = \mathbb{P}(Y = r | X = x)$, let

$$\mathcal{S}_{r_1,r_2} = \Big\{x \in \mathcal{R} \colon \operatorname*{argmax}_{r \in \{1,\dots,K\}} \eta_r(x) = \{r_1, r_2\}\Big\}$$

for distinct indices $r_1, r_2 \in \{1, \dots, K\}$. In addition to (A.1) and the obvious analogues of the conditions (A.2), (A.3) and (A.4), we require:

(A.5) For each $(r_1, r_2) \neq (r_3, r_4)$, the submanifolds $\mathcal{S}_{r_1,r_2}$ and $\mathcal{S}_{r_3,r_4}$ of $\mathbb{R}^d$ are transversal.

Condition (A.5) ensures that $\mathcal{S}_{r_1,r_2} \cap \mathcal{S}_{r_3,r_4} \cap (\mathcal{R} \setminus \partial\mathcal{R})$ is either empty or a $(d-2)$-dimensional submanifold of $\mathbb{R}^d$ [Guillemin and Pollack (1974), page 30]. Under these conditions, the conclusion of Theorem 1 holds, provided that the constants B_1 and B_2 are replaced with $\tilde{B}_1 = \sum_{r_1 \neq r_2} B_{1,r_1,r_2}$ and $\tilde{B}_2 = \sum_{r_1 \neq r_2} B_{2,r_1,r_2}$, respectively, where each term B_{1,r_1,r_2} and B_{2,r_1,r_2} is an integral over $\mathcal{S}_{r_1,r_2}$. Apart from the obvious notational changes involved in converting B_1 and B_2 to B_{1,r_1,r_2} and B_{2,r_1,r_2}, the only other change required is to replace the constant factor $1/4$ in the definition of B_1 with $\eta_{r_1,r_2}(x_0)\{1-\eta_{r_1,r_2}(x_0)\}$ where $\eta_{r_1,r_2}(x_0)$ denotes the common value that η_{r_1} and η_{r_2} take at $x_0 \in \mathcal{S}_{r_1,r_2}$. This change accounts for the fact that $\eta_{r_1,r_2}(x_0)$ is not necessarily equal to $1/2$ on $\mathcal{S}_{r_1,r_2}$.

It follows (provided also that $\tilde{B}_2 > 0$) that the asymptotically optimal weights are still of the form (1.1), but with the ratio B_1/B_2 in the expression for k^* in (2.4) replaced with $\tilde{B}_1/\tilde{B}_2$. Moreover, the conclusion of Corollary 3 and the subsequent discussion also remain true.

3. The bagged nearest neighbour classifier. Traditionally, the bagged nearest neighbour classifier is obtained by applying the 1-nearest neighbour classifier to many resamples from the training data. The final classification is made by a majority vote on the classifications obtained from the resamples. In the most common version of bagging where the resamples are drawn with replacement, and the resample size is the same as the original sample size, bagging the nearest neighbour classifier gives no improvement over the 1-nearest neighbour classifier [Hall and Samworth (2005)]. This is because the nearest neighbour occurs in more than half (in fact, roughly a proportion $1 - 1/e$) of the resamples.

Nevertheless, if a smaller resample size is used, then substantial improvements over the nearest neighbour classifier are possible, as has been verified empirically by Martínez-Muñoz and Suárez (2010). In fact, if the resample size is m, then the "infinite simulation" versions of the bagged nearest neighbour classifier in the with- and without-replacement resampling cases are weighted nearest neighbour classifiers with respective weights

$$w_{ni}^{\mathrm{b,with}} = \left(1 - \frac{i-1}{n}\right)^m - \left(1 - \frac{i}{n}\right)^m, \qquad i = 1, \dots, n \tag{3.1}$$

and

$$w_{ni}^{\mathrm{b,w/o}} = \begin{cases} \binom{n-i}{m-1} \Big/ \binom{n}{m}, & \text{for } i = 1, \dots, n-m+1, \\ 0, & \text{for } i = n-m+2, \dots, n. \end{cases} \tag{3.2}$$

Of course, the observations above render the resampling redundant, and we regard the weighted nearest neighbour classifiers with the weights above as defining the two versions of the bagged nearest neighbour classifier. It is convenient to let $q = m/n$ denote the resampling fraction. Intuitively, for large n, both versions of the bagged nearest neighbour classifier behave like the weighted nearest neighbour classifier with weights $(w_{ni}^{\mathrm{Geo}})_{i=1}^n$ which place a Geometric(q) distribution (conditioned on being in the set $\{1, \dots, n\}$) on the weights

$$w_{ni}^{\mathrm{Geo}} = \frac{q(1-q)^{i-1}}{1-(1-q)^n}, \qquad i = 1, \dots, n. \tag{3.3}$$

The reason for this is that, in order for the ith nearest neighbour of the training data to be the nearest neighbour of the resample, the nearest $i-1$ neighbours must not appear in the resample, while the ith nearest neighbour must appear, and these events are almost independent when n is large; see Hall and Samworth (2005). Naturally, the parameter q plays a crucial role in the performance of the bagged nearest neighbour classifier, and for small $\beta > 0$, the three vectors of weights given in (3.1), (3.2) and (3.3) belong to $W_{n,\beta}$ for all large n if $\max(\frac{1}{2}n^{-(1-\beta d/4)}, n^{-(1-2\beta)}) \leq q \leq 3n^{-\beta}$. In the following corollary of Theorem 1, we write $\hat{C}^{\text{bnn}}_{n,q}$ to denote either of the bagged nearest neighbour classifiers with weights (3.1), (3.2) or their approximation with weights (3.3).

COROLLARY 4. *Assume* (A.1)–(A.4). *For every* $\beta \in (0, 1/2)$,

$$R_{\mathcal{R}}(\hat{C}^{\text{bnn}}_{n,q}) - R_{\mathcal{R}}(C^{\text{Bayes}}) = \tilde{\gamma}_n(q)\{1 + o(1)\},$$

uniformly for $n^{-(1-\beta)} \leq q \leq n^{-\beta}$, *where*

$$\tilde{\gamma}_n(q) = \frac{B_1}{2}q + \frac{B_2\Gamma(2+2/d)^2}{n^{4/d}q^{4/d}}.$$

This result is somewhat related to Corollary 10 of Biau, Cérou and Guyader (2010). In that paper, the authors study the bagged nearest neighbour estimate $\hat{\eta}_n$ of the regression function η. They prove in particular that under regularity conditions (including a Lipschitz assumption on η) and for a suitable choice of resample size,

$$\mathbb{E}[\{\hat{\eta}_n(X) - \eta(X)\}^2] = O(n^{-2/(d+2)})$$

for $d \geq 3$. It is known [e.g., Ibragimov and Khasminskiĭ (1980, 1981, 1982)] that this is the minimax optimal rate for their problem.

Corollary 4 may also be applied to deduce that the asymptotically optimal choice of q in all three cases is

$$q^{\text{opt}} = \frac{8^{d/(d+4)}\Gamma(2+2/d)^{2d/(d+4)}}{d^{d/(d+4)}}\left(\frac{B_2}{B_1}\right)^{d/(d+4)} n^{-4/(d+4)}.$$

Thus, in an analogous fashion to Section 2, we can consider the performance of $\hat{C}^{\text{bnn}}_{n,q^{\text{opt}}}$ relative to that of $\hat{C}^{\text{nn}}_{n,k^{\text{opt}}}$.

COROLLARY 5. *Assume* (A.1)–(A.4) *and assume also that* $B_2 > 0$. *Then*

$$\frac{R_{\mathcal{R}}(\hat{C}^{\text{bnn}}_{n,q^{\text{opt}}}) - R_{\mathcal{R}}(C^{\text{Bayes}})}{R_{\mathcal{R}}(\hat{C}^{\text{nn}}_{n,k^{\text{opt}}}) - R_{\mathcal{R}}(C^{\text{Bayes}})} \to \frac{\Gamma(2+2/d)^{2d/(d+4)}}{2^{4/(d+4)}} \tag{3.4}$$

as $n \to \infty$.

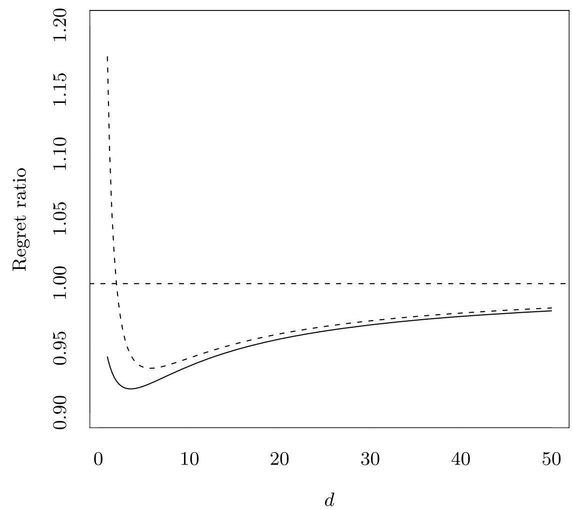

FIG. 3. *Asymptotic ratio of the regret of the bagged nearest neighbour classifier* (*dashed*) *to that of the* k*-nearest neighbour classifier, as a function of the dimension of the feature vectors. The asymptotic regret ratio for the optimally weighted nearest neighbour classifier compared with the* k*-nearest neighbour classifier is shown as a solid line for comparison.*

The limiting ratio in (3.4) is plotted as a function of d in Figure 3. The ratio is about 1.18 when $d = 1$, showing that the bagged nearest neighbour classifier has asymptotically worse performance than the k-nearest neighbour classifier in this case. The ratio is equal to 1 when $d = 2$, and is less than 1 for $d \geq 3$. The facts that the asymptotically optimal weights decay as illustrated in Figure 1 and that the bagged nearest neighbour weights decay approximately geometrically explain why the bagged nearest neighbour classifier has almost optimal performance among nonnegatively weighted nearest neighbour classifiers when d is large.

Similar to the discussion following Corollary 3, based on the expressions for k^{opt} and q^{opt}, there is a natural correspondence between the unweighted k-nearest neighbour classifier $\hat{C}^{\mathrm{nn}}_{n,\hat{k}}$ with data driven $\hat{k}$, and the bagged nearest neighbour classifier $\hat{C}^{\mathrm{bnn}}_{n,\hat{q}}$, where

$$\hat{q} = 2^{d/(d+4)}\Gamma\left(2 + \frac{2}{d}\right)^{2d/(d+4)} \frac{1}{\hat{k}}. \tag{3.5}$$

The same limit (3.4) holds for the regret ratio of these classifiers, again provided there exists $\beta \in (0, 1/2)$ such that $\mathbb{P}(n^{\beta} \leq \hat{k} \leq n^{1-\beta}) \to 1$.

4. Faster rates of convergence. If we allow negative weights, it is possible to choose weights satisfying $\sum_{i=1}^{n} \alpha_i w_{ni} = 0$. This means that we can eradicate the dominant squared bias term in the asymptotic expansion of Theorem 1. It follows that, subject to additional smoothness conditions, we can achieve faster rates of convergence with weighted nearest neighbour classifiers, as we now describe. The appropriate variant of condition (A.2), which we denote by (A.2)(r), is as follows:

(A.2)(r) The set $\mathcal{S} = \{x \in \mathcal{R} : \eta(x) = 1/2\}$ is nonempty. There exists an open subset U_0 of $\mathbb{R}^d$ that contains $\mathcal{S}$ and such that the following properties hold: first, η is continuous on $U \setminus U_0$, where U is an open set containing $\mathcal{R}$; second, the restrictions of P_1 and P_2 to U_0 are absolutely continuous with respect to Lebesgue measure, with $2r$-times continuously differentiable Radon–Nikodym derivatives f_1 and f_2, respectively.

Thus condition (A.2)(1) is identical to (A.2). Note that we are still in the setting of a margin condition with power parameter equal to 1. Let S denote the set of multi-indices $s = (s_1, \dots, s_d)$, so s is a d-tuple of nonnegative integers. For $s \in S$, we write $|s| = s_1 + \cdots + s_d$, and for $v = (v_1, \dots, v_d)^T \in \mathbb{R}^d$, we write $v^s = v_1^{s_1} v_2^{s_2} \cdots v_d^{s_d}$. Now, for $s \in S$, let $c_{s,d} = \int_{\|v\| \le 1} v^s \, dv$. It is convenient here to use multi-index notation for derivatives, so we write $g_s(x) = \frac{\partial^{|s|}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}} g(x)$. Now let

$$\bar{S}_r = \{(s^1, s^2) \in S \times S : |s^1| + |s^2| = 2r, |s^1| \ge 1, s_j^1 + s_j^2 \in 2\mathbb{Z} \ \forall j = 1, \dots, d\},$$

and let

$$a^{(r)}(x) = \frac{1}{a_d^{1+2r/d} \bar{f}(x)^{1+2r/d}} \sum_{(s^1, s^2) \in \bar{S}_r} \frac{c_{s^1+s^2,d} \, \eta_{s^1}(x) \bar{f}_{s^2}(x)}{|s^1|!|s^2|!};$$

thus $a^{(1)}(x) = a(x)$. Further, let

$$B_2^{(r)} = \int_{\mathcal{S}} \frac{\bar{f}(x_0)}{\|\dot{\eta}(x_0)\|} a^{(r)}(x_0)^2 \, d\mathrm{Vol}^{d-1}(x_0).$$

For $\ell \in \mathbb{N}$, define $\alpha_i^{(\ell)} = i^{1+2\ell/d} - (i-1)^{1+2\ell/d}$. We consider restrictions on the set of weight vectors analogous to those imposed on rth order kernels in kernel density estimation. Specifically, we let $W_{n,\beta,r}^{\dagger}$ denote the set of deterministic weight vectors $\mathbf{w}_n = (w_{ni})_{i=1}^n$ satisfying:

- $\sum_{i=1}^n w_{ni} = 1$, $n^{2r/d} \sum_{i=1}^n \alpha_i^{(\ell)} w_{ni} / n^{2\ell/d} \sum_{i=1}^n \alpha_i^{(r)} w_{ni} \le 1/\log n$ for $\ell = 1, \dots, r-1$;
- $\sum_{i=1}^n w_{ni}^2 \le n^{-\beta}$;
- $n^{-4r/d} (\sum_{i=1}^n \alpha_i^{(r)} w_{ni})^2 \le n^{-\beta}$;
- there exists $k_2 \le \lfloor n^{1-\beta} \rfloor$ such that $n^{2r/d} \sum_{i=k_2+1}^n |w_{ni}| / \sum_{i=1}^n \alpha_i^{(r)} w_{ni} \le 1/\log n$ and such that $\sum_{i=1}^{k_2} \alpha_i^{(r)} w_{ni} \ge \beta k_2^{2r/d}$;
- $\sum_{i=k_2+1}^n w_{ni}^2 / \sum_{i=1}^n w_{ni}^2 \le 1/\log n$;
- $\sum_{i=1}^n |w_{ni}|^3 / (\sum_{i=1}^n w_{ni}^2)^{3/2} \le 1/\log n$.

Finally, we are in a position to state the analogue of Theorem 1 for weight vectors in $W_{n,\beta,r}^{\dagger}$.

THEOREM 6. *Assume* (A.1), (A.2)(r), (A.3) *and* (A.4). *Then for each* $\beta \in (0, 1/2)$,

$$R_{\mathcal{R}}(\hat{C}_n^{\mathrm{wnn}}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}}) = \gamma_n^{(r)}(\mathbf{w}_n)\{1 + o(1)\} \tag{4.1}$$

as $n \to \infty$, *uniformly for* $\mathbf{w}_n \in W_{n,\beta,r}^{\dagger}$, *where*

$$\gamma_n^{(r)}(\mathbf{w}_n) = B_1 \sum_{i=1}^n w_{ni}^2 + B_2^{(r)} \Biggl(\sum_{i=1}^n \frac{\alpha_i^{(r)} w_{ni}}{n^{2r/d}}\Biggr)^2. \tag{4.2}$$

A consequence of Theorem 6 is that we can construct weighted nearest neighbour classifiers which, under conditions (A.1), (A.2)(r), (A.3) and (A.4), and provided that $B_2^{(r)} > 0$, achieve the rate of convergence $O(n^{-4r/(4r+d)})$ for the regret. To illustrate this, set $k^{*(r)} = \lfloor B^{*(r)} n^{4r/(4r+d)} \rfloor$, and in order to satisfy the restrictions on the allowable weights, consider weight vectors with $w_{ni} = 0$ for $i = k^{*(r)} + 1, \ldots, n$. Then, by mimicking the proof of Theorem 2 and seeking to minimise (4.2) subject to the constraints $\sum_{i=1}^{k^{*(r)}} w_{ni} = 1$ and $\sum_{i=1}^{k^{*(r)}} \alpha_i^{(\ell)} w_{ni} = 0$ for $\ell = 1, \ldots, r-1$, we obtain minimising weights of the form

$$w_{ni}^{*(r)} = \begin{cases} \dfrac{1}{k^{*(r)}}(b_0 + b_1 \alpha_i^{(1)} + \cdots + b_r \alpha_i^{(r)}), & \text{for } i = 1, \ldots, k^{*(r)}, \\ 0, & \text{for } i = k^{*(r)} + 1, \ldots, n. \end{cases} \tag{4.3}$$

The equations $\sum_{i=1}^n w_{ni} = 1$ and $\sum_{i=1}^n \alpha_i^{(\ell)} w_{ni} = 0$ for $\ell = 1, \ldots, r-1$ for weight vectors of the form (4.3) yield r linear equations in the $r+1$ unknowns $b_0, b_1, \ldots, b_r$. Although these equations can be solved directly in terms of b_0 say, simpler expressions are obtained by solving asymptotic approximations to these equations. In particular, since it is an elementary fact that for nonnegative integers ℓ_1 and ℓ_2,

$$\sum_{i=1}^k \alpha_i^{(\ell_1)} \alpha_i^{(\ell_2)} = \frac{(d + 2\ell_1)(d + 2\ell_2)}{d(d + 2\ell_1 + 2\ell_2)} k^{1 + 2(\ell_1 + \ell_2)/d}\{1 + O(k^{-2})\}$$

as $k \to \infty$, we can just deal with the dominant terms. As examples, when $r = 1$, we find

$$b_1 = \frac{1}{(k^{*(1)})^{2/d}}(1 - b_0),$$

and when $r = 2$, we should take

$$b_1 = \frac{1}{(k^{*(2)})^{2/d}} \biggl\{\frac{(d+4)^2}{4} - \frac{2(d+4)}{d+2} b_0\biggr\} \quad \text{and} \quad b_2 = \frac{1 - b_0 - (k^{*(2)})^{2/d} b_1}{(k^{*(2)})^{4/d}}.$$

Under the conditions of Theorem 6, and provided $B_2^{(r)} > 0$, these weighted nearest neighbour classifiers achieve the $O(n^{-4r/(4r+d)})$ convergence rate. The choice of b_0 involves a trade-off between the desire to keep the remaining squared bias term $B_2^{(r)}(\sum_{i=1}^{k^{*(r)}} \frac{\alpha_i^{(r)} w_{ni}^{*(r)}}{n^{2r/d}})^2$ small, and the need for it to be large enough to remain the dominant bias term. This reflects the fact that the asymptotic results of this section should be applied with some caution. Besides the discomfort many practitioners might feel in using negative weights, one would anticipate that rather large sample sizes would be needed for the leading terms in the asymptotic expansion (4.1) to dominate the error terms. This is also the reason why we do not pursue here methods such as Lepski's method [Lepskiĭ (1991)] that adapt to an unknown smoothness level around $\mathcal{S}$.

5. Empirical performance study. In this section, we assess the relative empirical performance of the k-nearest neighbour classifier, the optimally weighted nearest neighbour classifier of Section 2 and the bagged nearest neighbour classifier of Section 3 on simulated and real data sets. We consider four general simulation settings, designed to exhibit different distributional characteristics:

Setting 1: f_1 is the density of d independent components, each having a standard Laplace distribution, and f_2 is the density of the $N_d(\theta, I)$ distribution, where θ denotes a d-vector of ones.

Setting 2: f_1 is the density of d independent components, each having the mixture of normals distribution $\frac{1}{2}N(0, 1) + \frac{1}{2}N(3, 2)$. Likewise, f_2 is the density of d independent components, each having a $\frac{1}{2}N(1.5, 1) + \frac{1}{2}N(4.5, 2)$ distribution.

Setting 3: For $d \geq 2$, let Σ denote the $d \times d$ Toeplitz matrix whose jth entry of its first row is 0.6^{j-1}. Set f_1 to be the density of the $\frac{1}{2}N_d(0, \Sigma) + \frac{1}{2}N_d(3\theta, 2\Sigma)$ distribution, and f_2 to be the density of the $\frac{1}{2}N_d(3\theta/2, \Sigma) + \frac{1}{2}N_d(9\theta/2, 2\Sigma)$ distribution.

Setting 4: Both f_1 and f_2 are densities of independent components. For f_1, each component has a standard Cauchy density. For f_2, the first $\lfloor d/2 \rfloor$ components also have a standard Cauchy density, while the last $d - \lfloor d/2 \rfloor$ components have a standard Laplace density.

Setting 1 is a relatively benign classification problem. Setting 2 explores the effect of bimodality, and setting 3 combines bimodal marginals with dependence between the components. Setting 4 combines heavy-tailed distributions, a lack of location difference and introduces components which are irrelevant for classification as nuisance variables. For each setting, we examined the three sample sizes $n \in \{50, 200, 1000\}$, five dimensions $d \in \{1, 2, 3, 5, 10\}$ (except for setting 3, where the $d = 1$ case was omitted as it is covered in setting 1) and two prior probabilities $\pi \in \{1/2, 2/3\}$. Thus there were 114 simulation scenarios in total, and we used the Euclidean norm for computing distances throughout.

In each scenario, we took $\mathcal{R} = \mathbb{R}^d$ and computed the Bayes risk by Monte Carlo integration. For each data set of size n drawn from the relevant populations, we used a slight variant of a 5-fold cross validation algorithm to compute $\hat{k}$, the number of neighbours used by the k-nearest neighbour classifier. Specifically, we assigned each observation independently and uniformly at random to one of five groups, and found the minimiser, denoted $\tilde{k}$, of the cross-validation risk over a grid of 21 equally spaced points (up to integer rounding) from 5 to $n/2$. The variant arises from the observation that this minimiser targets the optimal value of k for a data set of size $4n/5$. Bearing in mind the expression for the optimal k^* in (2.4), we therefore set $\hat{k} = (\frac{5}{4})^{4/(d+4)}\tilde{k}$ as an appropriate choice for a data set of size n. The number of positive weights for the optimally weighted classifier was then chosen to be $\mu(\hat{k})$; cf (2.9). For the bagged nearest neighbour classifier, we used the "geometric" weights given in (3.3), with q given by $\hat{q}$ in (3.5). For each data set, we computed the proportion of misclassifications of $n_{\mathrm{test}} = 1000$ independent test points drawn from the appropriate distribution, and each simulation was repeated 1000 times to yield estimates of the risks of each of the three classifiers.

It is computationally convenient to evaluate the distance matrix between all $n_+ = n + n_{\mathrm{test}}$ points at the outset (even though some distances will not be used), and this takes $O(n_+^2 d)$ operations when $\|\cdot\|$ is an ℓ_p-norm. It then takes a further $O(n_+ n \log n)$ operations to choose $\hat{k}$ and classify the test points. In particular, the computational requirements are of the same order of magnitude for both the unweighted and weighted nearest neighbour classifiers.

An alternative to using a cross-validation method for choosing $\hat{k}$, as pointed out by an anonymous referee, is to estimate the constants B_1 and B_2 in (2.4) directly using a plug-in approach. We discuss this approach in the supplementary material [Samworth (2012)] following the proof of Theorem 6, but conclude that it seems awkward to propose a satisfactory algorithm for estimating B_1 and B_2 directly, and do not pursue it further here.

Our simulation results are presented in Tables 1 and 2. To save space, we have omitted the results for $\pi = 2/3$, which were qualitatively similar. As well as the risks for the three classifiers, we present in the final two columns estimates of the regret ratios

$$\frac{R(\hat{C}^{\mathrm{wnn}}_{n,\mathbf{w}_n^{\mu(\hat{k})}}) - R(C^{\mathrm{Bayes}})}{R(\hat{C}^{\mathrm{nn}}_{n,\hat{k}}) - R(C^{\mathrm{Bayes}})} \quad \text{and} \quad \frac{R(\hat{C}^{\mathrm{bnn}}_{n,\hat{q}}) - R(C^{\mathrm{Bayes}})}{R(\hat{C}^{\mathrm{nn}}_{n,\hat{k}}) - R(C^{\mathrm{Bayes}})}, \tag{5.1}$$

respectively. Standard errors for these estimates are also given, and were obtained using the delta method.

In 54 of the 57 scenarios in Tables 1 and 2, the risk of the optimally weighted nearest neighbour classifier is smaller than that of the k-nearest neighbour classifier. In one of the three exceptional cases, the difference is so small that it can easily be explained by the Monte Carlo error. The other cases are in setting 1 with $d = 10$ and $n = 50$, 200. Here it seems that in this relatively large dimension for nonparametric inference, these sample sizes are not large enough for the asymptotics to provide a good approximation.

TABLE 1

The estimated risks (multiplied by 100) of the Bayes, k-nearest neighbour, optimally weighted nearest neighbour and bagged nearest neighbour classifiers in settings 1 and 2. The final two columns give the regret ratios defined in (5.1). Standard errors are given in small script

d	Bayes	n	knn risk	ownn risk	bnn risk	ownn rr	bnn rr
Setting 1							
1	30.02	50	$33.93_{0.14}$	$33.77_{0.14}$	$34.71_{0.14}$	$0.96_{0.050}$	$1.20_{0.057}$
		200	$31.53_{0.066}$	$31.47_{0.067}$	$31.72_{0.075}$	$0.96_{0.061}$	$1.10_{0.071}$
		1000	$30.72_{0.046}$	$30.70_{0.046}$	$30.72_{0.046}$	$0.97_{0.093}$	$1.00_{0.094}$
2	24.21	50	$28.77_{0.11}$	$28.58_{0.11}$	$28.92_{0.11}$	$0.96_{0.034}$	$1.00_{0.035}$
		200	$26.50_{0.055}$	$26.42_{0.056}$	$26.51_{0.058}$	$0.97_{0.034}$	$1.00_{0.035}$
		1000	$25.67_{0.046}$	$25.62_{0.046}$	$25.60_{0.046}$	$0.96_{0.044}$	$0.95_{0.044}$
3	19.37	50	$25.35_{0.10}$	$25.03_{0.098}$	$25.23_{0.097}$	$0.95_{0.023}$	$0.98_{0.023}$
		200	$22.82_{0.052}$	$22.69_{0.053}$	$22.72_{0.053}$	$0.96_{0.021}$	$0.97_{0.021}$
		1000	$21.54_{0.045}$	$21.44_{0.045}$	$21.43_{0.046}$	$0.95_{0.029}$	$0.95_{0.029}$
5	13.17	50	$20.37_{0.093}$	$20.26_{0.095}$	$20.40_{0.093}$	$0.98_{0.018}$	$1.00_{0.018}$
		200	$17.74_{0.049}$	$17.54_{0.050}$	$17.55_{0.050}$	$0.96_{0.015}$	$0.96_{0.015}$
		1000	$16.21_{0.041}$	$16.03_{0.042}$	$16.07_{0.044}$	$0.94_{0.019}$	$0.96_{0.019}$
10	5.592	50	$13.89_{0.10}$	$14.59_{0.12}$	$14.63_{0.11}$	$1.10_{0.019}$	$1.10_{0.019}$
		200	$11.35_{0.050}$	$11.70_{0.051}$	$11.72_{0.050}$	$1.10_{0.013}$	$1.10_{0.013}$
		1000	$9.977_{0.033}$	$9.796_{0.033}$	$9.911_{0.033}$	$0.96_{0.010}$	$0.98_{0.011}$
Setting 2							
1	34.85	50	$38.96_{0.14}$	$38.78_{0.13}$	$39.01_{0.11}$	$0.96_{0.046}$	$1.00_{0.043}$
		200	$36.76_{0.073}$	$36.63_{0.073}$	$36.83_{0.075}$	$0.93_{0.052}$	$1.00_{0.056}$
		1000	$35.34_{0.052}$	$35.30_{0.052}$	$35.35_{0.052}$	$0.91_{0.14}$	$1.00_{0.15}$
2	26.83	50	$34.36_{0.13}$	$33.53_{0.13}$	$33.43_{0.12}$	$0.89_{0.023}$	$0.88_{0.022}$
		200	$30.00_{0.070}$	$29.63_{0.068}$	$29.66_{0.070}$	$0.88_{0.029}$	$0.89_{0.030}$
		1000	$27.56_{0.050}$	$27.48_{0.050}$	$27.49_{0.050}$	$0.89_{0.091}$	$0.90_{0.091}$
3	21.73	50	$31.07_{0.11}$	$30.07_{0.11}$	$30.01_{0.11}$	$0.89_{0.016}$	$0.89_{0.016}$
		200	$26.44_{0.063}$	$25.99_{0.063}$	$25.96_{0.065}$	$0.90_{0.018}$	$0.90_{0.018}$
		1000	$23.19_{0.045}$	$23.04_{0.046}$	$23.03_{0.046}$	$0.90_{0.042}$	$0.90_{0.042}$
5	15.23	50	$25.72_{0.12}$	$24.88_{0.11}$	$25.12_{0.11}$	$0.92_{0.015}$	$0.94_{0.015}$
		200	$21.51_{0.055}$	$20.92_{0.054}$	$20.93_{0.055}$	$0.91_{0.012}$	$0.91_{0.012}$
		1000	$18.69_{0.045}$	$18.34_{0.046}$	$18.33_{0.047}$	$0.90_{0.018}$	$0.90_{0.018}$
10	7.146	50	$16.87_{0.099}$	$16.57_{0.10}$	$16.88_{0.10}$	$0.97_{0.014}$	$1.00_{0.015}$
		200	$13.00_{0.048}$	$12.77_{0.051}$	$12.87_{0.051}$	$0.96_{0.012}$	$0.98_{0.012}$
		1000	$11.57_{0.034}$	$11.41_{0.033}$	$11.44_{0.034}$	$0.96_{0.011}$	$0.97_{0.011}$

The extent of the improvement of the optimally weighted nearest neighbour classifier is generally in close agreement with that predicted by the theory of Corollary 3 and the paragraph which follows it, even for small sample sizes. This theory tells us that the first regret ratio in (5.1) converges to 0.943, 0.924, 0.919, 0.920 and 0.936 in dimensions $d = 1, 2, 3, 5, 10$, respectively. Note that a few of the re-

TABLE 2

The estimated risks (multiplied by 100) of the Bayes, k-nearest neighbour, optimally weighted nearest neighbour and bagged nearest neighbour classifiers in settings 3 and 4. The final two columns give the regret ratios defined in (5.1). Standard errors are given in small script

d	Bayes	n	knn risk	ownn risk	bnn risk	ownn rr	bnn rr
Setting 3							
2	32.45	50	$37.91_{0.13}$	$37.40_{0.12}$	$37.54_{0.11}$	$0.91_{0.030}$	$0.93_{0.030}$
		200	$35.08_{0.065}$	$34.96_{0.065}$	$35.05_{0.068}$	$0.95_{0.034}$	$0.99_{0.036}$
		1000	$33.70_{0.052}$	$33.65_{0.051}$	$33.67_{0.051}$	$0.96_{0.057}$	$0.98_{0.057}$
3	30.00	50	$36.56_{0.13}$	$35.94_{0.11}$	$36.00_{0.11}$	$0.91_{0.024}$	$0.91_{0.024}$
		200	$33.61_{0.065}$	$33.52_{0.066}$	$33.58_{0.067}$	$0.97_{0.025}$	$0.99_{0.026}$
		1000	$32.03_{0.052}$	$31.96_{0.052}$	$31.94_{0.052}$	$0.96_{0.036}$	$0.96_{0.036}$
5	26.13	50	$34.10_{0.14}$	$33.41_{0.12}$	$33.47_{0.11}$	$0.91_{0.021}$	$0.92_{0.021}$
		200	$30.27_{0.068}$	$30.16_{0.070}$	$30.26_{0.070}$	$0.97_{0.023}$	$1.00_{0.024}$
		1000	$28.41_{0.051}$	$28.23_{0.051}$	$28.25_{0.052}$	$0.92_{0.030}$	$0.93_{0.031}$
10	18.26	50	$27.03_{0.13}$	$26.50_{0.11}$	$26.59_{0.11}$	$0.94_{0.019}$	$0.95_{0.019}$
		200	$22.86_{0.067}$	$22.90_{0.070}$	$23.01_{0.071}$	$1.00_{0.021}$	$1.00_{0.022}$
		1000	$21.07_{0.046}$	$20.91_{0.046}$	$20.92_{0.046}$	$0.94_{0.022}$	$0.95_{0.023}$
Setting 4							
1	41.95	50	$47.73_{0.099}$	$47.49_{0.10}$	$47.08_{0.094}$	$0.96_{0.024}$	$0.89_{0.022}$
		200	$45.64_{0.078}$	$45.45_{0.077}$	$45.24_{0.072}$	$0.95_{0.029}$	$0.89_{0.027}$
		1000	$43.38_{0.061}$	$43.28_{0.060}$	$43.32_{0.061}$	$0.93_{0.058}$	$0.96_{0.059}$
2	41.96	50	$48.36_{0.079}$	$48.05_{0.083}$	$47.85_{0.081}$	$0.95_{0.017}$	$0.92_{0.017}$
		200	$46.39_{0.074}$	$46.05_{0.072}$	$45.96_{0.070}$	$0.92_{0.022}$	$0.90_{0.022}$
		1000	$44.13_{0.060}$	$43.91_{0.060}$	$43.86_{0.060}$	$0.90_{0.037}$	$0.88_{0.037}$
3	36.37	50	$46.32_{0.10}$	$45.73_{0.10}$	$45.50_{0.10}$	$0.94_{0.014}$	$0.92_{0.014}$
		200	$42.92_{0.083}$	$42.38_{0.081}$	$42.29_{0.078}$	$0.92_{0.017}$	$0.90_{0.017}$
		1000	$39.36_{0.058}$	$39.04_{0.057}$	$39.03_{0.058}$	$0.89_{0.026}$	$0.89_{0.026}$
5	32.00	50	$45.66_{0.10}$	$44.80_{0.11}$	$44.57_{0.11}$	$0.94_{0.011}$	$0.92_{0.010}$
		200	$40.89_{0.085}$	$40.23_{0.080}$	$40.22_{0.078}$	$0.93_{0.013}$	$0.93_{0.012}$
		1000	$36.90_{0.056}$	$36.45_{0.056}$	$36.44_{0.056}$	$0.91_{0.015}$	$0.91_{0.015}$
10	25.40	50	$45.27_{0.099}$	$44.21_{0.10}$	$43.97_{0.098}$	$0.95_{0.0069}$	$0.93_{0.0068}$
		200	$39.51_{0.078}$	$38.84_{0.073}$	$38.83_{0.073}$	$0.95_{0.0074}$	$0.95_{0.0074}$
		1000	$36.03_{0.053}$	$35.61_{0.053}$	$35.76_{0.054}$	$0.96_{0.0069}$	$0.97_{0.0070}$

gret ratio estimates, particularly in settings 1 and 2 with small d and large n, have larger standard errors. This is caused by the fact that in these scenarios, the risks of all three classifiers are very close to the Bayes risk. In the more complex situations, the risks of the empirical classifiers are further from the Bayes risk, and the regret ratios can be estimated more precisely. The situation is similar for the bagged nearest neighbour classifier, whose relative performance also matches that predicted by the theory of Section 3 quite well.

TABLE 3

The estimated risks (multiplied by 100) of the Bayes, k-nearest neighbour, optimally weighted nearest neighbour and bagged nearest neighbour classifiers on three UCI repository data sets. Standard errors are given in small script. Recall here that K is the number of categories for the response Y

Data set	**Distance**	n	d	K	**knn risk**	**ownn risk**	**bnn risk**
Glass	L_1	163	9	2	$23.26_{0.15}$	$20.87_{0.15}$	$20.36_{0.15}$
Glass	L_2	163	9	2	$26.21_{0.15}$	$23.43_{0.14}$	$23.05_{0.14}$
Yeast	L_1	1136	8	3	$40.66_{0.059}$	$39.71_{0.062}$	$39.78_{0.063}$
Yeast	L_2	1136	8	3	$40.91_{0.057}$	$39.90_{0.058}$	$39.99_{0.059}$
Segmentation	L_1	2310	19	7	$12.04_{0.051}$	$10.05_{0.043}$	$9.882_{0.041}$
Segmentation	L_2	2310	19	7	$15.80_{0.062}$	$12.92_{0.049}$	$12.67_{0.049}$

We also applied all three classifiers to three benchmark data sets, referred to below as Glass, Yeast and Segmentation, from the UCI repository [Frank and Asuncion (2010)]. Detailed information on these data sets can be obtained from http://archive.ics.uci.edu/ml/datasets.html, but summary information is provided in Table 3. Following Athitsos and Sclaroff (2005), in each case we scaled each component of the covariates to have unit Euclidean length, and explored both the ℓ_1 and ℓ_2 norms for computing distances between observations. For the Glass and Yeast data sets, we randomly assigned each observation to a training or test set, each with probability $1/2$, while for the Segmentation data set, these probabilities were $1/11$ and $10/11$, respectively, since the original data were divided into a training and test set with these proportions. We then applied the same modified cross-validation algorithm as for the simulated data to choose the tuning parameters of the respective procedures. To estimate the risks of the three classifiers, we computed the proportion of misclassifications on the test set, and averaged these proportions over 1000 repetitions of the random assignment process.

The results are given in Table 3. In all cases, the optimally weighted nearest neighbour classifier outperforms the k-nearest neighbour classifier. Since the dimensions for the three data sets are $d = 9, 8$ and 19, it is not a surprise to see that the bagged nearest neighbour classifier also performs comparably well. The choice of distance appears to make little difference to the relative performance of the classifiers.

APPENDIX

PROOF OF THEOREM 1. The proof is rather lengthy, so we briefly outline the main ideas here. Write $P^\circ = \pi P_1 - (1-\pi)P_2$ and observe that

$$R_{\mathcal{R}}(\hat{C}_n^{\mathrm{wnn}}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})$$
$$= \int_{\mathcal{R}} \pi \bigl[\mathbb{P}\{\hat{C}_n^{\mathrm{wnn}}(x) = 2\} - \mathbb{1}_{\{C^{\mathrm{Bayes}}(x)=2\}}\bigr]\, dP_1(x) \tag{A.1}$$

$$+ \int_{\mathcal{R}} (1-\pi)\big[\mathbb{P}\{\hat{C}_n^{\mathrm{wnn}}(x)=1\} - \mathbb{1}_{\{C^{\mathrm{Bayes}}(x)=1\}}\big]\,dP_2(x)$$

$$= \int_{\mathcal{R}} \left\{\mathbb{P}\left(\sum_{i=1}^n w_{ni}\mathbb{1}_{\{Y_{(i)}=1\}} < \frac{1}{2}\right) - \mathbb{1}_{\{\eta(x)<1/2\}}\right\} dP^{\circ}(x).$$

For $\varepsilon > 0$, let

$$\mathcal{S}^{\varepsilon\varepsilon} = \{x \in \mathbb{R}^d : \eta(x) = 1/2 \text{ and } \mathrm{dist}(x, \mathcal{S}) < \varepsilon\}, \tag{A.2}$$

where $\mathrm{dist}(x, \mathcal{S}) = \inf_{x_0 \in \mathcal{S}} \|x - x_0\|$. Moreover, let

$$\mathcal{S}^{\varepsilon} = \left\{x_0 + t\frac{\dot{\eta}(x_0)}{\|\dot{\eta}(x_0)\|} : x_0 \in \mathcal{S}^{\varepsilon\varepsilon}, |t| < \varepsilon\right\}.$$

The dominant contribution to the integral in (A.1) comes from $\mathcal{R} \cap \mathcal{S}^{\varepsilon_n}$, where $\varepsilon_n = n^{-\beta/4d}$. Since the unit vector $\dot{\eta}(x_0)/\|\dot{\eta}(x_0)\|$ is orthogonal to the tangent space of $\mathcal{S}$ at x_0, we can decompose the integral over $\mathcal{R} \cap \mathcal{S}^{\varepsilon_n}$ as an integral along $\mathcal{S}$ and an integral in the perpendicular direction. We then apply a normal approximation to the integrand to deduce the result. This normal approximation requires asymptotic expansions to the mean and variance of the sum of independent random variables in (A.1), and these are developed in *step* 1 and *step* 2 below, respectively. In order to retain the flow of the main argument, we concentrate on the dominant terms in the first five steps of the argument, simply labelling the many remainder terms as $R_1, R_2, \ldots$. The sizes of these remainder terms are controlled in *step* 6 in the supplementary material [Samworth (2012)], where we also present an additional side calculation.

Step 1: Let $S_n(x) = \sum_{i=1}^n w_{ni}\mathbb{1}_{\{Y_{(i)}=1\}}$, let $\mu_n(x) = \mathbb{E}\{S_n(x)\}$, let $\varepsilon_n = n^{-\beta/4d}$ and write $t_n = n^{-2/d}\sum_{i=1}^n \alpha_i w_{ni}$. We show that

$$\sup_{x \in \mathcal{S}^{\varepsilon_n}} |\mu_n(x) - \eta(x) - a(x)t_n x| = o(t_n),$$

uniformly for $\mathbf{w}_n = (w_{ni})_{i=1}^n \in W_{n,\beta}$, where a is given in (2.2).

By a Taylor expansion,

$$\begin{aligned}\mu_n(x) &= \sum_{i=1}^n w_{ni}\mathbb{E}\{\eta(X_{(i)})\} \\ &= \eta(x) + \sum_{i=1}^{k_2} w_{ni}\mathbb{E}\{(X_{(i)} - x)^T\dot{\eta}(x)\} \\ &\quad + \frac{1}{2}\sum_{i=1}^{k_2} w_{ni}\mathbb{E}\{(X_{(i)} - x)^T\ddot{\eta}(x)(X_{(i)} - x)\} + R_1,\end{aligned} \tag{A.3}$$

where we show in *step* 6 that

$$\sup_{x \in \mathcal{S}^{\varepsilon_n}} |R_1| = o(t_n), \tag{A.4}$$

uniformly for $\mathbf{w}_n \in W_{n,\beta}$. Writing $p_t = p_t(x) = \mathbb{P}(\|X - x\| \le t)$, we also show in *step* 6 that for $x \in \mathcal{S}^{\varepsilon_n}$ and $i \le k_2$, the restriction of the distribution of $X_{(i)} - x$ to a sufficiently small ball about the origin is absolutely continuous with respect of Lebesgue measure, with Radon–Nikodym derivative given at $u = (u_1, \dots, u_d)^T$ by

$$\begin{aligned} f_{(i)}(u) &= n\bar{f}(x+u)\binom{n-1}{i-1} p_{\|u\|}^{i-1}(1 - p_{\|u\|})^{n-i} \\ &= n\bar{f}(x+u) p_{\|u\|}^{n-1}(i-1), \end{aligned} \tag{A.5}$$

say, where $p_{\|u\|}^{n}(i-1)$ denotes the probability that a $\mathrm{Bin}(n-1, p_{\|u\|})$ random variable is equal to $i-1$. Let $\delta_n = (k_2/n)^{1/2d}$. By examining the argument leading to (0.7) in the supplementary material [Samworth (2012)], we see that we can replace δ there with δ_n, to conclude that for all $M > 0$,

$$\sup_{x \in \mathcal{S}^{\varepsilon_n}} \sup_{1 \le i \le k_2} \mathbb{E}\{\|X_{(i)} - x\|^2 \mathbb{1}_{\{\|X_{(i)} - x\| > \delta_n\}}\} = O(n^{-M}).$$

It follows that

$$\begin{aligned} &\mathbb{E}\{(X_{(i)} - x)^T \dot{\eta}(x)\} \\ &\quad = \int_{\|u\| \le \delta_n} \dot{\eta}(x)^T u n\{\bar{f}(x+u) - \bar{f}(x)\} p_{\|u\|}^{n-1}(i-1)\, du + O(n^{-M}), \end{aligned} \tag{A.6}$$

uniformly for $x \in \mathcal{S}^{\varepsilon_n}$ and $1 \le i \le k_2$. Similarly,

$$\begin{aligned} &\mathbb{E}\{(X_{(i)} - x)^T \ddot{\eta}(x)(X_{(i)} - x)\} \\ &\quad = \int_{\|u\| \le \delta_n} u^T \ddot{\eta}(x) u n \bar{f}(x+u) p_{\|u\|}^{n-1}(i-1)\, du + O(n^{-M}), \end{aligned} \tag{A.7}$$

uniformly for $x \in \mathcal{S}^{\varepsilon_n}$ and $1 \le i \le k_2$. Let $k_1 = \lceil n^{\beta/4} \rceil$, and let $\Delta w_{ni} = w_{ni} - w_{n,i+1}$ with $w_{n,n+1} = 0$ (where we introduce the comma here for clarity). By a Taylor expansion, we have

$$\begin{aligned} &\sum_{i=1}^{k_2} w_{ni} \int_{\|u\| \le \delta_n} \Big[\dot{\eta}(x)^T u n\{\bar{f}(x+u) - \bar{f}(x)\} \\ &\qquad\qquad + \frac{1}{2} u^T \ddot{\eta}(x) u n \bar{f}(x+u)\Big] p_{\|u\|}^{n-1}(i-1)\, du \\ &\quad = \{1 + o(1)\} \sum_{i=k_1}^{k_2} n\Delta w_{ni} \sum_{j=1}^{d} \int_{\|u\| \le \delta_n} \Big\{\eta_j(x) u_j^2 \bar{f}_j(x) \\ &\qquad\qquad + \frac{1}{2}\eta_{jj}(x) u_j^2 \bar{f}(x)\Big\} q_{\|u\|}^{n-1}(i)\, du, \end{aligned} \tag{A.8}$$

uniformly for $x \in \mathcal{S}^{\varepsilon_n}$ and $\mathbf{w}_n \in W_{n,\beta}$, where $q_{\|u\|}^{n-1}(i)$ denotes the probability that a $\mathrm{Bin}(n-1, p_{\|u\|})$ random variable is less than i. Now, $q_{\|u\|}^{n-1}(i)$ is decreasing in

$\|u\|$ and is close to 1 when $\|u\|$ is small and close to zero when $\|u\|$ is large. To analyse this more precisely, note that $p_{\|u\|} = \bar{f}(x)a_d\|u\|^d\{1 + O(\|u\|^2)\}$ as $u \to 0$, uniformly for $x \in \mathcal{S}^{\epsilon_n}$, so it is convenient to let $b_n = (\frac{(n-1)a_d\bar{f}(x)}{i})^{1/d}$ and set $v = b_n u$. Then there exists n_0 such that for $n \geq n_0$, we have for all $x \in \mathcal{S}^{\epsilon_n}$, all $\|v\|^d \in (0, 1 - 2/\log n]$ and all $k_1 \leq i \leq k_2$ that

$$i - (n-1)p_{\|v\|/b_n} \geq \frac{i}{\log n}.$$

Thus by Bernstein's inequality [Shorack and Wellner (1986), page 440], for each $M > 0$ and for $n \geq n_0$,

$$\sup_{\|v\|^d \in (0,1-2/\log n]} \sup_{k_1 \leq i \leq k_2} \{1 - q^{n-1}_{\|v\|/b_n}(i)\} \leq \exp\biggl(-\frac{k_1}{3\log^2 n}\biggr) = O(n^{-M}). \tag{A.9}$$

Similarly, for $n \geq n_0$,

$$\sup_{\|v\|^d \in [1+2/\log n, b_n\delta_n]} \sup_{k_1 \leq i \leq k_2} q^{n-1}_{\|v\|/b_n}(i) \leq \exp\biggl(-\frac{k_1}{3\log^2 n}\biggr) = O(n^{-M}). \tag{A.10}$$

We deduce from (A.6)–(A.9) and (A.10) that

$$\begin{aligned}
&\sum_{i=1}^{k_2} w_{ni}\mathbb{E}\{(X_{(i)} - x)^T \dot{\eta}(x)\} \\
&\quad + \frac{1}{2}\sum_{i=1}^{k_2} w_{ni}\mathbb{E}\{(X_{(i)} - x)^T \ddot{\eta}(x)(X_{(i)} - x)\} \\
&= \{1 + o(1)\} \\
&\quad \times \sum_{i=1}^{n} \frac{n\Delta w_{ni}}{b_n^{d+2}} \sum_{j=1}^{d} \biggl\{\eta_j(x)\bar{f}_j(x) + \frac{1}{2}\eta_{jj}(x)\bar{f}(x)\biggr\} \int_{\|v\|\leq 1} v_j^2\,dv \qquad \text{(A.11)} \\
&= a(x)t_n + o(t_n),
\end{aligned}$$

uniformly for $x \in \mathcal{S}^{\epsilon_n}$ and $\mathbf{w}_n \in W_{n,\beta}$. Combining (A.3), (A.4) and (A.11), this completes *step* 1.

Step 2: Let $\sigma_n^2(x) = \operatorname{Var}\{S_n(x)\}$ and let $s_n^2 = \sum_{i=1}^n w_{ni}^2$. We claim that

$$\sup_{x \in \mathcal{S}^{\epsilon_n}} \biggl|\sigma_n^2(x) - \frac{1}{4}s_n^2\biggr| = o(s_n^2),$$

uniformly for $\mathbf{w}_n \in W_{n,\beta}$. To see this, note that

$$\begin{aligned}
\sigma_n^2(x) &= \sum_{i=1}^n w_{ni}^2 \mathbb{E}[\eta(X_{(i)})\{1 - \eta(X_{(i)})\}] + \sum_{i=1}^n w_{ni}^2 \operatorname{Var}\eta(X_{(i)}) \\
&= \sum_{i=1}^n w_{ni}^2 [\mathbb{E}\eta(X_{(i)}) - \{\mathbb{E}\eta(X_{(i)})\}^2].
\end{aligned}$$

But by a simplified version of the argument in *step* 1, we have

$$\sup_{x\in\mathcal{S}^{\varepsilon_n}}\sup_{1\le i\le k_2}\left|\mathbb{E}\eta(X_{(i)})-\eta(x)\right|\to 0.$$

It follows that

$$\begin{aligned}\sup_{x\in\mathcal{S}^{\varepsilon_n}}&\left|\sum_{i=1}^n w_{ni}^2\mathbb{E}\eta(X_{(i)})-\frac{1}{2}s_n^2\right|\\ &\le\sup_{x\in\mathcal{S}^{\varepsilon_n}}\sum_{i=1}^{k_2}w_{ni}^2\left|\mathbb{E}\eta(X_{(i)})-\eta(x)\right|\\ &\quad+\sum_{i=k_2+1}^n w_{ni}^2+s_n^2\sup_{x\in\mathcal{S}^{\varepsilon_n}}\left|\eta(x)-1/2\right|\\ &=o(s_n^2),\end{aligned}$$

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

$$\begin{aligned}&\left|\sum_{i=1}^n w_{ni}^2\{\mathbb{E}\eta(X_{(i)})\}^2-\frac{1}{4}s_n^2\right|\\ &\quad\le\sum_{i=1}^{k_2}w_{ni}^2\left|\mathbb{E}\eta(X_{(i)})-\eta(x)\right|\left|\mathbb{E}\eta(X_{(i)})+\eta(x)\right|\\ &\qquad+2\sum_{i=k_2+1}^n w_{ni}^2+s_n^2\left|\eta(x)^2-1/4\right|\\ &\quad=o(s_n^2),\end{aligned}$$

uniformly for $x\in\mathcal{S}^{\varepsilon_n}$ and $\mathbf{w}_n\in W_{n,\beta}$. This completes *step* 2.

Step 3: For $x_0\in\mathcal{S}$ and $t\in\mathbb{R}$, we write $x_0^t=x_0+t\dot\eta(x_0)/\|\dot\eta(x_0)\|$ for brevity. Moreover, we write $\psi=\pi f_1-(1-\pi)f_2$ for the Radon–Nikodym derivative with respect to Lebesgue measure of the restriction of P° to $\mathcal{S}^{\varepsilon_n}$ for large n. We show that

$$\begin{aligned}&\int_{\mathcal{R}\cap\mathcal{S}^{\varepsilon_n}}\left[\mathbb{P}\{S_n(x)<1/2\}-\mathbb{1}_{\{\eta(x)<1/2\}}\right]dP^\circ(x)\\ &\quad=\int_{\mathcal{S}}\int_{-\varepsilon_n}^{\varepsilon_n}\psi(x_0^t)\left[\mathbb{P}\{S_n(x_0^t)<1/2\}\right.\\ &\qquad\qquad\left.-\mathbb{1}_{\{t<0\}}\right]dt\,d\mathrm{Vol}^{d-1}(x_0)\{1+o(1)\},\end{aligned}\tag{A.12}$$

uniformly for $\mathbf{w}_n\in W_{n,\beta}$. Recalling the definition of $\mathcal{S}^{\varepsilon_n\varepsilon_n}$ in (A.2), note that for large n, the map

$$\phi\left(x_0,t\frac{\dot\eta(x_0)}{\|\dot\eta(x_0)\|}\right)=x_0^t$$

is a diffeomorphism from $\{(x_0, t\dot{\eta}(x_0)/\|\dot{\eta}(x_0)\|) : x_0 \in \mathcal{S}^{\varepsilon_n\varepsilon_n}, |t| < \varepsilon_n\}$ onto $\mathcal{S}^{\varepsilon_n}$ [Gray (2004), pages 32–33]. Observe that

$$\{x \in \mathbb{R}^d : \mathrm{dist}(x, \mathcal{S}) < \varepsilon_n\} \subseteq \mathcal{S}^{\varepsilon_n} \subseteq \{x \in \mathbb{R}^d : \mathrm{dist}(x, \mathcal{S}) < 2\varepsilon_n\}. \tag{A.13}$$

Moreover, for large n and $|t| < \varepsilon_n$, we have $\mathrm{sgn}\{\eta(x_0^t) - 1/2\} = \mathrm{sgn}\{\psi(x_0^t)\} = \mathrm{sgn}(t)$. The pullback of the d-form dx is given at $(x_0, t\dot{\eta}(x_0)/\|\dot{\eta}(x_0)\|)$ by

$$\det \dot{\phi}\left(x_0, t\frac{\dot{\eta}(x_0)}{\|\dot{\eta}(x_0)\|}\right) dt\, d\mathrm{Vol}^{d-1}(x_0) = \{1 + o(1)\}\, dt\, d\mathrm{Vol}^{d-1}(x_0),$$

where the error term is uniform in $(x_0, t\dot{\eta}(x_0)/\|\dot{\eta}(x_0)\|)$ for $x_0 \in \mathcal{S}$ and $|t| < \varepsilon_n$. It follows from the theory of integration on manifolds, as described in Guillemin and Pollack (1974), page 168 and Gray (2004), Theorems 3.15 and 4.7 [see also Moore (1992)], that

$$\begin{aligned} &\int_{\mathcal{S}^{\varepsilon_n}} [\mathbb{P}\{S_n(x) < 1/2\} - \mathbb{1}_{\{\eta(x)<1/2\}}]\, dP^\circ(x) \\ &\quad = \int_{\mathcal{S}^{\varepsilon_n\varepsilon_n}} \int_{-\varepsilon_n}^{\varepsilon_n} \psi(x_0^t)[\mathbb{P}\{S_n(x_0^t) < 1/2\} \\ &\qquad - \mathbb{1}_{\{t<0\}}]\, dt\, d\mathrm{Vol}^{d-1}(x_0)\{1 + o(1)\}, \end{aligned} \tag{A.14}$$

uniformly for $\mathbf{w}_n \in W_{n,\beta}$. But $\mathcal{S}^{\varepsilon_n} \setminus \mathcal{R} \subseteq \{x \in \mathbb{R}^d : \mathrm{dist}(x, \partial\mathcal{S}) < \varepsilon_n\}$, and this latter set has volume $O(\varepsilon_n^2)$ by Weyl's tube formula [Gray (2004), Theorem 4.8]. Thus the integral over $\mathcal{S}^{\varepsilon_n}$ in (A.14) may be replaced with an integral over $\mathcal{R} \cap \mathcal{S}^{\varepsilon_n}$ and, similarly, the integral over $\mathcal{S}^{\varepsilon_n\varepsilon_n}$ may be replaced with an integral over $\mathcal{S}$, without changing the order of the error term in (A.14). Thus (A.12) holds, and this completes *step* 3.

Step 4: We now return to the main argument to bound the contribution to the risk (A.1) from $\mathcal{R} \setminus \mathcal{S}^{\varepsilon_n}$. In particular, we show that

$$\sup_{\mathbf{w}_n \in W_{n,\beta}} \int_{\mathcal{R}\setminus\mathcal{S}^{\varepsilon_n}} [\mathbb{P}\{S_n(x) < 1/2\} - \mathbb{1}_{\{\eta(x)<1/2\}}]\, dP^\circ(x) = O(n^{-M}) \tag{A.15}$$

for all $M > 0$. To see this, recall that $|\eta(x) - 1/2|$ is assumed to be bounded away from zero on the set $\mathcal{R} \setminus \mathcal{S}^\varepsilon$ (for fixed $\varepsilon > 0$), and $\|\dot{\eta}(x_0)\|$ is bounded away from zero for $x_0 \in \mathcal{S}$. Hence, by (A.13) in *step* 3, there exists $c_1 > 0$ such that, for sufficiently small $\varepsilon > 0$,

$$\inf_{x\in\mathcal{R}\setminus\mathcal{S}^\varepsilon} |\eta(x) - 1/2| \geq c_1\varepsilon. \tag{A.16}$$

We also claim that $\mu_n(x) = \mathbb{E}\{S_n(x)\}$ is similarly bounded away from 1/2 uniformly for $x \in \mathcal{R} \setminus \mathcal{S}^{\varepsilon_n}$. In fact, we have by Hoeffding's inequality that

$$\mathbb{P}(\|X_{(k_2)} - x\| > \varepsilon_n/2) = q_{\varepsilon_n/2}^n(k_2) \leq e^{-(2/n)(np_{\varepsilon_n/2}-k_2)^2}.$$

It follows that

$$\sup_{\substack{x\in\mathcal{R}\setminus\mathcal{S}^{\varepsilon_n}:\\ \eta(x)\le 1/2}} \mu_n(x)-\frac{1}{2}$$

$$\text{(A.17)}\qquad \le \sup_{\substack{x\in\mathcal{R}\setminus\mathcal{S}^{\varepsilon_n}:\\ \eta(x)\le 1/2}} \Biggl\{\sum_{i=1}^{k_2} w_{ni}\mathbb{P}\bigl(Y_{(i)}=1\cap\|X_{(k_2)}-x\|\le\varepsilon_n/2\bigr) - \frac{1}{2}+e^{-(2/n)(np_{\varepsilon_n/2}-k_2)^2}+\sum_{i=k_2+1}^{n} w_{ni}\Biggr\}$$

$$\text{(A.18)}\qquad \le \sum_{i=1}^{k_2} w_{ni}\biggl(\frac{1}{2}-\frac{c_1\varepsilon_n}{2}\biggr)-\frac{1}{2}+e^{-(2/n)(np_{\varepsilon_n/2}-k_2)^2}+n^{-\beta}\le -\frac{c_1\varepsilon_n}{4}$$

for sufficiently large n. Similarly,

$$\inf_{\substack{x\in\mathcal{R}\setminus\mathcal{S}^{\varepsilon_n}:\\ \eta(x)\ge 1/2}} \mu_n(x)-\frac{1}{2}$$

$$\text{(A.19)}\qquad \ge \inf_{\substack{x\in\mathcal{R}\setminus\mathcal{S}^{\varepsilon_n}:\\ \eta(x)\ge 1/2}} \sum_{i=1}^{k_2} w_{ni}\mathbb{P}\bigl(Y_{(i)}=1\cap\|X_{(k_2)}-x\|\le\varepsilon_n/2\bigr)-\frac{1}{2}$$

$$\ge \bigl(1-n^{-\beta/2}\bigr)\biggl(\frac{1}{2}+\frac{c_1\varepsilon_n}{2}\biggr)\bigl(1-e^{-(2/n)(np_{\varepsilon_n/2}-k_2)^2}\bigr)-\frac{1}{2}\ge\frac{c_1\varepsilon_n}{4}$$

for large n.

Now we may apply Hoeffding's inequality again, this time to $S_n(x)$, to deduce that

$$\bigl|\mathbb{P}\{S_n(x)<1/2\}-\mathbb{1}_{\{\eta(x)<1/2\}}\bigr|\le e^{(-2(\mu_n(x)-1/2)^2)/s_n^2}=O\bigl(n^{-M}\bigr)$$

for each $M>0$, uniformly for $\mathbf{w}_n\in W_{n,\beta}$ and $x\in\mathcal{R}\setminus\mathcal{S}^{\varepsilon_n}$, using (A.17) and (A.19) and the fact that $s_n^2\le n^{-\beta}$ for $\mathbf{w}_n\in W_{n,\beta}$. This completes *step* 4.

Step 5. We now show that

$$\int_{\mathcal{S}}\int_{-\varepsilon_n}^{\varepsilon_n}\psi\bigl(x_0^t\bigr)\bigl[\mathbb{P}\bigl\{S_n\bigl(x_0^t\bigr)<1/2\bigr\}-\mathbb{1}_{\{t<0\}}\bigr]\,dt\,d\mathrm{Vol}^{d-1}(x_0)$$
$$=B_1s_n^2+B_2t_n^2+o\bigl(s_n^2+t_n^2\bigr),$$

uniformly for $\mathbf{w}_n\in W_{n,\beta}$, where B_1 and B_2 were defined in (2.3). When combined with (A.1) and the results of *step* 3 and *step* 4 [in particular, (A.12) and (A.15)], this will complete the proof of Theorem 1.

First observe that

$$
\begin{aligned}
&\int_{\mathcal{S}} \int_{-\varepsilon_n}^{\varepsilon_n} \psi(x_0^t) [\mathbb{P}\{S_n(x_0^t) < 1/2\} - \mathbb{1}_{\{t<0\}}] \, dt \, d\mathrm{Vol}^{d-1}(x_0) \\
&\quad = \int_{\mathcal{S}} \int_{-\varepsilon_n}^{\varepsilon_n} t \|\dot{\psi}(x_0)\| [\mathbb{P}\{S_n(x_0^t) < 1/2\} \\
&\qquad\qquad - \mathbb{1}_{\{t<0\}}] \, dt \, d\mathrm{Vol}^{d-1}(x_0) \{1 + o(1)\}.
\end{aligned} \tag{A.20}
$$

Now, $S_n(x)$ is a sum of independent, bounded random variables, so by the nonuniform version of the Berry–Esseen theorem, there exists $C_1 > 0$ such that for all $y \in \mathbb{R}$,

$$
\begin{aligned}
\sup_{x_0 \in \mathcal{S}} \sup_{t \in [-\varepsilon_n, \varepsilon_n]} &\left| \mathbb{P}\left(\frac{S_n(x_0^t) - \mu_n(x_0^t)}{\sigma_n(x_0^t)} \le y \right) - \Phi(y) \right| \\
&\le \frac{C_1}{n^{1/2}(1 + |y|^3)},
\end{aligned}
$$

where Φ denotes the standard normal distribution function. Thus

$$
\begin{aligned}
&\int_{\mathcal{S}} \int_{-\varepsilon_n}^{\varepsilon_n} t \|\dot{\psi}(x_0)\| \{\mathbb{P}\{S_n(x_0^t) < 1/2\} - \mathbb{1}_{\{t<0\}}\} \, dt \, d\mathrm{Vol}^{d-1}(x_0) \\
&\quad = \int_{\mathcal{S}} \int_{-\varepsilon_n}^{\varepsilon_n} t \|\dot{\psi}(x_0)\| \left\{ \Phi\left(\frac{1/2 - \mu_n(x_0^t)}{\sigma_n(x_0^t)} \right) \right. \\
&\qquad\qquad \left. - \mathbb{1}_{\{t<0\}} \right\} dt \, d\mathrm{Vol}^{d-1}(x_0) + R_2,
\end{aligned}
$$

where we show in *step* 6 that

$$
|R_2| = o(s_n^2 + t_n^2), \tag{A.21}
$$

uniformly for $\mathbf{w}_n \in W_{n,\beta}$. Moreover, by a Taylor expansion and *step* 1 and *step* 2,

$$
\begin{aligned}
&\int_{\mathcal{S}} \int_{-\varepsilon_n}^{\varepsilon_n} t \|\dot{\psi}(x_0)\| \left\{ \Phi\left(\frac{1/2 - \mu_n(x_0^t)}{\sigma_n(x_0^t)} \right) - \mathbb{1}_{\{t<0\}} \right\} dt \, d\mathrm{Vol}^{d-1}(x_0) \\
&\quad = \int_{\mathcal{S}} \int_{-\varepsilon_n}^{\varepsilon_n} t \|\dot{\psi}(x_0)\| \left\{ \Phi\left(\frac{-2t\|\dot{\eta}(x_0)\| - 2a(x_0)t_n}{s_n} \right) \right. \\
&\qquad\qquad \left. - \mathbb{1}_{\{t<0\}} \right\} dt \, d\mathrm{Vol}^{d-1}(x_0) + R_3,
\end{aligned}
$$

where we show in *step* 6 that

$$
|R_3| = o(s_n^2 + t_n^2), \tag{A.22}
$$

uniformly for $\mathbf{w}_n \in W_{n,\beta}$. Finally, we can make the substitution $r = t/s_n$ to conclude that

$$\begin{aligned}
&\int_{\mathcal{S}} \int_{-\varepsilon_n}^{\varepsilon_n} t \|\dot{\psi}(x_0)\| \left\{ \Phi\left(\frac{-2t\|\dot{\eta}(x_0)\| - 2a(x_0)t_n}{s_n} \right) - \mathbb{1}_{\{t<0\}} \right\} dt \, d\mathrm{Vol}^{d-1}(x_0) \\
&\quad = \frac{s_n^2}{4} \int_{\mathcal{S}} \int_{-\infty}^{\infty} u \|\dot{\psi}(x_0)\| \left\{ \Phi\left(-u\|\dot{\eta}(x_0)\| - \frac{2t_n a(x_0)}{s_n} \right) \right. \\
&\qquad\qquad\qquad\qquad\qquad \left. - \mathbb{1}_{\{u<0\}} \right\} du \, d\mathrm{Vol}^{d-1}(x_0) + R_4 \\
&\quad = B_1 s_n^2 + B_2 t_n^2 + R_4,
\end{aligned}$$

where B_1 and B_2 were defined in (2.3). Here, we have used the fact that $\|\dot{\psi}(x_0)\|/\|\dot{\eta}(x_0)\| = 2\bar{f}(x_0)$ for $x_0 \in \mathcal{S}$ in the final step of this calculation. Once we have shown in *step* 6 that

$$|R_4| = o(s_n^2), \tag{A.23}$$

uniformly for $\mathbf{w}_n \in W_{n,\beta}$, this will complete *Step* 5 and hence the proof of Theorem 1. □

Acknowledgements. I am very grateful to Peter Hall for introducing me to this topic, and to the anonymous referees and Associate Editor for their constructive comments, which significantly helped to improve the paper. I thank the Statistical and Applied Mathematical Sciences Institute (SAMSI) in North Carolina for kind hospitality when I attended the "Program on Analysis of Object Data" from September to November 2010, during which time part of this research was carried out.

SUPPLEMENTARY MATERIAL

Supplement to "Optimal weighted nearest neighbour classifiers" (DOI: 10.1214/12-AOS1049SUPP; .pdf). We complete the proof of Theorem 1, and give the proofs of the other results in the paper. We also discuss minimax properties of weighted nearest neighbour classifiers and a plug-in approach to estimating k^*.

REFERENCES

ATHITSOS, V. and SCLAROFF, S. (2005). Boosting nearest neighbour classifiers for multiclass recognition. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 45–55. IEEE Computer Society, Washington, DC.

AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** 608–633. MR2336861

BAILEY, T. and JAIN, A. (1978). A note on distance-weighted k-nearest neighbour rules. *Transactions on Systems, Man, and Cybernetics* **8** 311–313.

BIAU, G., CÉROU, F. and GUYADER, A. (2010). On the rate of convergence of the bagged nearest neighbour estimate. *J. Mach. Learn. Res.* **11** 687–712. MR2600626

BIAU, G. and DEVROYE, L. (2010). On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *J. Multivariate Anal.* **101** 2499–2518. MR2719877

BLANCHARD, G., BOUSQUET, O. and MASSART, P. (2008). Statistical performance of support vector machines. *Ann. Statist.* **36** 489–531. MR2396805

BOUCHERON, S., BOUSQUET, O. and LUGOSI, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM Probab. Stat.* **9** 323–375. MR2182250

BREIMAN, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24** 2350–2383. MR1425957

BREIMAN, L. (1999). Using adaptive bagging to debias regressions. Technical report, Dept. Statistics, Univ. California, Berkeley.

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

CHANDA, K. C. and RUYMGAART, F. H. (1989). Asymptotic estimate of probability of misclassification for discriminant rules based on density estimates. *Statist. Probab. Lett.* **8** 81–88. MR1006427

CORTES, C. and VAPNIK, V. (1995). Support-vector networks. *Mach. Learn.* **20** 273–297.

COVER, T. M. and HART, P. E. (1967). Nearest neighbour pattern classification. *IEEE Trans. Inform. Theory* **13** 21–27.

DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition. Applications of Mathematics* (*New York*) **31**. Springer, New York. MR1383093

FIX, E. and HODGES, J. L. (1951). Discriminatory analysis—Nonparametric discrimination: Consistency properties. Technical Report 4, Project no. 21-29-004, USAF School of Aviation Medicine, Randolph Field, Texas.

FIX, E. and HODGES, J. L. (1989). Discriminatory analysis—Nonparametric discrimination: Consistency properties. *Internat. Statist. Rev.* **57** 238–247.

FRANK, A. and ASUNCION, A. (2010). UCI machine learning repository. School of Information and Computer Sciences, Univ. California, Irvine, CA. Available at http://archive.ics.uci.edu/ml.

GORDON, A. D. (1999). *Classification*, 2nd ed. Chapman & Hall, London. MR0637465

GRAY, A. (2004). *Tubes*, 2nd ed. *Progress in Mathematics* **221**. Birkhäuser, Basel. MR2024928

GUILLEMIN, V. and POLLACK, A. (1974). *Differential Topology*. Prentice-Hall, Englewood Cliffs, NJ. MR0348781

HALL, P. and KANG, K.-H. (2005). Bandwidth choice for nonparametric classification. *Ann. Statist.* **33** 284–306. MR2157804

HALL, P., PARK, B. U. and SAMWORTH, R. J. (2008). Choice of neighbour order in nearest-neighbour classification. *Ann. Statist.* **36** 2135–2152. MR2458182

HALL, P. and SAMWORTH, R. J. (2005). Properties of bagged nearest neighbour classifiers. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 363–379. MR2155343

HAND, D. J. (1981). *Discrimination and Classification*. Wiley, Chichester. MR0634676

IBRAGIMOV, I. A. and KHASMINSKIĬ, R. Z. (1980). Nonparametric regression estimation. *Dokl. Akad. Nauk SSSR* **252** 780–784. MR0580829

IBRAGIMOV, I. A. and KHASMINSKIĬ, R. Z. (1981). *Statistical Estimation. Applications of Mathematics* **16**. Springer, New York. MR0620321

IBRAGIMOV, I. A. and KHASMINSKIĬ, R. Z. (1982). On the bounds for quality of nonparametric regression function estimation. *Theory Probab. Appl.* **27** 81–94.

LEPSKIĬ, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** 682–697.

MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. MR1765618

MARTÍNEZ-MUÑOZ, G. and SUÁREZ, A. (2010). Out-of-bag estimation of the optimal sample size in bagging. *Pattern Recognition* **43** 143–152.

MOORE, J. D. (1992). Book review: Tubes. *Bull. Amer. Math. Soc.* (*N.S.*) **27** 311–313. MR1567997

POLONIK, W. (1995). Measuring mass concentrations and estimating density contour clusters—An excess mass approach. *Ann. Statist.* **23** 855–881. MR1345204

RAUDYS, Š. and YOUNG, D. M. (2004). Results in statistical discriminant analysis: A review of the former Soviet Union literature. *J. Multivariate Anal.* **89** 1–35. MR2041207

RIGOLLET, P. and VERT, R. (2009). Optimal rates for plug-in estimators of density level sets. *Bernoulli* **15** 1154–1178. MR2597587

ROYALL, R. (1966). A class of nonparametric estimators of a smooth regression function. Ph.D. thesis, Stanford Univ., Stanford, CA.

SAMWORTH, R. J. (2012). Supplement to "Optimal weighted nearest neighbour classifiers." DOI:10.1214/12-AOS1049SUPP.

SAMWORTH, R. J. and WAND, M. P. (2010). Asymptotics and optimal bandwidth selection for highest density region estimation. *Ann. Statist.* **38** 1767–1792. MR2662359

SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York. MR0838963

STEELE, B. M. (2009). Exact bootstrap k-nearest neighbour learners. *Mach. Learn.* **74** 235–255.

STEINWART, I. and CHRISTMANN, A. (2008). *Support Vector Machines*. Springer, New York. MR2450103

STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645. MR0443204

TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002

STATISTICAL LABORATORY
UNIVERSITY OF CAMBRIDGE
WILBERFORCE ROAD
CAMBRIDGE
CB3 0WB
UNITED KINGDOM
E-MAIL: r.samworth@statslab.cam.ac.uk
URL: http://www.statslab.cam.ac.uk/~rjs57