The Annals of Statistics
2020, Vol. 48, No. 3, 1789–1814
https://doi.org/10.1214/19-AOS1868

LOCAL NEAREST NEIGHBOUR CLASSIFICATION WITH APPLICATIONS TO SEMI-SUPERVISED LEARNING

BY TIMOTHY I. CANNINGS[1], THOMAS B. BERRETT[2,*] AND RICHARD J. SAMWORTH[2,**]

[1]*School of Mathematics, University of Edinburgh, timothy.cannings@ed.ac.uk*
[2]*Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, *t.berrett@statslab.cam.ac.uk; **r.samworth@statslab.cam.ac.uk*

We derive a new asymptotic expansion for the global excess risk of a local-k-nearest neighbour classifier, where the choice of k may depend upon the test point. This expansion elucidates conditions under which the dominant contribution to the excess risk comes from the decision boundary of the optimal Bayes classifier, but we also show that if these conditions are not satisfied, then the dominant contribution may arise from the tails of the marginal distribution of the features. Moreover, we prove that, provided the d-dimensional marginal distribution of the features has a finite ρth moment for some $\rho > 4$ (as well as other regularity conditions), a local choice of k can yield a rate of convergence of the excess risk of $O(n^{-4/(d+4)})$, where n is the sample size, whereas for the standard k-nearest neighbour classifier, our theory would require $d \geq 5$ and $\rho > 4d/(d-4)$ finite moments to achieve this rate. These results motivate a new k-nearest neighbour classifier for semi-supervised learning problems, where the unlabelled data are used to obtain an estimate of the marginal feature density, and fewer neighbours are used for classification when this density estimate is small. Our worst-case rates are complemented by a minimax lower bound, which reveals that the local, semi-supervised k-nearest neighbour classifier attains the minimax optimal rate over our classes for the excess risk, up to a subpolynomial factor in n. These theoretical improvements over the standard k-nearest neighbour classifier are also illustrated through a simulation study.

1. Introduction. Supervised classification problems represent some of the most frequently-occurring statistical challenges in a wide variety of fields, including fraud detection, medical diagnoses and targeted advertising, to name just a few. The area has received an enormous amount of attention within both the statistics and machine learning communities; for an excellent survey with pointers to much of the relevant literature, see Boucheron, Bousquet and Lugosi (2005).

The k-nearest neighbour classifier, which assigns the test point according to a majority vote over the classes of its k nearest points in the training set, was introduced in the seminal work of Fix and Hodges (1951) (later republished as Fix and Hodges (1989)), and is arguably the simplest and most intuitive nonparametric classifier. Cover and Hart (1967) provided mild conditions under which the asymptotic risk of the 1-nearest neighbour classifier is bounded above by twice the risk of the optimal Bayes classifier. Stone (1977) proved that if $k = k_n$ is chosen such that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$, then the k-nearest neighbour classifier is universally consistent, in the sense that under any data generating mechanism, its risk converges to the Bayes risk. Further recent contributions, some of which treat the k-nearest neighbour classifier as a special case of a plug-in classifier, include Kulkarni and Posner

Received August 2018; revised May 2019.
MSC2010 subject classifications. 62G20.
Key words and phrases. Classification problems, nearest neighbours, nonparametric classification, semi-supervised learning.

(1995), Audibert and Tsybakov (2007), Hall, Park and Samworth (2008), Biau, Cérou and Guyader (2010), Samworth (2012), Chaudhuri and Dasgupta (2014) and Celisse and Mary-Huard (2018). Nearest neighbour methods have also been extensively used in other statistical problems, including density estimation (Loftsgaarden and Quesenberry (1965), Mack (1983), Mack and Rosenblatt (1979)), nonparametric clustering (Heckel and Bölcskei (2015)), entropy and other functional estimation (Berrett and Samworth (2019a), Berrett, Samworth and Yuan (2019), Kozachenko and Leonenko (1987)) and testing problems (Berrett and Samworth (2019b), Schilling (1986)); see also the recent book Biau and Devroye (2015).

Despite these aforementioned works, the behaviour of the k-nearest neighbour classifier in the tails of a distribution remains poorly understood. Indeed, writing (X, Y) for a generic data pair, where the d-dimensional feature vector X has marginal density $\bar{f}$ and Y denotes a binary class label, most of the results in the papers mentioned in the previous paragraph pertain either to situations where $\bar{f}$ is compactly supported and bounded away from zero on its support, or where the excess risk over that of the Bayes classifier is computed only over a compact subset of $\mathbb{R}^d$. As such, many questions remain regarding the effect of tail behaviour on the excess risk.

In this paper, we consider classes of distributions that allow the feature vectors to have unbounded support. Our first goal is to provide a new asymptotic expansion for the global excess risk of a k-nearest neighbour classifier, whose error term can be bounded uniformly over our classes (Theorem 1). This expansion elucidates conditions under which the dominant contribution to the excess risk comes from the decision boundary of the Bayes classifier, but we also show that if these conditions are not satisfied, then the dominant contribution may arise from the tails of the marginal distribution of the features. The threshold for these two different regimes is governed by a parameter ρ that controls the number of finite moments of the marginal feature distribution: if $d \geq 5$ and $\rho > 4d/(d-4)$, then we obtain a rate of $O(n^{-4/(d+4)})$ uniformly over our classes, while if $d \leq 4$ or $d \geq 5$ and $\rho \leq 4d/(d-4)$ then our rate is slower, namely $O(n^{-\frac{\rho}{2\rho+d}+\epsilon})$, for every $\epsilon > 0$.

The proof of Theorem 1 also reveals a local bias-variance trade-off that motivates a modification of the standard k-nearest neighbour classifier in semi-supervised learning settings, where, as well as the labelled training data, we have access to another, independent, sample of unlabelled observations. Such semi-supervised problems occur in a wide range of applications, especially where it is expensive or time-consuming to obtain the labels associated with observations; in fact, it is often the case that unlabelled observations may vastly outnumber labelled ones. For an overview of semi-supervised learning applications and techniques, see Chapelle, Zien and Schölkopf (2006).

Our second contribution is to propose to allow the choice of k in k-nearest neighbour classification to depend on an estimate of $\bar{f}$ at the test point $x \in \mathbb{R}^d$ in semi-supervised settings. Such a local choice of k is analagous to the use of local bandwidths in the context of kernel density estimation, as studied by, for example, Breiman, Meisel and Purcell (1977), Abramson (1982) and Giné and Sang (2010). However, for density estimation, it is more common to choose a family of bandwidths $\{h(X_i) : i = 1, \ldots, n\}$ rather than $h = h(x)$, to ensure that the resulting estimate is itself a density. Moreover, theory there suggests that one should then choose $h(X_i) \propto \bar{f}^{-1/2}(X_i)$ in order to cancel the leading term in the asymptotic bias expansion (Abramson (1982)). By contrast, we find that when choosing $k = k(x)$, by using *fewer* neighbours in low density regions, we are able to achieve a better balance in the local bias-variance trade-off for estimating our main quantity of interest, namely the regression function. In particular, we initially study an oracle choice of $k = k(x)$ that depends on $\bar{f}(x)$, and show that the excess risk of the resulting classifier, computed over the whole of $\mathbb{R}^d$, is $O(n^{-4/(d+4)})$, again uniformly over our classes, for every $d \in \mathbb{N}$ and provided only that $\rho > 4$. Moreover, in the more challenging case where $\rho \leq 4$, we obtain a

rate of $O(n^{-\frac{\rho}{\rho+d}+\epsilon})$, for every $\epsilon > 0$, which still reflects an improvement through the locally-adaptive choice of k. Assuming further that $\bar{f}$ has Hölder smoothness $\gamma \in (0, 2]$, we show that if m additional, unlabelled observations are used to estimate $\bar{f}$ by $\hat{f}_m$, and if $m = m_n$ satisfies $\liminf_{n\to\infty} m_n/n^{2+d/\gamma} > 0$, then our semi-supervised k-nearest-neighbour classifier mimics the asymptotic performance of the oracle.

Finally, we consider corresponding minimax lower bounds. We show in particular that the rates of convergence achieved by our semi-supervised, local-k-nearest neighbour classifier are optimal up to subpolynomial factors in n. Interestingly, our arguments also reveal that these rates cannot be improved with the additional knowledge of $\bar{f}$.

As mentioned previously, studies of global excess risk rates of convergence in nonparametric classification for unbounded feature vector distributions are comparatively rare. Hall and Kang (2005) studied the tail error properties of a classifier based on kernel density estimates of the class conditional densities for univariate data. As an illustrative example, they showed that if, for large x, one class has density $ax^{-\alpha}$, while the other has density $bx^{-\beta}$, for some $a, b > 0$ and $1 < \alpha < \beta < \alpha + 1 < \infty$, then the excess risk from the right tail is of larger order than that in the body of the distribution.

Perhaps most closely related to this work, Gadat, Klein and Marteau (2016) recently obtained upper bounds on the supremum excess risk of the k-nearest neighbour classifier, over classes where η is Lipschitz, the well-known margin assumption of Mammen and Tsybakov (1999) is satisfied with parameter $\alpha > 0$, and assuming the tail condition that $\mathbb{P}\{\bar{f}(X) < \delta\} \leq \psi(\delta)$ is satisfied for some function ψ and sufficiently small $\delta > 0$. Gadat, Klein and Marteau (2016) obtained a minimax lower bound over these classes, as well as providing an upper bound for the rate of the standard k-nearest neighbour classifier. Since these rates do not match, they further introduced regions of the form $\{\bar{f}^{-1}((a_{j+1}, a_j]) : j \in \mathbb{N}\}$ with $a_{j+1} = a_j/2$, and proved that when we choose $k = k(j)$ and specialise to the case where ψ is the identity function, the resulting sliced k-nearest neighbour classifier attains the minimax optimal rate of $n^{-(1+\alpha)/(2+\alpha+d)}$ up to a polylogarithmic factor in n. Neither our smoothness and tail assumptions, nor our conclusions are directly comparable with the work of Gadat, Klein and Marteau (2016). In particular, we make a stronger smoothness assumption on η in a neighbourhood of the Bayes decision boundary, implying that the margin assumption holds with parameter $\alpha = 1$; see Lemma 1 in the online supplement (Cannings, Berrett and Samworth (2019)). This enables us to show that our semi-supervised classifier attains faster rates than are achievable under just a Lipschitz condition, and that these rates are minimax optimal up to subpolynomial factors in n, over all possible values of our tail parameter ρ; moreover, we are also able to provide the leading constants in the asymptotic expansion of the excess risk in some cases.

The remainder of this paper is organised as follows. After introducing our setting in Section 2, we present in Section 3 our main results for the standard k-nearest neighbour classifier. This leads on, in Section 4, to our study of the semi-supervised setting, where we derive asymptotic results of the excess risk of our local-k-nearest neighbour classifier. Our minimax lower bound in presented in Section 5. The main arguments of the proofs of our theoretical results are given in Section 6, while in the online supplement (Cannings, Berrett and Samworth (2019)), we prove several claims made in the main text, bound various remainder terms, illustrate the finite-sample benefits of the semi-supervised classifier over the standard k-nearest neighbour classifier in a simulation study and provide an introduction to the ideas of differential geometry that underpin much of our analysis.

Finally, we fix here some notation used throughout the paper. Let $\|\cdot\|$ denote the Euclidean norm and, for $r > 0$ and $x \in \mathbb{R}^d$, let $B_r(x) := \{z \in \mathbb{R}^d : \|x - z\| < r\}$ and $\bar{B}_r(x) := \{z \in \mathbb{R}^d : \|x - z\| \leq r\}$ denote respectively the open and closed Euclidean balls of radius r centred at x. Let $a_d := \frac{2\pi^{d/2}}{d\Gamma(d/2)}$ denote the d-dimensional Lebesgue measure of $B_1(0)$.

For a real-valued function g defined on $A \subseteq \mathbb{R}^d$ that is twice differentiable at x, write $\dot{g}(x) = (g_1(x), \ldots, g_d(x))^T$ and $\ddot{g}(x) = (g_{jk}(x))$ for its gradient vector and Hessian matrix at x, and let $\|g\|_\infty = \sup_{x \in A} |g(x)|$. We write $\|\cdot\|_{\mathrm{op}}$ for the operator norm of a matrix.

2. Statistical setting. Let $(X, Y), (X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})$ be independent and identically distributed random pairs taking values in $\mathbb{R}^d \times \{0, 1\}$. Let $\pi_r := \mathbb{P}(Y = r)$, for $r = 0, 1$, and $X|Y = r \sim P_r$, for $r = 0, 1$, where P_r is a probability measure on $\mathbb{R}^d$. Let $\eta(x) := \mathbb{P}(Y = 1|X = x)$ denote the regression function and $P_X := \pi_0 P_0 + \pi_1 P_1$ denote the marginal distribution of X. We observe *labelled training data*, $\mathcal{T}_n := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, and *unlabelled training data*, $\mathcal{T}'_m := \{X_{n+1}, \ldots, X_{n+m}\}$, and are presented with the task of assigning the *test point* X to either class 0 or 1.

A *classifier* is a Borel measurable function $C : \mathbb{R}^d \to \{0, 1\}$, with the interpretation that C assigns $x \in \mathbb{R}^d$ to the class $C(x)$. Given a Borel measurable set $\mathcal{R} \subseteq \mathbb{R}^d$, the misclassification rate, or *risk*, over $\mathcal{R}$ is

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

When $\mathcal{R} = \mathbb{R}^d$, we drop the subscript for convenience. The *Bayes classifier*

$$C^{\mathrm{Bayes}}(x) := \begin{cases} 1 & \text{if } \eta(x) \geq 1/2; \\ 0 & \text{otherwise,} \end{cases}$$

minimises the risk over any region $\mathcal{R}$ (Devroye, Györfi and Lugosi (1996, p. 20)). The performance of a classifier C is therefore measured via its *excess risk*, $R_{\mathcal{R}}(C) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})$.

We can now formally define the local-k-nearest neighbour classifier, which allows the number of neighbours considered to vary depending on the location of the test point. Suppose $k_{\mathrm{L}} : \mathbb{R}^d \to \{1, \ldots, n\}$ is measurable. Given the test point $x \in \mathbb{R}^d$, let $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ be a reordering of the training data such that $\|X_{(1)} - x\| \leq \cdots \leq \|X_{(n)} - x\|$. We will later assume that P_X is absolutely continuous with respect to d-dimensional Lebesgue measure, which ensures that ties occur with probability zero; where helpful for clarity, we also write $X_{(i)}(x)$ for the ith nearest neighbour of x. Let $\hat{S}_n(x) := k_{\mathrm{L}}(x)^{-1} \sum_{i=1}^{k_{\mathrm{L}}(x)} \mathbb{1}_{\{Y_{(i)}=1\}}$. Then the *local-$k$-nearest neighbour ($k_{\mathrm{L}}$nn) classifier* is defined to be

$$\hat{C}_n^{k_{\mathrm{L}}\mathrm{nn}}(x) := \begin{cases} 1 & \text{if } \hat{S}_n(x) \geq 1/2; \\ 0 & \text{otherwise.} \end{cases}$$

Given $k \in \{1, \ldots, n\}$, let k_0 denote the constant function $k_0(x) := k$ for all $x \in \mathbb{R}^d$. Using $k_{\mathrm{L}} = k_0$, the definition above reduces to the standard *k-nearest neighbour classifier* ($k$nn), and we will write $\hat{C}_n^{k\mathrm{nn}}$ in place of $\hat{C}_n^{k_0\mathrm{nn}}$. For $\beta \in (0, 1/2)$, let

$$K_\beta \equiv K_{\beta,n} := \big\{\lceil (n-1)^\beta \rceil, \lceil (n-1)^\beta \rceil + 1, \ldots, \lfloor (n-1)^{1-\beta} \rfloor\big\}$$

denote a range of values of k that will be of interest to us. Note that $K_{\beta_1} \supseteq K_{\beta_2}$, for $\beta_1 < \beta_2$. Moreover, when β is small, the restriction that $k \in K_\beta$ is only a slightly stronger requirement than the consistency conditions of Stone (1977), namely that $k = k_n \to \infty$, $k_n/n \to 0$ as $n \to \infty$.

3. Global risk of the k-nearest neighbour classifier. In this section, we provide an asymptotic expansion for the global risk of the standard (nonlocal) k-nearest neighbour classifier. We first define the classes of data generating mechanisms over which our results will hold. Let $\mathcal{L}$ denote the class of decreasing functions $\ell : (0, \infty) \to [1, \infty)$ such that

$\ell(\delta) = o(\delta^{-\tau})$ as $\delta \searrow 0$, for every $\tau > 0$. Let $\mathcal{G}$ denote the class of strictly increasing functions $g : (0, 1) \to (0, 1)$ with $g(\epsilon) = o(\epsilon^M)$ as $\epsilon \searrow 0$, for every $M > 0$. Recall from Section 2 that, to any distribution P on $\mathbb{R}^d \times \{0, 1\}$, we associate conditional distributions P_0, P_1, a regression function η, marginal probabilities π_0, π_1 and a marginal distribution P_X. Now, for $\Theta := (0, \infty) \times [1, \infty) \times (0, \infty) \times \mathcal{L} \times \mathcal{G}$, and $\theta = (\epsilon_0, M_0, \rho, \ell, g) \in \Theta$, let $\mathcal{P}_{d,\theta}$ denote the class of distributions P on $\mathbb{R}^d \times \{0, 1\}$ such that the probability measures P_0 and P_1 are absolutely continuous with respect to Lebesgue measure, with Radon–Nikodym derivatives f_0 and f_1, respectively. Moreover, we assume that there exist versions of f_0 and f_1 for which the following conditions hold:

(A.1) The marginal density of X, namely $\bar{f} := \pi_0 f_0 + \pi_1 f_1$, is continuous P_X-almost everywhere and the set $\mathcal{X}_{\bar{f}}$ of continuity points of $\bar{f}$ is open.

Thus $\eta(x) := \pi_1 f_1(x)/\{\pi_0 f_0(x) + \pi_1 f_1(x)\}$, where we define $0/0 := 0$. Let $\mathcal{S} := \{x \in \mathbb{R}^d : \eta(x) = 1/2\}$ and, for $\epsilon > 0$, let $\mathcal{S}^\epsilon := \mathcal{S} + B_\epsilon(0)$. In our assumptions below, we will place further assumptions on $\mathcal{S}$, which ensure not only that this set is nonempty, but in fact that it is a $(d-1)$-dimensional, orientable manifold.

(A.2) The set $\mathcal{S} \cap \{x \in \mathbb{R}^d : \bar{f}(x) > 0\}$ is nonempty and $\sup_{x_0 \in \mathcal{S}} \bar{f}(x_0) \le M_0$. The function $\bar{f}$ is twice continuously differentiable on $\mathcal{S}^{\epsilon_0}$, and

$$\max\Big\{\|\dot{\bar{f}}(x_0)\|, \sup_{u \in B_{\epsilon_0}(0)} \|\ddot{\bar{f}}(x_0 + u)\|_{\mathrm{op}}\Big\} \le \bar{f}(x_0)\ell\big(\bar{f}(x_0)\big), \tag{1}$$

for all $x_0 \in \mathcal{S}$. Furthermore, writing $p_r(x) := P_X(B_r(x))$, we have for all $x \in \mathbb{R}^d \setminus \mathcal{S}^{\epsilon_0}$ and $r \in (0, \epsilon_0]$ that

$$p_r(x) \ge \epsilon_0 a_d r^d \bar{f}(x).$$

(A.3) We have that η is twice differentiable on $\mathcal{S}^{2\epsilon_0}$ with $\inf_{x_0 \in \mathcal{S}} \|\dot{\eta}(x_0)\| \ge \epsilon_0 M_0$. Moreover, $\sup_{x \in \mathcal{S}^{2\epsilon_0}} \|\dot{\eta}(x)\| \le M_0$, $\sup_{x \in \mathcal{S}^{2\epsilon_0}} \|\ddot{\eta}(x)\|_{\mathrm{op}} \le M_0$ and given $\epsilon > 0$,

$$\sup_{x, z \in \mathcal{S}^{2\epsilon_0} : \|z - x\| \le g(\epsilon)} \|\ddot{\eta}(z) - \ddot{\eta}(x)\|_{\mathrm{op}} \le \epsilon.$$

Finally, the function η is continuous on $\{x : \bar{f}(x) > 0\}$, and

$$|\eta(x) - 1/2| \ge \frac{1}{\ell(\bar{f}(x))}$$

for all $x \in \mathbb{R}^d \setminus \mathcal{S}^{\epsilon_0}$.

(A.4) We have $\int_{\mathbb{R}^d} \|x\|^\rho \, dP_X(x) \le M_0$.

EXAMPLE 1. Consider the distribution P on $\mathbb{R}^d \times \{0, 1\}$ for which $\bar{f}(x) = \frac{\Gamma(3+d/2)}{2\pi^{d/2}}(1 - \|x\|^2)^2 \mathbb{1}_{\{x \in B_1(0)\}}$ and $\eta(x) = \min(\|x\|^2, 1)$. In Section 2 of the online supplement, we show that $P \in \mathcal{P}_{d,\theta}$ with $\theta = (\epsilon_0, M_0, \rho, \ell, g) \in \Theta$ for any $\rho > 0$, $g \in \mathcal{G}$, and provided that $M_0 \ge \max\{2, \frac{\Gamma(3+d/2)}{8\pi^{d/2}}\}$, $\epsilon_0 \le \min(\frac{1}{10}, 2^{-d}, \frac{2^{1/2}}{M_0})$ and $\ell \in \mathcal{L}$ satisfies $\ell(\delta) \ge \max(48, \epsilon_0^{-1})$ for all $\delta > 0$.

Asking for P_X to have a Lebesgue density allows us to define the tail of the distribution as the region where $\bar{f}$ is smaller than some threshold. Condition (A.1) ensures that for all $\delta > 0$ sufficiently small, the set $\mathcal{R} := \{x : \bar{f}(x) > \delta\} \cap \mathcal{X}_{\bar{f}}$ is a d-dimensional manifold, and $P_X(\mathcal{R}^c) \le \mathbb{P}\{\bar{f}(X) \le \delta\}$, where the latter quantity can be bounded using (A.4). The first part of (A.2) asks for a certain level of smoothness for $\bar{f}$ in a neighbourhood of $\mathcal{S}$, and

controls the behaviour of its first and second derivatives there relative to the original density. In particular, the greater degree of regularity asked of these derivatives in the tails of the marginal density in (1) allows us still to control the error of a Taylor approximation even in this region. The condition (1) is satisfied by all Gaussian and multivariate-t densities, for example, for appropriate choices of ϵ_0 and ℓ. The last part of (A.2) concerns the behaviour of the marginal feature distribution away from $\mathcal{S}^{\epsilon_0}$ and is often referred to as the strong minimal mass assumption (e.g., Gadat, Klein and Marteau (2016)). It requires that the mass of the marginal feature distribution is not concentrated in the neighbourhood of a point and is a rather weaker condition than we ask for on $\mathcal{S}^{\epsilon_0}$; in particular, we do not insist that derivatives of $\bar{f}$ exist in this region.

The condition $\inf_{x_0 \in \mathcal{S}} \|\dot{\eta}(x_0)\| \geq \epsilon_0 M_0$ in (A.3) asks for the class conditional densities, when weighted by their respective prior probabilities, to cross at an angle; in particular, this ensures that $\mathcal{S}$ is a $(d-1)$-dimensional, orientable manifold (cf. Section 7.3 of the online supplement). Moreover, the bounds on the first and second derivatives of η in a neighbourhood of $\mathcal{S}$ ensure that we can estimate η sufficiently well. The last part of (A.3) asks that η does not approach the critical value of $1/2$ too fast on the complement of $\mathcal{S}^{\epsilon_0}$. Assumption (A.4) is a simple moment condition that, together with (A.2), ensures that the constants B_1 and B_2 in (2) below are finite where needed.

Let $d\,\mathrm{Vol}^{d-1}$ denote the $(d-1)$-dimensional volume form on $\mathcal{S}$ (cf. Section 7.3 of the online supplement). Now let

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

where

$$a(x) := \frac{\sum_{j=1}^{d}\{\eta_j(x)\bar{f}_j(x) + \frac{1}{2}\eta_{jj}(x)\bar{f}(x)\}}{(d+2)a_d^{2/d}\bar{f}(x)}. \tag{3}$$

We are now in a position to present our asymptotic expansion for the global excess risk of the standard k-nearest neighbour classifier.

THEOREM 1. *Fix $d \in \mathbb{N}$ and $\theta = (\epsilon_0, M_0, \rho, \ell, g) \in \Theta$ such that $\mathcal{P}_{d,\theta} \neq \varnothing$.*

(i) *Suppose that $d \geq 5$ and $\rho > \frac{4d}{d-4}$. Then for each $\beta \in (0, 1/2)$,*

$$\sup_{P \in \mathcal{P}_{d,\theta}} \left| R(\hat{C}_n^{k\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) - \frac{B_1}{k} - B_2\left(\frac{k}{n}\right)^{4/d} \right| = o\left(\frac{1}{k} + \left(\frac{k}{n}\right)^{4/d}\right)$$

as $n \to \infty$, uniformly for $k \in K_\beta$.

(ii) *Suppose that either $d \leq 4$, or, $d \geq 5$ and $\rho \leq \frac{4d}{d-4}$. Then for each $\beta \in (0, 1/2)$ and each $\epsilon > 0$ we have*

$$\sup_{P \in \mathcal{P}_{d,\theta}} \left| R(\hat{C}_n^{k\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) - \frac{B_1}{k} \right| = o\left(\frac{1}{k} + \left(\frac{k}{n}\right)^{\frac{\rho}{\rho+d}-\epsilon}\right)$$

as $n \to \infty$, uniformly for $k \in K_\beta$.

Theorem 1 reveals an interesting dichotomy: when $d \geq 5$ and $\rho > 4d/(d-4)$, the dominant contribution to the excess risk arises from the difficulty of classifying points close

to the Bayes decision boundary $\mathcal{S}$. In such settings, the excess risk of the standard k-nearest neighbour classifier converges to zero at rate $O(n^{-4/(d+4)})$ when k is chosen proportional to $n^{4/(d+4)}$. On the other hand, part (ii) shows that when either $d \leq 4$ or $d \geq 5$ and $\rho \leq 4d/(d-4)$, the dominant contribution to the excess risk when k is large may come from the challenge of classifying points in the tails of the distribution. Indeed, Example 2 below provides one simple setting where this dominant contribution does come from the tails of the distribution.

EXAMPLE 2. Suppose that the joint density of X at $x = (x_1, x_2) \in (0,1) \times \mathbb{R}$ is given by $\bar{f}(x) = 2x_1 f_2(x_2)$, where f_2 is a positive, twice continuously differentiable density with $f_2(x_2) = e^{-|x_2|}/2$ for $|x_2| > 1$. Suppose also that $\eta(x) = x_1$. Then the corresponding joint distribution P belongs to $\mathcal{P}_{2,\theta}$ provided $\theta = (\epsilon_0, M_0, \rho, \ell, g)$ is such that M_0 is sufficiently large, $\epsilon_0 \leq \min(1/8, 1/M_0)$ and ℓ is a sufficiently large constant ($\rho > 0$ and $g \in \mathcal{G}$ can be chosen arbitrarily). We prove in Section 3 in the supplementary material that for every $\beta \in (0, 1/2)$ and $\epsilon > 0$,

$$\liminf_{n\to\infty} \inf_{k \in K_\beta} \left\{ k + \left(\frac{n}{k}\right)^{1+\epsilon} \right\} \{ R(\hat{C}_n^{k\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) \} > 0 \tag{4}$$

as $n \to \infty$. Thus the rate of convergence in this example is at best $n^{-1/2}$, up to subpolynomial factors, whereas a rate of $n^{-2/3}$ is achievable over any compact set.

The proof of Theorem 1, and indeed the proofs of Theorems 2 and 3 that follow in Section 4 below, depend crucially on Theorem 5 in Section 6. This result provides an asymptotic expansion for the excess risk of a general (local or global) k-nearest neighbour classifier over a region $\mathcal{R}_n \subseteq \{x \in \mathbb{R}^d : \bar{f}(x) \geq \delta_n(x)\}$, where $\delta_n(x)$, defined in (7) below, shrinks to zero at a rate slow enough to ensure that $X_{(k)}(x)$ concentrates around x uniformly over $\mathcal{R}_n$. The intuition regarding the behaviour of the excess risk, then is that when $x \in \mathcal{R}_n$ and x is not close to $\mathcal{S}$, with high probability the k nearest neighbours of x are on the same side of $\mathcal{S}$ as x; that is, $\mathrm{sgn}(\eta(X_{(i)}) - 1/2) = \mathrm{sgn}(\eta(x) - 1/2)$ for $i = 1, \ldots, k$. The probability of classifying x differently from the Bayes classifier can therefore be shown to be $O(n^{-M})$ for every $M > 0$, using Hoeffding's inequality. Thus, the challenging regions for classification consist of neighbourhoods of $\mathcal{S}$, where η is close to $1/2$, together with $\mathcal{R}_n^c$, where we no longer enjoy the same nearest neighbour concentration properties. For the first of these regions, we exploit our smoothness assumptions to derive asymptotic expansions for the bias and variance of $\hat{S}_n(x)$, uniformly over appropriate neighbourhoods of $\mathcal{S}$, and using a normal approximation, we can deduce an asymptotic expansion for the excess risk, uniformly over our classes of distributions and an appropriate set of nearest neighbour classifiers. For $\mathcal{R}_n^c$, we are unable to bound the probability of classifying differently from the Bayes classifier with anything other than a trivial bound, but we can control $P_X(\mathcal{R}_n^c)$ using (A.4).

Finally in this section, we mention that Samworth (2012) obtained a similar expansion to that in Theorem 1(i) for a fixed distribution P satisfying certain smoothness conditions. However, there the risk was computed only over a compact set, so the analysis failed to elucidate the important effects of tail behaviour on the excess risk. Another key difference is that here we define classes $\mathcal{P}_{d,\theta}$, and show that the remainder terms in our asymptotic expansion hold *uniformly* over these classes; the introduction of these classes further facilitates the study of corresponding minimax lower bounds in Section 5 below.

4. Local-k-nearest neighbour classifiers. In this section, we explore the consequences of a local choice of k, compared with the global choice in Theorem 1. Initially, we consider an oracle choice, where k is allowed to depend on the marginal feature density $\bar{f}$ (Section 4.1), but we then relax this to semi-supervised settings, where $\bar{f}$ can be estimated from unlabelled training data (Section 4.2).

4.1. *Oracle classifier.* Suppose for now that the marginal density $\bar{f}$ is known. For $\beta \in (0, 1/2)$ and $B > 0$, let

$$k_{\mathrm{O}}(x) := \max\bigl[\bigl\lceil (n-1)^{\beta} \bigr\rceil, \min\bigl\{\bigl\lfloor B\bigl\{\bar{f}(x)(n-1)\bigr\}^{4/(d+4)} \bigr\rfloor, \bigl\lfloor (n-1)^{1-\beta} \bigr\rfloor\bigr\}\bigr], \tag{5}$$

where the subscript O refers to the fact that this is an oracle choice of the function k_{L}, since it depends on $\bar{f}$. This choice aims to balance the local bias and variance of $\hat{S}_n(x)$.

THEOREM 2. *Fix $d \in \mathbb{N}$ and $\theta = (\epsilon_0, M_0, \rho, \ell, g) \in \Theta$ such that $\mathcal{P}_{d,\theta} \neq \varnothing$. For each $0 < B_* \leq B^* < \infty$,*

(i) *if $\rho > 4$ then for $\beta < 4d(\rho-4)/\{\rho(d+4)^2\}$,*

$$\sup_{P \in \mathcal{P}_{d,\theta}} \bigl| R\bigl(\hat{C}_n^{k_{\mathrm{O}}\mathrm{nn}}\bigr) - R\bigl(C^{\mathrm{Bayes}}\bigr) - B_3 n^{-4/(d+4)} \bigr| = o\bigl(n^{-4/(d+4)}\bigr),$$

uniformly for $B \in [B_*, B^*]$ as $n \to \infty$, where

$$B_3 := \int_{\mathcal{S}} \frac{\bar{f}(x_0)^{d/(d+4)}}{\|\dot{\eta}(x_0)\|} \biggl\{ \frac{1}{4B} + B^{4/d} a(x_0)^2 \biggr\} d\,\mathrm{Vol}^{d-1}(x_0) < \infty.$$

(ii) *if $\rho \leq 4$ and $\beta < \min\{1/2, 4/(d+4)\}$, then for every $\epsilon > 0$,*

$$\sup_{P \in \mathcal{P}_{d,\theta}} \bigl\{ R\bigl(\hat{C}_n^{k_{\mathrm{O}}\mathrm{nn}}\bigr) - R\bigl(C^{\mathrm{Bayes}}\bigr) \bigr\} = o\bigl(n^{-\rho/(\rho+d)+\beta+\epsilon}\bigr),$$

uniformly for $B \in [B_*, B^*]$, as $n \to \infty$.

Comparing Theorem 2(i) and Theorem 1(i), we see that, unlike for the global k-nearest neighbour classifier, we can guarantee a $O(n^{-4/(d+4)})$ rate of convergence for the excess risk of the oracle classifier, both in low dimensions ($d \leq 4$), and under a weaker condition on ρ when $d \geq 5$. In particular, the condition on ρ no longer depends on the dimension of the covariates. The guarantees in Theorem 2(ii) are also stronger than those provided by Theorem 1(ii) for any global choice of k. Examining the proof of Theorem 2, we find that the key difference with the proof of Theorem 1 is that we can now choose the region $\mathcal{R}_n$ (cf. the discussion of the proof of Theorem 1 in Section 3) to be larger.

4.2. *The semi-supervised nearest neighbour classifier.* Now consider the more realistic setting where the marginal density $\bar{f}$ of X is unknown, but where we have access to an estimate $\hat{f}_m$ based on the unlabelled training set $\mathcal{T}'_m$. Of course, many different techniques are available, but for simplicity, we focus here on a kernel method. Let K be a bounded kernel with $\int_{\mathbb{R}^d} K(x)\,dx = 1$, $\int_{\mathbb{R}^d} xK(x)\,dx = 0$, $\int_{\mathbb{R}^d} \|x\|^2 |K(x)|\,dx < \infty$, and let $R(K) := \int_{\mathbb{R}^d} K(x)^2\,dx$. We further assume that $K(x) = Q(p(x))$, where p is a polynomial and Q is a function of bounded variation. Now define a kernel density estimator of $\bar{f}$, given by

$$\hat{f}_m(x) = \hat{f}_{m,h}(x) := \frac{1}{mh^d} \sum_{j=1}^{m} K\biggl(\frac{x - X_{n+j}}{h}\biggr).$$

Motivated by the oracle local choice of k in (5), for $\beta \in (0, 1/2)$ and $B > 0$, let

$$k_{\mathrm{SS}}(x) := \max\bigl[\bigl\lceil (n-1)^{\beta} \bigr\rceil, \min\bigl\{\bigl\lfloor B\bigl\{\hat{f}_m(x)(n-1)\bigr\}^{4/(d+4)} \bigr\rfloor, \bigl\lfloor (n-1)^{1-\beta} \bigr\rfloor\bigr\}\bigr].$$

Our main result in this setting will require an additional smoothness condition on the marginal feature density $\bar{f}$ in order to ensure that $\hat{f}_m$ estimates it well. For $d \in \mathbb{N}$, $\gamma \in (0, 1]$ and $\lambda > 0$, let $\mathcal{Q}_{d,\gamma,\lambda}$ denote the class of distributions P on $\mathbb{R}^d \times \{0, 1\}$ whose marginal distribution P_X

is absolutely continuous with respect to Lebesgue measure with Radon–Nikodym derivative $\bar{f}$ satisfying $\|\bar{f}\|_\infty \le \lambda$ and

$$\|\bar{f}(y) - \bar{f}(x)\| \le \lambda \|y - x\|^\gamma \quad \text{for all } x, y \in \mathbb{R}^d.$$

If $\gamma \in (1, 2]$, then we define $\mathcal{Q}_{d,\gamma,\lambda}$ to consist of distributions P on $\mathbb{R}^d \times \{0, 1\}$ whose marginal distribution P_X is again absolutely continuous with Radon–Nikodym derivative $\bar{f}$ satisfying $\|\bar{f}\|_\infty \le \lambda$, but we now ask that $\bar{f}$ be differentiable, and that

$$\|\dot{\bar{f}}(y) - \dot{\bar{f}}(x)\| \le \lambda \|y - x\|^{\gamma - 1} \quad \text{for all } x, y \in \mathbb{R}^d.$$

In Section 2 of the online supplement, we show that the distribution considered in Example 1 belongs to $\mathcal{Q}_{d,\gamma,\lambda}$ with $\gamma = 2$ provided that $\lambda \ge 6\pi^{-d/2}\Gamma(3 + d/2)$.

THEOREM 3. *Fix $d \in \mathbb{N}$, $\theta = (\epsilon_0, M_0, \rho, \ell, g) \in \Theta$, $\gamma \in (0, 2]$ and $\lambda > 0$ such that $\mathcal{P}_{d,\theta} \cap \mathcal{Q}_{d,\gamma,\lambda} \neq \varnothing$. Let $m_0 > 0$, let $0 < A_* \le A^* < \infty$ and $0 < B_* \le B^* < \infty$, and let $h = h_m := Am^{-1/(d+2\gamma)}$ for some $A > 0$.*

(i) *If $\rho > 4$ and $\beta < 4d(\rho - 4)/\{\rho(d + 4)^2\}$,*

$$\sup_{P \in \mathcal{P}_{d,\theta} \cap \mathcal{Q}_{d,\gamma,\lambda}} \left| R\big(\hat{C}_n^{k_{\mathrm{SS}}\mathrm{nn}}\big) - R\big(C^{\mathrm{Bayes}}\big) - B_3 n^{-4/(d+4)} \right| = o\big(n^{-4/(d+4)}\big)$$

uniformly for $A \in [A_*, A^*]$, $B \in [B_*, B^*]$ and $m = m_n \ge m_0(n - 1)^{2+d/\gamma}$, where B_3 was defined in Theorem 2(i).

(ii) *if $\rho \le 4$ and $\beta < \min\{1/2, 4/(d + 4)\}$, then for every $\epsilon > 0$,*

$$\sup_{P \in \mathcal{P}_{d,\theta} \cap \mathcal{Q}_{d,\gamma,\lambda}} \left\{ R\big(\hat{C}_n^{k_{\mathrm{SS}}\mathrm{nn}}\big) - R\big(C^{\mathrm{Bayes}}\big) \right\} = o\big(n^{-\rho/(\rho+d)+\beta+\epsilon}\big),$$

uniformly for $A \in [A_*, A^*]$, $B \in [B_*, B^*]$ and $m = m_n \ge m_0(n - 1)^{2+d/\gamma}$.

Examination of the proof of Theorem 3 reveals that the key property of our kernel estimator $\hat{f}_m$ of $\bar{f}$ is that there exists $\alpha > (1 + d/4)\beta$ such that

$$\sup_{P \in \mathcal{P}_{d,\theta} \cap \mathcal{Q}_{d,\gamma,\lambda}} \mathbb{P}\left(\|\hat{f}_m - \bar{f}\|_\infty \ge \frac{1}{(n - 1)^{1-\alpha/2}} \right) = o\big(n^{-4/(d+4)}\big). \tag{6}$$

This observation would allow similar results to Theorem 3 to be proved for other versions of the semi-supervised nearest neighbour classifier, with alternative estimators of $\bar{f}$ in the definition of $\hat{k}_{\mathrm{SS}}(\cdot)$, subject potentially to suitable modifications of the class $\mathcal{Q}_{d,\gamma,\lambda}$. It is therefore not our intention to argue that the kernel density approach is superior to other methods of estimating the marginal density $\bar{f}$.

5. Minimax lower bounds. Our main minimax lower bound is the following.

THEOREM 4. *Fix $d \in \mathbb{N}$, $\rho > 0$, $g \in \mathcal{G}$ with $r \mapsto r/g^{-1}(r)$ increasing for sufficiently small $r > 0$, and $\gamma \in (0, 2]$. There exist $\lambda_* > 0$, $\epsilon_* > 0$ and $M_* > 0$, depending only on d, such that for $\lambda \ge \lambda_*$, $M_0 \ge M_*$, $\epsilon_0 \in (0, \min(\epsilon_*, 1/(4M_0))]$ and $\ell \in \mathcal{L}$ with $\ell(\delta) \ge 2/\epsilon_0$ for all $\delta \in (0, \infty)$, writing $\theta = (\epsilon_0, M_0, \rho, \ell, g) \in \Theta$, we can find $c = c(d, \theta, \gamma, \lambda) > 0$ such that for all $n \in \mathbb{N}$ and all $\nu \ge 0$, we have*

$$\inf_{C_n} \sup_{P \in \mathcal{P}_{d,\theta} \cap \mathcal{Q}_{d,\gamma,\lambda}} \left\{ R(C_n) - R\big(C^{\mathrm{Bayes}}\big) \right\} \ge c g^{-1}(1/q)^{\frac{2d(1+\nu)}{4+d+\nu(\rho+d)}} n^{-\frac{4+\nu\rho}{4+d+\nu(\rho+d)}},$$

where $q = q_n \in (1/\|g\|_\infty, \infty)$ *is the unique solution to* $\frac{q^{4+d+\nu(\rho+d)}}{g^{-1}(1/q)^2} = n$ *and the infimum is taken over all measurable functions* $C_n : (\mathbb{R}^d \times \{0,1\})^{\times n} \times \mathbb{R}^d \to \{0,1\}$. *In particular, for every* $\epsilon > 0$, *there exists* $c = c(d, \theta, \gamma, \lambda, \epsilon) > 0$ *such that*

$$\inf_{C_n} \sup_{P \in \mathcal{P}_{d,\theta} \cap \mathcal{Q}_{d,\gamma,\lambda}} \{R(C_n) - R(C^{\mathrm{Bayes}})\} \geq c n^{-(\min\{\frac{4}{4+d}, \frac{\rho}{\rho+d}\}+\epsilon)}.$$

REMARK 1. The proof of this result also reveals that the lower bound holds if the classifier is allowed to depend on some unlabelled data or even the true marginal X density $\bar{f}$.

EXAMPLE 3. Consider the case where $g(\epsilon) = \exp(-1/\epsilon)$, so $g \in \mathcal{G}$. Then for $q \in (1, \infty)$, we have $g^{-1}(1/q) = 1/\log q$, so for $n \in \mathbb{N}$,

$$g^{-1}(1/q_n)^{\frac{2d(1+\nu)}{4+d+\nu(\rho+d)}} n^{-\frac{4+\nu\rho}{4+d+\nu(\rho+d)}} \geq \frac{1}{\{1 + \frac{\log n}{4+d+\nu(\rho+d)}\}^{\frac{2d(1+\nu)}{4+d+\nu(\rho+d)}}} n^{-\frac{4+\nu\rho}{4+d+\nu(\rho+d)}}.$$

Thus, if $\rho > 4$, then we can take $\nu = 0$ in Theorem 4 to obtain a minimax lower bound of order $n^{-4/(4+d)}/\log^2 n$; on the other hand, if $\rho \leq 4$, then we can take $\nu = \log^{1/2} n$ to obtain a minimax lower bound of order $n^{-(\frac{\rho}{\rho+d}+\epsilon)}$, for every $\epsilon > 0$. Combining this result with Theorem 3, we see that for every $\rho \in (0, \infty)$, our semi-supervised local-k-nearest neighbour classifier attains the minimax optimal rate over the class $\mathcal{P}_{d,\theta} \cap \mathcal{Q}_{d,\gamma,\lambda}$ up to polylogarithmic factors when $\rho > 4$ and up to subpolynomial factors when $\rho \leq 4$.

6. Proofs. The proofs of Theorems 1, 2 and 3 rely on the general asymptotic expansion presented in Theorem 5 below. We begin with some further notation. Define the $d \times n$ matrices $X^n := (X_1, \ldots, X_n)$ and $x^n := (x_1, \ldots, x_n)$. Write

$$\hat{\mu}_n(x) = \hat{\mu}_n(x, x^n) := \mathbb{E}\{\hat{S}_n(x) | X^n = x^n\} = \frac{1}{k_{\mathrm{L}}(x)} \sum_{i=1}^{k_{\mathrm{L}}(x)} \eta(x_{(i)}),$$

and

$$\hat{\sigma}_n^2(x) = \hat{\sigma}_n^2(x, x^n) := \mathrm{Var}\{\hat{S}_n(x) | X^n = x^n\} = \frac{1}{k_{\mathrm{L}}(x)^2} \sum_{i=1}^{k_{\mathrm{L}}(x)} \eta(x_{(i)})\{1 - \eta(x_{(i)})\}.$$

Here, we have used the fact that the ordered labels $Y_{(1)}, \ldots, Y_{(n)}$ are independent given X^n, satisfying $\mathbb{P}(Y_{(i)} = 1 | X^n) = \eta(X_{(i)})$. Since η takes values in $[0,1]$ it is clear that $0 \leq \hat{\sigma}_n^2(x) \leq \frac{1}{4k_{\mathrm{L}}(x)}$ for all $x \in \mathbb{R}^d$. Further, write $\mu_n(x) := \mathbb{E}\{\hat{S}_n(x)\} = \frac{1}{k_{\mathrm{L}}(x)} \sum_{i=1}^{k_{\mathrm{L}}(x)} \mathbb{E}\eta(X_{(i)})$ for the unconditional expectation of $\hat{S}_n(x)$. Recall also that $p_r(x) = P_X(B_r(x))$.

6.1. *A general asymptotic expansion.* Let

$$c_n := \sup_{x_0 \in \mathcal{S}} \ell\left(\frac{k_{\mathrm{L}}(x_0)}{n-1}\right).$$

Further, for $x \in \mathbb{R}^d$, let

$$\delta_n(x) = \delta_{n,\mathrm{L}}(x) := \frac{k_{\mathrm{L}}(x)}{n-1} c_n^d \log^d\left(\frac{n-1}{k_{\mathrm{L}}(x)}\right). \tag{7}$$

Recall that $\mathcal{S} = \{x \in \mathbb{R}^d : \eta(x) = 1/2\}$, and note that by Proposition 2 in the online supplement, for $\epsilon > 0$, we can write

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

Let

$$\epsilon_n := \frac{1}{c_n \beta^{1/2} \log^{1/2}(n-1)}, \tag{8}$$

and recall the definition of the function $a(\cdot)$ in (3).

THEOREM 5. *Fix $d \in \mathbb{N}$ and $\theta = (\epsilon_0, M_0, \rho, \ell, g) \in \Theta$ such that $\mathcal{P}_{d,\theta} \neq \varnothing$. For n sufficiently large, let $\mathcal{R}_n \subseteq \{x \in \mathbb{R}^d : \bar{f}(x) \geq \delta_n(x)\}$ be a d-dimensional manifold. Write $\partial\mathcal{R}_n$ for the topological boundary of $\mathcal{R}_n$, let $(\partial\mathcal{R}_n)^\epsilon := \partial\mathcal{R}_n + \epsilon \bar{B}_1(0)$, and let $\mathcal{S}_n := \mathcal{S} \cap \mathcal{R}_n$. For $\beta \in (0, 1/2)$ and $\tau > 0$, define the class of functions*

$$K_{\beta,\tau} \equiv K_{\beta,\tau,n} := \left\{ k_{\mathrm{L}} : \mathbb{R}^d \to K_\beta : \sup_{x_0 \in \mathcal{S}_n} \sup_{|t| < \epsilon_n} \left| \frac{k_{\mathrm{L}}(x_0 + t \frac{\dot{\eta}(x_0)}{\|\dot{\eta}(x_0)\|})}{k_{\mathrm{L}}(x_0)} - 1 \right| \leq \tau \right\}.$$

Then for each $\beta \in (0, 1/2)$ and each $\tau = \tau_n$ with $\tau_n \searrow 0$, we have

$$\begin{aligned} &R_{\mathcal{R}_n}(\hat{C}_n^{k_{\mathrm{L}}\mathrm{nn}}) - R_{\mathcal{R}_n}(C^{\mathrm{Bayes}}) \\ &\quad = \int_{\mathcal{S}_n} \frac{\bar{f}(x_0)}{\|\dot{\eta}(x_0)\|} \left\{ \frac{1}{4k_{\mathrm{L}}(x_0)} + \left(\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)} \right)^{4/d} a(x_0)^2 \right\} d\,\mathrm{Vol}^{d-1}(x_0) \\ &\qquad + W_{n,1} + W_{n,2} \end{aligned}$$

as $n \to \infty$, where $\sup_{P \in \mathcal{P}_{d,\theta}} \sup_{k_{\mathrm{L}} \in K_{\beta,\tau}} |W_{n,1}|/\gamma_n(k_{\mathrm{L}}) \to 0$ with

$$\gamma_n(k_{\mathrm{L}}) := \int_{\mathcal{S}_n} \frac{\bar{f}(x_0)}{\|\dot{\eta}(x_0)\|} \left\{ \frac{1}{4k_{\mathrm{L}}(x_0)} + \left(\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)} \right)^{4/d} \ell\big(\bar{f}(x_0)\big)^2 \right\} d\,\mathrm{Vol}^{d-1}(x_0),$$

and where $\limsup_{n\to\infty} \sup_{P \in \mathcal{P}_{d,\theta}} \sup_{k_{\mathrm{L}} \in K_{\beta,\tau}} |W_{n,2}|/P_X((\partial\mathcal{R}_n)^{\epsilon_n} \cap \mathcal{S}^{\epsilon_n}) \leq 1$.

PROOF OF THEOREM 5. First, observe that

$$\begin{aligned} &R_{\mathcal{R}_n}(\hat{C}_n^{k_{\mathrm{L}}\mathrm{nn}}) - R_{\mathcal{R}_n}(C^{\mathrm{Bayes}}) \\ &\quad = \int_{\mathcal{R}_n} \big[\mathbb{P}\{\hat{S}_n(x) < 1/2\} - \mathbb{1}_{\{\eta(x) < 1/2\}} \big] \{2\eta(x) - 1\} \bar{f}(x)\, dx. \end{aligned} \tag{9}$$

The proof is presented in seven steps. We will see that the dominant contribution to the integral in (9) arises from a small neighbourhood about the Bayes decision boundary, that is, the region $\mathcal{S}^{\epsilon_n} \cap \mathcal{R}_n$. On $\mathcal{R}_n \setminus \mathcal{S}^{\epsilon_n}$, the k_{L}nn classifier agrees with the Bayes classifier with high probability (asymptotically). More precisely, we show in Step 4 that

$$\sup_{P \in \mathcal{P}_{d,\theta}} \sup_{k_{\mathrm{L}} \in K_{\beta,\tau}} \sup_{x \in \mathcal{R}_n \setminus \mathcal{S}^{\epsilon_n}} \big| \mathbb{P}\{\hat{S}_n(x) < 1/2\} - \mathbb{1}_{\{\eta(x) < 1/2\}} \big| = O\big(n^{-M}\big),$$

for each $M > 0$, as $n \to \infty$. In Steps 1, 2 and 3, we derive the key asymptotic properties of the bias, conditional (on X^n) bias and variance of $\hat{S}_n(x)$, respectively. In Step 5, we show that the integral over $\mathcal{S}^{\epsilon_n} \cap \mathcal{R}_n$ can be decomposed into an integral over $\mathcal{S}_n$ and one perpendicular to $\mathcal{S}$. Step 6 is dedicated to combining the results of Steps 1–5; we derive the leading order terms in the asymptotic expansion of the integral in (9). Finally, we bound the remaining error terms to conclude the proof in Step 7, which is presented in the supplementary material. To ease notation, where it is clear from the context, we write k_{L} in place of $k_{\mathrm{L}}(x)$.

Step 1: Let $\mu_n(x) := \mathbb{E}\{\hat{S}_n(x)\}$, and for $x_0 \in \mathcal{S}$ and $t \in \mathbb{R}$, write $x = x(x_0, t) := x_0 + t \frac{\dot{\eta}(x_0)}{\|\dot{\eta}(x_0)\|}$. We show that

$$\mu_n(x) - \eta(x) - \left(\frac{k_{\mathrm{L}}(x)}{n\bar{f}(x)} \right)^{2/d} a(x) = o\left(\left(\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)} \right)^{2/d} \ell\big(\bar{f}(x_0)\big) \right),$$

uniformly for $P \in \mathcal{P}_{d,\theta}$, $k_{\mathrm{L}} \in K_{\beta,\tau}$, $x_0 \in \mathcal{S}_n$ and $|t| < \epsilon_n$. Write

$$\begin{aligned}\mu_n(x) - \eta(x) &= \frac{1}{k_{\mathrm{L}}(x)} \sum_{i=1}^{k_{\mathrm{L}}(x)} \mathbb{E}\{\eta(X_{(i)}) - \eta(x)\} \\ &= \frac{1}{k_{\mathrm{L}}(x)} \sum_{i=1}^{k_{\mathrm{L}}(x)} \mathbb{E}\{(X_{(i)} - x)^T \dot{\eta}(x)\} \\ &\quad + \frac{1}{2}\mathbb{E}\{(X_{(i)} - x)^T \ddot{\eta}(x)(X_{(i)} - x)\} + R_1,\end{aligned}$$

where we show in Step 7 that

$$|R_1| = o\left\{\left(\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\right)^{2/d}\right\} \tag{10}$$

uniformly for $P \in \mathcal{P}_{d,\theta}$, $k_{\mathrm{L}} \in K_{\beta,\tau}$, $x_0 \in \mathcal{S}_n$ and $|t| < \epsilon_n$.

The density of $X_{(i)} - x$ at $u \in \mathbb{R}^d$ is given by

$$\begin{aligned}f_{(i)}(u) &:= n\bar{f}(x+u)\binom{n-1}{i-1} p_{\|u\|}^{i-1}(1 - p_{\|u\|})^{n-i} \\ &= n\bar{f}(x+u) p_{\|u\|}^{n-1}(i-1),\end{aligned} \tag{11}$$

where $p_{\|u\|} = p_{\|u\|}(x)$ and $p_{\|u\|}^{n-1}(i-1)$ denotes the probability that a $\mathrm{Bin}(n-1, p_{\|u\|})$ random variable equals $i-1$. Now let

$$r_n = r_n(x) := \left\{\frac{2k_{\mathrm{L}}(x)}{(n-1)\bar{f}(x)a_d}\right\}^{1/d}. \tag{12}$$

We show in Step 7 that

$$\begin{aligned}R_2 &:= \sup_{P \in \mathcal{P}_{d,\theta}} \sup_{k_{\mathrm{L}} \in K_{\beta,\tau}} \sup_{x_0 \in \mathcal{S}_n} \sup_{|t| < \epsilon_n} \mathbb{E}\{\|X_{(k_{\mathrm{L}})} - x\|^2 \mathbb{1}_{\{\|X_{(k_{\mathrm{L}})} - x\| \geq r_n\}}\} \\ &= O(n^{-M}),\end{aligned} \tag{13}$$

for each $M > 0$, as $n \to \infty$. It follows from (11) and (13), together with the upper bound on $\sup_{x \in \mathcal{S}^{2\epsilon_0}} \|\dot{\eta}(x)\|$ in (A.3) that

$$\begin{aligned}&\mathbb{E}\{(X_{(i)} - x)^T \dot{\eta}(x)\} \\ &= \int_{B_{r_n}(0)} \dot{\eta}(x)^T u n\{\bar{f}(x+u) - \bar{f}(x)\} p_{\|u\|}^{n-1}(i-1)\, du + O(n^{-M}),\end{aligned}$$

uniformly for $P \in \mathcal{P}_{d,\theta}$, $k_{\mathrm{L}} \in K_{\beta,\tau}$, $i \in \{1, \ldots, k_{\mathrm{L}}\}$, $x_0 \in \mathcal{S}_n$ and $|t| < \epsilon_n$. Similarly, using the upper bound on $\sup_{x \in \mathcal{S}^{2\epsilon_0}} \|\ddot{\eta}(x)\|_{\mathrm{op}}$ in (A.3),

$$\begin{aligned}&\mathbb{E}\{(X_{(i)} - x)^T \ddot{\eta}(x)(X_{(i)} - x)\} \\ &= \int_{B_{r_n}(0)} u^T \ddot{\eta}(x) u n \bar{f}(x+u) p_{\|u\|}^{n-1}(i-1)\, du + O(n^{-M}),\end{aligned}$$

uniformly for $P \in \mathcal{P}_{d,\theta}$, $k_{\mathrm{L}} \in K_{\beta,\tau}$, $i \in \{1, \ldots, k_{\mathrm{L}}\}$, $x_0 \in \mathcal{S}_n$ and $|t| < \epsilon_n$. Hence, summing over i, we see that

$$\frac{1}{k_{\mathrm{L}}} \sum_{i=1}^{k_{\mathrm{L}}} \mathbb{E}\{(X_{(i)} - x)^T \dot{\eta}(x)\} + \frac{1}{2k_{\mathrm{L}}} \sum_{i=1}^{k_{\mathrm{L}}} \mathbb{E}\{(X_{(i)} - x)^T \ddot{\eta}(x)(X_{(i)} - x)\}$$

$$= \int_{B_{r_n}(0)} \left[\dot{\eta}(x)^T un\{\bar{f}(x+u) - \bar{f}(x)\} + \frac{1}{2} u^T \ddot{\eta}(x) un \bar{f}(x+u) \right] q_{\|u\|}^{n-1}(k_{\mathrm{L}})\, du + O(n^{-M}),$$

uniformly for $P \in \mathcal{P}_{d,\theta}$, $k_{\mathrm{L}} \in K_{\beta,\tau}$, $i \in \{1, \ldots, k_{\mathrm{L}}\}$, $x_0 \in \mathcal{S}_n$ and $|t| < \epsilon_n$, where $q_{\|u\|}^{n-1}(k_{\mathrm{L}})$ denotes the probability that a $\mathrm{Bin}(n-1, p_{\|u\|})$ random variable is less than k_{L}. Let $n_0 \in \mathbb{N}$ be large enough that

$$\epsilon_n + \sup_{x_0 \in \mathcal{S}_n} \sup_{|t| < \epsilon_n} r_n(x) < \epsilon_0$$

for $n \geq n_0$. That this is possible follows from the fact that, for $\epsilon_n < \epsilon_0$,

$$\begin{aligned} &\sup_{P \in \mathcal{P}_{d,\theta}} \sup_{k_{\mathrm{L}} \in K_{\beta,\tau}} \sup_{x_0 \in \mathcal{S}_n} \sup_{|t| < \epsilon_n} \max\left\{ \left| \frac{k_{\mathrm{L}}(x)}{k_{\mathrm{L}}(x_0)} - 1 \right|, \left| \frac{\bar{f}(x)}{\bar{f}(x_0)} - 1 \right| \right\} \\ &\leq \sup_{P \in \mathcal{P}_{d,\theta}} \max\left\{ \tau, c_n \epsilon_n + \frac{c_n \epsilon_n^2}{2} \right\} \qquad (14) \\ &\leq \max\left\{ \tau, \frac{1}{\beta^{1/2} \log^{1/2}(n-1)} + \frac{1}{2\beta \log(n-1)} \right\} \to 0. \end{aligned}$$

By a Taylor expansion of $\bar{f}$ and assumption (A.2), for all $x_0 \in \mathcal{S}_n$, $|t| < \epsilon_n$, $\|u\| < r_n$ and $n \geq n_0$,

$$\begin{aligned} |\bar{f}(x+u) - \bar{f}(x) - u^T \dot{\bar{f}}(x)| &\leq \frac{\|u\|^2}{2} \sup_{s \in B_{\|u\|}(0)} \|\ddot{\bar{f}}(x+s)\|_{\mathrm{op}} \\ &\leq \frac{\|u\|^2}{2} \bar{f}(x_0) \ell(\bar{f}(x_0)). \end{aligned}$$

Hence, for $x_0 \in \mathcal{S}_n$, $|t| < \epsilon_n$, $r < r_n$ and $n \geq n_0$,

$$\begin{aligned} |p_r(x) - \bar{f}(x) a_d r^d| &\leq \int_{B_r(0)} |\bar{f}(x+u) - \bar{f}(x) - u^T \dot{\bar{f}}(x)|\, du \\ &\leq \frac{1}{2} \bar{f}(x_0) \ell(\bar{f}(x_0)) \int_{B_r(0)} \|u\|^2\, du \qquad (15) \\ &= \frac{d a_d}{2(d+2)} \bar{f}(x_0) \ell(\bar{f}(x_0)) r^{d+2}. \end{aligned}$$

Now, for $v \in B_1(0)$, $x_0 \in \mathcal{S}_n$, $|t| < \epsilon_n$ and $n \geq n_0$,

$$\begin{aligned} k_{\mathrm{L}}(x) - (n-1) p_{\|v\| r_n} &= k_{\mathrm{L}}(x) - (n-1) \bar{f}(x) a_d \|v\|^d r_n^d + R_3 \\ &= k_{\mathrm{L}}(x)(1 - 2\|v\|^d) + R_3, \end{aligned}$$

where

$$\begin{aligned} |R_3| &\leq \frac{d a_d (n-1) \bar{f}(x_0) \ell(\bar{f}(x_0)) \|v\|^{d+2} r_n^{d+2}}{2(d+2)} \\ &\leq \frac{2^{2/d}\, d k_{\mathrm{L}}(x)}{a_d^{2/d} (d+2) \log^2(\frac{n-1}{k_{\mathrm{L}}(x_0)})} \left(\frac{\bar{f}(x_0)}{\bar{f}(x)} \right)^{1+2/d} \left(\frac{k_{\mathrm{L}}(x)}{k_{\mathrm{L}}(x_0)} \right)^{2/d}. \end{aligned}$$

It follows from (14) that there exists $n_1 \in \mathbb{N}$ such that, for all $x_0 \in \mathcal{S}_n$, $|t| < \epsilon_n$, $\|v\|^d \in (0, 1/2 - 1/\log((n-1)/k_{\mathrm{L}}(x_0))]$ and $n \geq n_1$,

$$k_{\mathrm{L}}(x) - (n-1)p_{\|v\|r_n} \geq \frac{k_{\mathrm{L}}(x)}{\log((n-1)/k_{\mathrm{L}}(x_0))}.$$

Similarly, for all $\|v\|^d \in [1/2 + 1/\log((n-1)/k_{\mathrm{L}}(x_0)), 1)$ and $n \geq n_1$,

$$(n-1)p_{\|v\|r_n} - k_{\mathrm{L}}(x) \geq \frac{k_{\mathrm{L}}(x)}{\log((n-1)/k_{\mathrm{L}}(x_0))}.$$

Hence, by Bernstein's inequality, we have that for each $M > 0$,

$$\sup_{P \in \mathcal{P}_{d,\theta}} \sup_{k_{\mathrm{L}} \in K_{\beta,\tau}} \sup_{x_0 \in \mathcal{S}_n} \sup_{|t| < \epsilon_n} \sup_{\|v\|^d \in (0, \frac{1}{2} - \frac{1}{\log((n-1)/k_{\mathrm{L}}(x_0))}]} 1 - q^{n-1}_{\|v\|r_n}\big(k_{\mathrm{L}}(x)\big) = O\big(n^{-M}\big),$$

and

$$\sup_{P \in \mathcal{P}_{d,\theta}} \sup_{k_{\mathrm{L}} \in K_{\beta,\tau}} \sup_{x_0 \in \mathcal{S}_n} \sup_{|t| < \epsilon_n} \sup_{\|v\|^d \in [\frac{1}{2} - \frac{1}{\log((n-1)/k_{\mathrm{L}}(x_0))}, 1)} q^{n-1}_{\|v\|r_n}\big(k_{\mathrm{L}}(x)\big) = O\big(n^{-M}\big). \tag{16}$$

We conclude that

$$\begin{aligned}
&\frac{1}{k_{\mathrm{L}}(x)} \int_{B_{r_n}(0)} \bigg[\dot{\eta}(x)^T u n \{\bar{f}(x+u) - \bar{f}(x)\} \\
&\qquad + \frac{1}{2} u^T \ddot{\eta}(x) u n \bar{f}(x+u) \bigg] q^{n-1}_{\|u\|}\big(k_{\mathrm{L}}(x)\big)\, du \\
&= \frac{1}{k_{\mathrm{L}}(x)} \int_{B_{2^{-1/d} r_n}(0)} \bigg[\dot{\eta}(x)^T u n \{\bar{f}(x+u) - \bar{f}(x)\} \\
&\qquad + \frac{1}{2} u^T \ddot{\eta}(x) u n \bar{f}(x+u) \bigg] du + R_{41} \\
&= \bigg(\frac{k_{\mathrm{L}}(x)}{n}\bigg)^{2/d} \frac{\sum_{j=1}^d \{\eta_j(x)\bar{f}_j(x) + \frac{1}{2}\eta_{jj}(x)\bar{f}(x)\}}{(d+2)a_d^{2/d}\bar{f}(x)^{1+2/d}} + R_{41} + R_{42} \\
&= \bigg(\frac{k_{\mathrm{L}}(x)}{n\bar{f}(x)}\bigg)^{2/d} a(x) + R_{41} + R_{42},
\end{aligned} \tag{17}$$

where

$$|R_{41}| + |R_{42}| = o\bigg(\bigg(\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\bigg)^{2/d} \ell\big(\bar{f}(x_0)\big)\bigg),$$

uniformly for $P \in \mathcal{P}_{d,\theta}$, $k_{\mathrm{L}} \in K_{\beta,\tau}$, $x_0 \in \mathcal{S}_n$ and $|t| < \epsilon_n$.

Step 2: Recall that $\hat{\sigma}_n^2(x, x^n) = \mathrm{Var}\{\hat{S}_n(x) | X^n = x^n\}$. We show that

$$\bigg|\hat{\sigma}_n^2\big(x, X^n\big) - \frac{1}{4k_{\mathrm{L}}}\bigg| = o_p(1/k_{\mathrm{L}}), \tag{18}$$

uniformly for $P \in \mathcal{P}_{d,\theta}$, $k_{\mathrm{L}} \in K_{\beta,\tau}$, $x_0 \in \mathcal{S}_n$ and $|t| < \epsilon_n$. Recall that

$$\hat{\sigma}_n^2\big(x, X^n\big) = \frac{1}{k_{\mathrm{L}}^2} \sum_{i=1}^{k_{\mathrm{L}}} \eta(X_{(i)})\{1 - \eta(X_{(i)})\}.$$

Let $n_2 \in \mathbb{N}$ be large enough that $1 - c_n\epsilon_n - \frac{d+1}{d+2}c_n\epsilon_n^2 \geq \epsilon_0$ for $n \geq n_2$. Then for $n \geq \max\{n_0, n_2\}$, $P \in \mathcal{P}_{d,\theta}$, $r < \epsilon_n$, $x_0 \in \mathcal{S}_n$ and $|t| < \epsilon_n$, we have by (A.2) and a very similar argument to that in (15) that

$$p_r(x) \geq \epsilon_0 a_d r^d \bar{f}(x_0) \geq \epsilon_0 a_d r^d \delta_n(x_0). \tag{19}$$

Now suppose that $z_1, \dots, z_N \in \mathcal{R}_n \cup \mathcal{S}_n^{\epsilon_n}$ are such that $\|z_j - z_\ell\| \geq \epsilon_n/6$ for all $j \neq \ell$, but $\sup_{x \in \mathcal{R}_n \cup \mathcal{S}_n^{\epsilon_n}} \min_{j=1,\dots,N} \|x - z_j\| < \epsilon_n/6$. We have by (A.2) that

$$1 = P_X(\mathbb{R}^d) \geq \sum_{j=1}^{N} p_{\epsilon_n/12}(z_j) \geq \frac{N\epsilon_0 a_d \beta^{d/2} \log^{d/2}(n-1)}{12^d (n-1)^{1-\beta}}.$$

For each $j = 1, \dots, N$, choose

$$z_j' \in \operatorname*{argmax}_{z \in B_{z_j}(\epsilon_n/6) \cap (\mathcal{R}_n \cup \mathcal{S}_n^{\epsilon_n})} k_{\mathrm{L}}(z).$$

Now, given $x \in \mathcal{R}_n \cup \mathcal{S}_n^{\epsilon_n}$, let $j_0 := \operatorname{argmin}_j \|x - z_j\|$, so that $B_{\epsilon_n/6}(z_{j_0}') \subseteq B_{\epsilon_n/2}(x)$. Thus, if there are at least $k_{\mathrm{L}}(z_j')$ points among $\{x_1, \dots, x_n\}$ inside each of the balls $B_{\epsilon_n/6}(z_j')$, then for every $x \in \mathcal{R}_n \cup \mathcal{S}_n^{\epsilon_n}$ there are at least $k_{\mathrm{L}}(x)$ of them in $B_{\epsilon_n/2}(x)$. Moreover by (14), (19) and (A.2),

$$\min_{j=1,\dots,N} \{np_{\epsilon_n/6}(z_j') - 2k_{\mathrm{L}}(z_j')\} \geq (n-1)^{\beta}$$

for all $P \in \mathcal{P}_{d,\theta}$, $k_{\mathrm{L}} \in K_{\beta,\tau}$ and $n \geq n_3$, say. Define $A_{k_{\mathrm{L}}} := \{\|X_{(k_{\mathrm{L}})}(x) - x\| < \epsilon_n/2 \text{ for all } x \in \mathcal{R}_n \cup \mathcal{S}_n^{\epsilon_n}\}$. Then by a standard binomial tail bound (Shorack and Wellner) (1986, equation (6), p. 440), for $n \geq n_3$ and any $M > 0$,

$$\begin{aligned}
\mathbb{P}(A_{k_{\mathrm{L}}}^c) &= \mathbb{P}\Big\{ \sup_{x \in \mathcal{R}_n \cup \mathcal{S}_n^{\epsilon_n}} \|X_{(k_{\mathrm{L}}(x))}(x) - x\| \geq \epsilon_n/2 \Big\} \\
&\leq \mathbb{P}\Big\{ \max_{j=1,\dots,N} \|X_{(k_{\mathrm{L}}(z_j))}(z_j') - z_j'\| \geq \epsilon_n/6 \Big\} \\
&\leq \sum_{j=1}^{N} \mathbb{P}\{\|X_{(k_{\mathrm{L}}(z_j))}(z_j') - z_j'\| \geq \epsilon_n/6\} \\
&\leq N \max_{j=1,\dots,N} \exp\Big(-\frac{1}{2} np_{\epsilon_n/6}(z_j') + k_{\mathrm{L}}(z_j')\Big) = O(n^{-M}),
\end{aligned} \tag{20}$$

uniformly for $P \in \mathcal{P}_{d,\theta}$ and $k_{\mathrm{L}} \in K_{\beta,\tau}$. Now, for $3\epsilon_n/2 < 2\epsilon_0$,

$$\begin{aligned}
&\sup_{P \in \mathcal{P}_{d,\theta}} \sup_{k_{\mathrm{L}} \in K_{\beta,\tau}} \sup_{x_0 \in \mathcal{S}_n} \sup_{|t| < \epsilon_n} \sup_{x^n \in A_{k_{\mathrm{L}}}} \max_{1 \leq i \leq k_{\mathrm{L}}(x)} \big|\eta(x_{(i)}(x)) - 1/2\big| \\
&\quad \leq 3M_0 \sup_{P \in \mathcal{P}_{d,\theta}} \sup_{k_{\mathrm{L}} \in K_{\beta,\tau}} \frac{\epsilon_n}{2} \leq \frac{3M_0}{2\beta^{1/2} \log^{1/2}(n-1)} \to 0.
\end{aligned}$$

It follows that

$$\sup_{x^n \in A_{k_{\mathrm{L}}}} \Big| \frac{1}{k_{\mathrm{L}}(x)^2} \sum_{i=1}^{k_{\mathrm{L}}(x)} \eta(x_{(i)}(x))\{1 - \eta(x_{(i)}(x))\} - \frac{1}{4k_{\mathrm{L}}(x)} \Big| = o\Big(\frac{1}{k_{\mathrm{L}}(x)}\Big) \tag{21}$$

as $n \to \infty$, uniformly for $P \in \mathcal{P}_{d,\theta}$, $k_{\mathrm{L}} \in K_{\beta,\tau}$, $x_0 \in \mathcal{S}_n$ and $|t| < \epsilon_n$. The claim (18) follows from (20) and (21).

Step 3: In this step, we emphasise the dependence of $\hat{\mu}_n(x, x^n) = \mathbb{E}\{\hat{S}_n(x) | X^n = x^n\}$ on k_{L} by writing it as $\hat{\mu}_n^{(k_{\mathrm{L}})}(x, x^n)$. We show that

$$\operatorname{Var}\{\hat{\mu}_n^{(k_{\mathrm{L}})}(x, X^n)\} = O\Big\{ \frac{1}{k_{\mathrm{L}}(x_0)} \Big(\frac{k_{\mathrm{L}}(x_0)}{n \bar{f}(x_0)} \Big)^{2/d} \Big\} \tag{22}$$

uniformly for $P \in \mathcal{P}_{d,\theta}$, $k_{\mathrm{L}} \in K_{\beta,\tau}$, $x_0 \in \mathcal{S}_n$ and $|t| < \epsilon_n$. We will write $X^{n,j} := (X_1, \ldots, X_{j-1}X_{j+1}, \ldots, X_n)$, considered as a random $d \times (n-1)$ matrix, so that

$$\hat{\mu}_n^{(k_{\mathrm{L}})}(x, X^n) - \hat{\mu}_{n-1}^{(k_{\mathrm{L}})}(x, X^{n,(i)}) = \frac{1}{k_{\mathrm{L}}}\{\eta(X_{(i)}) - \eta(X_{(k_{\mathrm{L}}+1)})\}\mathbb{1}_{\{i \le k_{\mathrm{L}}\}}.$$

It follows from the Efron–Stein inequality (e.g., Boucheron, Lugosi and Massart (2013), Theorem 3.1) that

$$\begin{aligned}
\mathrm{Var}\{\hat{\mu}_n^{(k_{\mathrm{L}})}(x, X^n)\} &\le \sum_{i=1}^{n} \mathbb{E}\big[\{\hat{\mu}_n^{(k_{\mathrm{L}})}(x, X^n) - \hat{\mu}_{n-1}^{(k_{\mathrm{L}})}(x, X^{n,(i)})\}^2\big] \\
&= \frac{1}{k_{\mathrm{L}}^2} \sum_{i=1}^{k_{\mathrm{L}}} \mathbb{E}\big[\{\eta(X_{(i)}) - \eta(X_{(k_{\mathrm{L}}+1)})\}^2\big] \qquad (23) \\
&\le \frac{2}{k_{\mathrm{L}}^2} \sum_{i=1}^{k_{\mathrm{L}}} \mathbb{E}\big[\{\eta(X_{(i)}) - \eta(x)\}^2 + \{\eta(X_{(k_{\mathrm{L}}+1)}) - \eta(x)\}^2\big].
\end{aligned}$$

Recall the definition of r_n given in (12). Now observe that, for $\max(\epsilon_n, r_n) \le \epsilon_0$ and all $M > 0$ we have that

$$\begin{aligned}
&\max_{i \in \{1,\ldots,k_{\mathrm{L}}+1\}} \mathbb{E}\big[\{\eta(X_{(i)}) - \eta(x)\}^2\big] \\
&\le \max_{i \in \{1,\ldots,k_{\mathrm{L}}+1\}} \mathbb{E}\big[\{\eta(X_{(i)}) - \eta(x)\}^2 \mathbb{1}_{\{\|X_{(i)} - x\| \le r_n\}}\big] \qquad (24) \\
&\quad + \mathbb{P}\big(\|X_{(k_{\mathrm{L}}+1)} - x\| > r_n\big) \\
&\le r_n^2 M_0 + O\big(n^{-M}\big),
\end{aligned}$$

uniformly for $P \in \mathcal{P}_{d,\theta}$, $k_{\mathrm{L}} \in K_{\beta,\tau}$, $x_0 \in \mathcal{S}_n$ and $|t| < \epsilon_n$. The final inequality here follows from similar arguments to those used to bound R_1. Now (22) follows from (23) and (24).

Step 4: We show that

$$\sup_{P \in \mathcal{P}_{d,\theta}} \sup_{k_{\mathrm{L}} \in K_{\beta,\tau}} \sup_{x \in \mathcal{R}_n \setminus \mathcal{S}^{\epsilon_n}} \big|\mathbb{P}\big\{\hat{S}_n(x) < 1/2\big\} - \mathbb{1}_{\{\eta(x) < 1/2\}}\big| = O\big(n^{-M}\big),$$

for each $M > 0$, as $n \to \infty$. First, by (A.3) and Proposition 2 in Section 7.2 in the online supplement, there exists $c_0 > 0$ such that for every $r \in (0, \epsilon_0]$, $P \in \mathcal{P}_{d,\theta}$ and $k_{\mathrm{L}} \in K_{\beta,\tau}$,

$$\inf_{x \in \mathcal{R}_n \setminus \mathcal{S}^r} |\eta(x) - 1/2| \ge c_0 \min\Big\{r, \inf_{x \in \mathcal{R}_n \setminus \mathcal{S}^{\epsilon_0}} \delta_n(x)^{\beta/2}\Big\}.$$

Hence, on the event $A_{k_{\mathrm{L}}}$, for $\epsilon_n < \epsilon_0$ and $x \in \mathcal{R}_n \setminus \mathcal{S}^{\epsilon_n}$, all of the k_{L} nearest neighbours of x are on the same side of $\mathcal{S}$, so

$$\begin{aligned}
|\hat{\mu}_n(x, X^n) - 1/2| &= \Big|\frac{1}{k_{\mathrm{L}}} \sum_{i=1}^{k_{\mathrm{L}}} \eta(X_{(i)}) - 1/2\Big| \\
&\ge \inf_{z \in B_{\epsilon_n/2}(x)} |\eta(z) - 1/2| \\
&\ge c_0 \min\Big\{\frac{\epsilon_n}{2}, \inf_{x \in \mathcal{R}_n \setminus \mathcal{S}^{\epsilon_0}} \delta_n(x)^{\beta/2}\Big\}.
\end{aligned}$$

Now, conditional on X^n, $\hat{S}_n(x)$ is the sum of $k_{\mathrm{L}}(x)$ independent terms. Therefore, by Hoeffding's inequality,

$$\sup_{P \in \mathcal{P}_{d,\theta}} \sup_{k_{\mathrm{L}} \in K_{\beta,\tau}} \sup_{x \in \mathcal{R}_n \setminus \mathcal{S}^{\epsilon_n}} \big|\mathbb{P}\big\{\hat{S}_n(x) < 1/2\big\} - \mathbb{1}_{\{\eta(x) < 1/2\}}\big|$$

$$
\begin{aligned}
&= \sup_{P\in\mathcal{P}_{d,\theta}} \sup_{k_{\mathrm{L}}\in K_{\beta,\tau}} \sup_{x\in\mathcal{R}_n\setminus\mathcal{S}^{\epsilon_n}} \bigl|\mathbb{E}\{\mathbb{P}\{\hat{S}_n(x) < 1/2|X^n\} - \mathbb{1}_{\{\eta(x)<1/2\}}\bigr| \\
&\le \sup_{P\in\mathcal{P}_{d,\theta}} \sup_{k_{\mathrm{L}}\in K_{\beta,\tau}} \sup_{x\in\mathcal{R}_n\setminus\mathcal{S}^{\epsilon_n}} \bigl\{\mathbb{E}\bigl[e^{-2k_{\mathrm{L}}\{\hat{\mu}_n(x,X^n)-1/2\}^2}\mathbb{1}_{A_{k_{\mathrm{L}}}}\bigr] + \mathbb{P}\bigl(A_{k_{\mathrm{L}}}^c\bigr)\bigr\} \\
&= O\bigl(n^{-M}\bigr)
\end{aligned}
$$

for every $M > 0$. This completes Step 4.

Step 5: It is now convenient to be more explicit in our notation, by writing $x_0^t := x_0 + t\dot{\eta}(x_0)/\|\dot{\eta}(x_0)\|$. We also let

$$\psi(x) := \{2\eta(x) - 1\}\bar{f}(x) = \pi_1 f_1(x) - \pi_0 f_0(x).$$

Recall that $\mathcal{S}_n := \mathcal{S} \cap \mathcal{R}_n$ and let

$$W_{n,2} := \biggl(\int_{\mathcal{S}^{\epsilon_n}\cap\mathcal{R}_n} - \int_{\mathcal{S}_n^{\epsilon_n}}\biggr)\psi(x)\bigl[\mathbb{P}\{\hat{S}_n(x) < 1/2\} - \mathbb{1}_{\{\eta(x)<1/2\}}\bigr]\,dx.$$

We show that

$$
\begin{aligned}
&\int_{\mathcal{S}^{\epsilon_n}\cap\mathcal{R}_n} \psi(x)\bigl[\mathbb{P}\{\hat{S}_n(x) < 1/2\} - \mathbb{1}_{\{\eta(x)<1/2\}}\bigr]\,dx \\
&\quad= \int_{\mathcal{S}_n}\int_{-\epsilon_n}^{\epsilon_n} \psi\bigl(x_0^t\bigr)\bigl[\mathbb{P}\{\hat{S}_n\bigl(x_0^t\bigr) < 1/2\} - \mathbb{1}_{\{t<0\}}\bigr]\,dt\,d\,\mathrm{Vol}^{d-1}(x_0)\{1 + o(1)\} \\
&\qquad+ W_{n,2}
\end{aligned}
$$

uniformly for $P \in \mathcal{P}_{d,\theta}$ and $k_{\mathrm{L}} \in K_{\beta,\tau}$, and that for all $n \ge 2$,

$$\sup_{P\in\mathcal{P}_{d,\theta}} \sup_{k_{\mathrm{L}}\in K_{\beta,\tau}} \frac{|W_{n,2}|}{P_X((\partial\mathcal{R}_n)^{\epsilon_n}\cap\mathcal{S}^{\epsilon_n})} \le 1. \tag{25}$$

Now by Proposition 3 in Section 7.2 of the online supplement, for $\epsilon_n \le \epsilon_0$, the map $x(x_0, t) = x_0^t$ is a diffeomorphism from $\mathcal{S}_n \times (-\epsilon_n, \epsilon_n)$ to $\mathcal{S}_n^{\epsilon_n}$, where

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

Furthermore, for such n, and $|t| < \epsilon_n$, $\operatorname{sgn}\{\eta(x_0^t) - 1/2\} = \operatorname{sgn}(t)$. It follows from this and (62) in Section 7.3 in the online supplement that

$$
\begin{aligned}
&\int_{\mathcal{S}^{\epsilon_n}\cap\mathcal{R}_n} \psi(x)\bigl[\mathbb{P}\{\hat{S}_n(x) < 1/2\} - \mathbb{1}_{\{\eta(x)<1/2\}}\bigr]\,dx \\
&\quad= \int_{\mathcal{S}_n^{\epsilon_n}} \psi(x)\bigl[\mathbb{P}\{\hat{S}_n(x) < 1/2\} - \mathbb{1}_{\{\eta(x)<1/2\}}\bigr]\,dx + W_{n,2} \\
&\quad= \int_{\mathcal{S}_n}\int_{-\epsilon_n}^{\epsilon_n} \det(I + tB)\psi\bigl(x_0^t\bigr)\bigl[\mathbb{P}\{\hat{S}_n\bigl(x_0^t\bigr) < 1/2\} - \mathbb{1}_{\{t<0\}}\bigr]\,dt\,d\,\mathrm{Vol}^{d-1}(x_0) \\
&\qquad+ W_{n,2},
\end{aligned}
$$

where B is defined in (55) in the online supplement, and $\det(I + tB) = 1 + o(1)$ as $n \to \infty$, uniformly for $P \in \mathcal{P}_{d,\theta}$, $x_0 \in \mathcal{S}$ and $t \in (-\epsilon_n, \epsilon_n)$. Now observe that $(\mathcal{S}^{\epsilon_n} \cap \mathcal{R}_n) \setminus \mathcal{S}_n^{\epsilon_n} \subseteq (\partial\mathcal{R}_n)^{\epsilon_n} \cap \mathcal{S}^{\epsilon_n}$ and $\mathcal{S}_n^{\epsilon_n} \setminus (\mathcal{S}^{\epsilon_n} \cap \mathcal{R}_n) \subseteq (\partial\mathcal{R}_n)^{\epsilon_n} \cap \mathcal{S}^{\epsilon_n}$. We deduce from this and the definition of $W_{n,2}$ that (25) holds.

Step 6: The last step in the main argument is to show that

$$\begin{aligned}\tilde{W}_{n,1} &:= \int_{\mathcal{S}_n}\int_{-\epsilon_n}^{\epsilon_n} \psi(x_0^t)\big[\mathbb{P}\big\{\hat{S}_n(x_0^t) < 1/2\big\} - \mathbb{1}_{\{t<0\}}\big]\,dt\,d\,\mathrm{Vol}^{d-1}(x_0) \\ &\quad - \int_{\mathcal{S}_n} \frac{\bar{f}(x_0)}{\|\dot{\eta}(x_0)\|}\bigg\{\frac{1}{4k_{\mathrm{L}}(x_0)} + \bigg(\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\bigg)^{4/d} a(x_0)^2\bigg\}\,d\,\mathrm{Vol}^{d-1}(x_0) \\ &= o\big(\gamma_n(k_{\mathrm{L}})\big)\end{aligned}$$

as $n \to \infty$, uniformly for $P \in \mathcal{P}_{d,\theta}$ and $k_{\mathrm{L}} \in K_{\beta,\tau}$. First, observe that

$$\begin{aligned}&\int_{\mathcal{S}_n}\int_{-\epsilon_n}^{\epsilon_n} \psi(x_0^t)\big[\mathbb{P}\big\{\hat{S}_n(x_0^t) < 1/2\big\} - \mathbb{1}_{\{t<0\}}\big]\,dt\,d\,\mathrm{Vol}^{d-1}(x_0) \\ &\quad = \int_{\mathcal{S}_n}\int_{-\epsilon_n}^{\epsilon_n} t\|\dot{\psi}(x_0)\|\big[\mathbb{P}\big\{\hat{S}_n(x_0^t) < 1/2\big\} - \mathbb{1}_{\{t<0\}}\big]\,dt\,d\,\mathrm{Vol}^{d-1}(x_0)\big\{1 + o(1)\big\},\end{aligned}$$

uniformly for $P \in \mathcal{P}_{d,\theta}$ and $k_{\mathrm{L}} \in K_{\beta,\tau}$. Now, write $\mathbb{P}\{\hat{S}_n(x_0^t) < 1/2\} - \mathbb{1}_{\{t<0\}} = \mathbb{E}[\mathbb{P}\{\hat{S}_n(x_0^t) < 1/2|X^n\} - \mathbb{1}_{\{t<0\}}]$. Note that, given X^n, $\hat{S}_n(x) = \frac{1}{k_{\mathrm{L}}(x)}\sum_{i=1}^{k_{\mathrm{L}}(x)} \mathbb{1}_{\{Y_{(i)}=1\}}$ is the sum of $k_{\mathrm{L}}(x)$ independent Bernoulli variables, satisfying $\mathbb{P}(Y_{(i)} = 1|X^n) = \eta(X_{(i)})$. Let Φ be the standard normal distribution function, and let

$$\begin{aligned}\hat{\theta}(x) \equiv \hat{\theta}_n(x) &:= -\big\{\hat{\mu}_n\big(x, X^n\big) - 1/2\big\}/\hat{\sigma}_n\big(x, X^n\big), \\ \bar{\theta}(x_0, t) \equiv \bar{\theta}_n(x_0, t) &:= -2k_{\mathrm{L}}(x_0)^{1/2}\bigg\{t\|\dot{\eta}(x_0)\| + \bigg(\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\bigg)^{2/d} a(x_0)\bigg\}.\end{aligned}$$

We can write

$$\begin{aligned}&\int_{-\epsilon_n}^{\epsilon_n} t\|\dot{\psi}(x_0)\|\big[\mathbb{P}\big\{\hat{S}_n(x_0^t) < 1/2\big\} - \mathbb{1}_{\{t<0\}}\big]\,dt \\ &\quad = \int_{-\epsilon_n}^{\epsilon_n} t\|\dot{\psi}(x_0)\|\mathbb{E}\big\{\Phi\big(\hat{\theta}(x_0^t)\big) - \mathbb{1}_{\{t<0\}}\big\}\,dt + R_5(x_0) \\ &\quad = \int_{-\epsilon_n}^{\epsilon_n} t\|\dot{\psi}(x_0)\|\big\{\Phi\big(\bar{\theta}(x_0, t)\big) - \mathbb{1}_{\{t<0\}}\big\}\,dt + R_5(x_0) + R_6(x_0),\end{aligned}$$

where we show in Step 7 that

$$\bigg|\int_{\mathcal{S}_n}\big\{R_5(x_0) + R_6(x_0)\big\}\,d\,\mathrm{Vol}^{d-1}(x_0)\bigg| = o\big(\gamma_n(k_{\mathrm{L}})\big) \tag{26}$$

uniformly for $P \in \mathcal{P}_{d,\theta}$ and $k_{\mathrm{L}} \in K_{\beta,\tau}$. Then, substituting $u = 2k_{\mathrm{L}}(x_0)^{1/2}t$, we see that

$$\begin{aligned}&\int_{-\epsilon_n}^{\epsilon_n} t\|\dot{\psi}(x_0)\|\big[\Phi\big(\bar{\theta}(x_0, t)\big) - \mathbb{1}_{\{t<0\}}\big]\,dt \\ &\quad = \frac{1}{4k_{\mathrm{L}}(x_0)}\int_{-2k_{\mathrm{L}}(x_0)^{1/2}\epsilon_n}^{2k_{\mathrm{L}}(x_0)^{1/2}\epsilon_n} u\|\dot{\psi}(x_0)\|\bigg\{\Phi\bigg(\bar{\theta}\bigg(x_0, \frac{u}{2k_{\mathrm{L}}(x_0)^{1/2}}\bigg)\bigg) - \mathbb{1}_{\{u<0\}}\bigg\}\,du \\ &\quad = \bigg\{\frac{\bar{f}(x_0)}{4k_{\mathrm{L}}(x_0)\|\dot{\eta}(x_0)\|} + \bigg(\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\bigg)^{4/d}\frac{\bar{f}(x_0)a(x_0)^2}{\|\dot{\eta}(x_0)\|}\bigg\}\big\{1 + o(1)\big\},\end{aligned}$$

uniformly for $P \in \mathcal{P}_{d,\theta}$, $k_{\mathrm{L}} \in K_{\beta,\tau}$ and $x_0 \in \mathcal{S}_n$. The conclusion follows by integrating with respect to $d\,\mathrm{Vol}^{d-1}$ over $\mathcal{S}_n$.

Step 7: It remains to bound the error terms R_1, R_2, R_5 and R_6—these bounds are presented in Section 5 of the supplementary material. □

6.2. *Proof of Theorem 1.*

PROOF OF THEOREM 1. Let $k \in K_\beta$, and note that since $k_{\mathrm{L}}(x) = k$ is constant, we have that $c_n = \ell(k/(n-1))$, and $\delta_n = \frac{k}{n-1} c_n^d \log^d(\frac{n-1}{k})$. Now let

$$\mathcal{R}_n = \{x \in \mathbb{R}^d : \bar{f}(x) > \delta_n\} \cap \mathcal{X}_{\bar{f}},$$

and observe that by Berrett, Samworth and Yuan ((2019), Lemma 10(i)), for $P \in \mathcal{P}_{d,\theta}$,

$$\|\bar{f}\|_\infty^\rho \geq \frac{\rho^\rho d^d}{a_d^\rho M_0^d (\rho + d)^{\rho+d}}. \tag{27}$$

It follows that we can find $n_0 \in \mathbb{N}$ be large enough that $\mathcal{R}_n$ is nonempty for all $P \in \mathcal{P}_{d,\theta}$, $k \in K_\beta$ and $n \geq n_0$, so that, by Assumption (A.1), for $n \geq n_0$ it is an open subset of $\mathbb{R}^d$ and, therefore, a d-dimensional manifold. Let $\mathcal{S}_n := \mathcal{S} \cap \mathcal{R}_n$,

$$B_{1,n} := \int_{\mathcal{S}_n} \frac{\bar{f}(x_0)}{4\|\dot{\eta}(x_0)\|} \, d \, \mathrm{Vol}^{d-1}(x_0)$$

and

$$B_{2,n} := \int_{\mathcal{S}_n} \frac{\bar{f}(x_0)^{1-4/d}}{\|\dot{\eta}(x_0)\|} a(x_0)^2 \, d \, \mathrm{Vol}^{d-1}(x_0).$$

Recalling the definition of ϵ_n in (8), for $n \geq n_0$, we may apply Theorem 5 with $k_{\mathrm{L}}(x) = k$ for all $x \in \mathbb{R}^d$ to deduce that

$$R_{\mathcal{R}_n}(\hat{C}_n^{k\mathrm{nn}}) - R_{\mathcal{R}_n}(C^{\mathrm{Bayes}}) = B_{1,n}\frac{1}{k} + B_{2,n}\left(\frac{k}{n}\right)^{4/d} + W_{n,1} + W_{n,2},$$

where $\sup_{P \in \mathcal{P}_{d,\theta}} \sup_{k \in K_\beta} |W_{n,1}|/\gamma_n(k) \to 0$ and where

$$\limsup_{n\to\infty} \sup_{P \in \mathcal{P}_{d,\theta}} \sup_{k \in K_\beta} \frac{|W_{n,2}|}{P_X((\partial\mathcal{R}_n)^{\epsilon_n} \cap \mathcal{S}^{\epsilon_n})} \leq 1.$$

We now show that, under the conditions of part (i), $B_{1,n}$ and $B_{2,n}$ are well approximated by integrals over the whole of the manifold $\mathcal{S}$, and that these integrals are uniformly bounded. Given $x_0 \in \mathcal{S} \cap \{x \in \mathbb{R}^d : \bar{f}(x) > 0\}$, define $\epsilon_0(x_0) := \min\{1, \frac{\epsilon_0 \log 2}{2d}, \frac{1}{4\ell(\bar{f}(x_0))}\}$. Then for any $t \in [-\epsilon_0(x_0), \epsilon_0(x_0)]$ we have by (A.2) and Cauchy–Schwarz that

$$\begin{aligned}\left|\frac{\bar{f}(x_0^t)}{\bar{f}(x_0)} - 1\right| &= \left|\frac{\bar{f}(x_0^t) - \bar{f}(x_0) - (x_0^t - x_0)^T \nabla \bar{f}(x_0)}{\bar{f}(x_0)} + \frac{(x_0^t - x_0)^T \nabla \bar{f}(x_0)}{\bar{f}(x_0)}\right| \\ &\leq \frac{t^2}{2}\ell\big(\bar{f}(x_0)\big) + |t|\ell\big(\bar{f}(x_0)\big) \leq \frac{1}{2}.\end{aligned}$$

Moreover, writing $\lambda_1, \ldots, \lambda_d$ for the eigenvalues of the matrix B defined in (55), for $t \in [-\epsilon_0(x_0), \epsilon_0(x_0)]$, we have

$$\begin{aligned}\big|\log\det(I + tB)\big| &= \left|\sum_{j=1}^d \log(1 + t\lambda_j)\right| \\ &\leq 2|t| \sum_{j=1}^d |\lambda_j| \leq 2|t| d \|B\|_{\mathrm{op}} \leq \frac{2|t|d}{\epsilon_0},\end{aligned}$$

so $\det(I + tB) \geq 1/2$. Hence, for any $\tau \in (d/(\rho+d), 1]$ there exists $A_\tau = A_\tau(d, \theta) > 0$ such that, writing $\bar{\tau} := \frac{1}{2}(\tau + \frac{d}{\rho+d})$, by (62), Hölder's inequality and (A.4), we have

$$
\begin{aligned}
&\int_{\mathcal{S}} \bar{f}(x_0)^\tau \, d\,\mathrm{Vol}^{d-1}(x_0) \\
&\quad = \int_{\mathcal{S}} \frac{1}{2\epsilon_0(x_0)} \int_{-\epsilon_0(x_0)}^{\epsilon_0(x_0)} \bar{f}(x_0)^\tau \, dt \, d\,\mathrm{Vol}^{d-1}(x_0) \\
&\quad \leq 2^{\tau-1} \\
&\qquad \times \int_{\mathcal{S}} \int_{-\epsilon_0(x_0)}^{\epsilon_0(x_0)} \max\Big\{1, \frac{2d}{\epsilon_0 \log 2}, 4\ell\big(2\bar{f}(x_0^t)/3\big)\Big\} \bar{f}(x_0^t)^\tau \, dt \, d\,\mathrm{Vol}^{d-1}(x_0) \\
&\quad \leq 2^\tau \int_{\mathcal{S}^{\epsilon_0}} \max\Big\{1, \frac{2d}{\epsilon_0 \log 2}, 4\ell\big(2\bar{f}(x)/3\big)\Big\} \bar{f}(x)^\tau \, dx \\
&\quad \leq A_\tau \int_{\mathbb{R}^d} \bar{f}(x)^{\bar{\tau}} \, dx \\
&\quad \leq A_\tau (1 + M_0)^{\bar{\tau}} \Big\{ \int_{\mathbb{R}^d} \big(1 + \|x\|^\rho\big)^{-\frac{\bar{\tau}}{1-\bar{\tau}}} \, dx \Big\}^{1-\bar{\tau}} \\
&\quad =: A'_\tau < \infty.
\end{aligned} \tag{28}
$$

Now, by Assumption (A.3), for any $P \in \mathcal{P}_{d,\theta}$,

$$
B_1 = \int_{\mathcal{S}} \frac{\bar{f}(x_0)}{4\|\dot{\eta}(x_0)\|} \, d\,\mathrm{Vol}^{d-1}(x_0) \leq \frac{1}{4\epsilon_0 M_0} \int_{\mathcal{S}} \bar{f}(x_0) \, d\,\mathrm{Vol}^{d-1}(x_0) \leq \frac{A'_1}{4\epsilon_0 M_0}.
$$

Moreover, writing $\bar{\tau} := \frac{1}{2}(1 + \frac{d}{\rho+d})$,

$$
\begin{aligned}
\sup_{P \in \mathcal{P}_{d,\theta}} \sup_{k \in K_\beta} (B_1 - B_{1,n}) &= \sup_{P \in \mathcal{P}_{d,\theta}} \sup_{k \in K_\beta} \int_{\mathcal{S} \setminus \mathcal{R}_n} \frac{\bar{f}(x_0)}{4\|\dot{\eta}(x_0)\|} \, d\,\mathrm{Vol}^{d-1}(x_0) \\
&\leq \sup_{P \in \mathcal{P}_{d,\theta}} \sup_{k \in K_\beta} \frac{1}{4\epsilon_0 M_0} \int_{\mathcal{S} \setminus \mathcal{R}_n} \bar{f}(x_0) \, d\,\mathrm{Vol}^{d-1}(x_0) \\
&\leq \sup_{P \in \mathcal{P}_{d,\theta}} \sup_{k \in K_\beta} \frac{\delta_n^{1-\bar{\tau}}}{4\epsilon_0 M_0} \int_{\mathcal{S} \setminus \mathcal{R}_n} \bar{f}(x_0)^{\bar{\tau}} \, d\,\mathrm{Vol}^{d-1}(x_0) \\
&\leq \frac{\ell^{d(1-\bar{\tau})}(1/(n-1)) \log^{d(1-\bar{\tau})}(n-1)}{4\epsilon_0 M_0 (n-1)^{\beta(1-\bar{\tau})}} A'_{\bar{\tau}} \to 0.
\end{aligned}
$$

By Assumptions (A.2), (A.3), (28) and the fact that $\rho/(\rho+d) > 4/d$, we have, writing $\bar{\tau} := \frac{1}{2}(1 - 4/d + \frac{d}{\rho+d})$, that

$$
\begin{aligned}
\sup_{P \in \mathcal{P}_{d,\theta}} B_2 &= \sup_{P \in \mathcal{P}_{d,\theta}} \int_{\mathcal{S}} \frac{\bar{f}(x_0)^{1-4/d}}{\|\dot{\eta}(x_0)\|} a(x_0)^2 \, d\,\mathrm{Vol}^{d-1}(x_0) \\
&\leq \sup_{P \in \mathcal{P}_{d,\theta}} \sup_{x_0 \in \mathcal{S}} \Big\{ \frac{a(x_0)^2 \bar{f}(x_0)^{\frac{\rho/(\rho+d)-4/d}{2}}}{\|\dot{\eta}(x_0)\|} \Big\} \int_{\mathcal{S}} \bar{f}(x_0)^{\bar{\tau}} \, d\,\mathrm{Vol}^{d-1}(x_0) \\
&\leq \sup_{\delta \in (0, M_0]} \frac{M_0 \delta^{\frac{\rho/(\rho+d)-4/d}{2}} \{\ell(\delta) + 1/2\}^2}{(d+2)^2 a_d^{4/d} \epsilon_0} A'_{\bar{\tau}} < \infty.
\end{aligned}
$$

Similarly,

$$\sup_{P\in\mathcal{P}_{d,\theta}}\sup_{k\in K_\beta}(B_2-B_{2,n}) = \sup_{P\in\mathcal{P}_{d,\theta}}\sup_{k\in K_\beta}\int_{\mathcal{S}\setminus\mathcal{R}_n}\frac{\bar{f}(x_0)^{1-4/d}}{\|\dot{\eta}(x_0)\|}a(x_0)^2\,d\,\mathrm{Vol}^{d-1}(x_0)$$
$$\leq \sup_{k\in K_\beta}\sup_{\delta\in(0,\delta_n]}\frac{M_0\delta^{\frac{\rho/(\rho+d)-4/d}{2}}\{\ell(\delta)+1/2\}^2}{(d+2)^2a_d^{4/d}\epsilon_0}A'_{\bar{\tau}}\to 0.$$

A similar argument shows that $\gamma_n(k)=O(1/k+(k/n)^{4/d})$, uniformly for $P\in\mathcal{P}_{d,\theta}$ and $k\in K_\beta$.

Finally, we bound $P_X((\partial\mathcal{R}_n)^{\epsilon_n}\cap\mathcal{S}^{\epsilon_n})$ and $R_{\mathcal{R}_n^c}(\hat{C}_n^{k\mathrm{nn}})-R_{\mathcal{R}_n^c}(C^{\mathrm{Bayes}})$. Suppose that $x\in(\partial\mathcal{R}_n)^{\epsilon_n}\cap\mathcal{S}^{\epsilon_n}$. Then there exists $z\in\partial\mathcal{R}_n\cap B_{\epsilon_n}(x)\cap\mathcal{S}^{2\epsilon_n}$ with $\bar{f}(z)=\delta_n$. By Assumption (A.2), we have that

$$\begin{aligned}\left|\frac{\bar{f}(x)}{\bar{f}(z)}-1\right| &\leq \ell\big(\bar{f}(z)\big)\|x-z\|+\frac{1}{2}\ell\big(\bar{f}(z)\big)\|x-z\|^2\\ &\leq \frac{1+\epsilon_n/2}{\beta^{1/2}\log^{1/2}(n-1)}.\end{aligned}\tag{29}$$

Thus there exists $n_1\in\mathbb{N}$ such that $(\partial\mathcal{R}_n)^{\epsilon_n}\cap\mathcal{S}^{\epsilon_n}\subseteq\{x\in\mathbb{R}^d:\bar{f}(x)\leq 2\delta_n\}$ for $n\geq n_1$. By the moment assumption in (A.4) and Hölder's inequality, observe that for any $\alpha\in(0,1)$, $P\in\mathcal{P}_{d,\theta}$, $n\geq n_1$ and $\epsilon>0$,

$$\begin{aligned}P_X\big((\partial\mathcal{R}_n)^{\epsilon_n}\cap\mathcal{S}^{\epsilon_n}\big) &\leq \mathbb{P}\big\{\bar{f}(X)\leq 2\delta_n\big\}\\ &\leq (2\delta_n)^{\frac{\rho(1-\alpha)}{\rho+d}}\int_{x:\bar{f}(x)\leq 2\delta_n}\bar{f}(x)^{1-\frac{\rho(1-\alpha)}{\rho+d}}\,dx\\ &\leq (2\delta_n)^{\frac{\rho(1-\alpha)}{\rho+d}}\bigg\{\int_{\mathbb{R}^d}\big(1+\|x\|^\rho\big)\bar{f}(x)\,dx\bigg\}^{1-\frac{\rho(1-\alpha)}{\rho+d}}\\ &\quad\times\bigg\{\int_{\mathbb{R}^d}\frac{1}{(1+\|x\|^\rho)^{\frac{d+\rho\alpha}{\rho(1-\alpha)}}}\,dx\bigg\}^{\frac{\rho(1-\alpha)}{\rho+d}}\\ &\leq (2\delta_n)^{\frac{\rho(1-\alpha)}{\rho+d}}(1+M_0)^{1-\frac{\rho(1-\alpha)}{\rho+d}}\\ &\quad\times\bigg\{\int_{\mathbb{R}^d}\frac{1}{(1+\|x\|^\rho)^{\frac{d+\rho\alpha}{\rho(1-\alpha)}}}\,dx\bigg\}^{\frac{\rho(1-\alpha)}{\rho+d}}\\ &= o\bigg(\Big(\frac{k}{n}\Big)^{\frac{\rho(1-\alpha)}{\rho+d}-\epsilon}\bigg)\end{aligned}\tag{30}$$

uniformly for $k\in K_\beta$. Moreover,

$$R_{\mathcal{R}_n^c}\big(\hat{C}_n^{k\mathrm{nn}}\big)-R_{\mathcal{R}_n^c}\big(C^{\mathrm{Bayes}}\big)\leq P_X\big(\mathcal{R}_n^c\big)\leq\mathbb{P}\big\{\bar{f}(X)\leq 2\delta_n\big\},$$

so the same bound (30) applies. Since $\rho/(\rho+d)>4/d$ and $\alpha\in(0,1)$ was arbitrary, this completes the proof of part (i).

For part (ii), in contrast to part (i), the dominant contribution to the excess risk could now arise from the tail of the distribution. First, as in part (i), we have $B_{1,n}\to B_1\leq A'_1/(4\epsilon_0M_0)$, uniformly for $P\in\mathcal{P}_{d,\theta}$ and $k\in K_\beta$. Furthermore, using Assumption (A.3), (28) and the fact

that $4/d \geq \rho/(\rho+d)$, we see that, for any $\epsilon' \in (0, \rho/(\rho+d)]$,

$$B_{2,n}\Big(\frac{k}{n}\Big)^{4/d} \leq \delta_n^{\rho/(\rho+d)-\epsilon'} \int_{\mathcal{S}_n} \frac{\delta_n^{4/d-\rho/(\rho+d)}\, \bar{f}(x_0)^{1-4/d+\epsilon'}}{c_n^4 \log^4((n-1)/k)\|\dot{\eta}(x_0)\|} a(x_0)^2 \, d\,\mathrm{Vol}^{d-1}(x_0)$$

$$\leq \sup_{x_0 \in \mathcal{S}_n} a(x_0)^2 \frac{\delta_n^{\rho/(\rho+d)-\epsilon'} A'_{d/(\rho+d)+\epsilon'}}{\epsilon_0 M_0 c_n^4 \log^4((n-1)/k)} = o\big((k/n)^{\rho/(\rho+d)-\epsilon}\big),$$

for every $\epsilon \in (\epsilon', \rho/(\rho+d)]$, uniformly for $P \in \mathcal{P}_{d,\theta}$ and $k \in K_\beta$, where the final conclusion follows from the fact that $\sup_{P\in\mathcal{P}_{d,\theta}} \sup_{x_0\in\mathcal{S}_n} a^2(x_0)/c_n^2$ is bounded. We can also bound $\gamma_n(k)$ by the same argument, so the result follows in the same way as in part (i). □

6.3. *Proofs of results from Section 4.*

PROOF OF THEOREM 2. Recall that

$$k_{\mathrm{O}}(x) = \max\big[\lceil (n-1)^\beta \rceil, \min\big\{\lfloor B\{\bar{f}(x)(n-1)\}^{4/(d+4)} \rfloor, \lfloor (n-1)^{1-\beta} \rfloor\big\}\big],$$

and define

$$\delta_{n,\mathrm{O}}(x) := \frac{k_{\mathrm{O}}(x)}{n-1} c_n^d \log^d\Big(\frac{n-1}{k_{\mathrm{O}}(x)}\Big),$$

where $c_n := \sup_{x_0\in\mathcal{S}:\bar{f}(x_0)\geq k_{\mathrm{O}}(x_0)/(n-1)} \ell(\bar{f}(x_0))$. For $\alpha \in ((1+d/4)\beta, 1)$, let

$$\mathcal{R}_n = \big\{x \in \mathbb{R}^d : \bar{f}(x) > (n-1)^{-(1-\alpha)}\big\} \cap \mathcal{X}_{\bar{f}}.$$

Then there exists $n_0 \in \mathbb{N}$ such that for $n \geq n_0$ we have $\mathcal{R}_n \subseteq \{x \in \mathbb{R}^d : \bar{f}(x) \geq \delta_{n,\mathrm{O}}(x)\}$ for all $P \in \mathcal{P}_{d,\theta}$ and $B \in [B_*, B^*]$, and by Assumption (A.1) and (27), we then have that $\mathcal{R}_n$ is a d-dimensional manifold. There exists $n_1 \in \mathbb{N}$ such that for all $n \geq n_1$, $P \in \mathcal{P}_{d,\theta}$, $B \in [B_*, B^*]$ and $x \in \mathcal{R}_n \cap \mathcal{S}^{\epsilon_0}$ we have that $k_{\mathrm{O}}(x) = \lfloor B\{\bar{f}(x)(n-1)\}^{4/(d+4)} \rfloor$. By (A.2), we therefore have that $k_{\mathrm{O}} \in K_{\beta,\tau}$ for some $\tau = \tau_n$ (which does not depend on $P \in \mathcal{P}_{d,\theta}$ or $B \in [B_*, B^*]$) with $\tau_n \searrow 0$.

By a similar argument to that in (29), there exists $n_2 \in \mathbb{N}$ such that for $n \geq n_2$, $P \in \mathcal{P}_{d,\theta}$, $B \in [B_*, B^*]$ and $x \in (\partial\mathcal{R}_n)^{\epsilon_n} \cap \mathcal{S}^{\epsilon_n}$, we have $\bar{f}(x) \leq 2(n-1)^{-(1-\alpha)}$. But, by Markov's inequality and Hölder's inequality, for $\tilde{\alpha} \in (0,1)$ and any $P \in \mathcal{P}_{d,\theta}$,

$$\begin{aligned}
&\mathbb{P}\big\{\bar{f}(X) \leq 2(n-1)^{-(1-\alpha)}\big\} \\
&\leq \big\{2(n-1)^{-(1-\alpha)}\big\}^{\frac{\rho(1-\tilde{\alpha})}{\rho+d}} \int_{\mathbb{R}^d} \bar{f}(x)^{1-\frac{\rho(1-\tilde{\alpha})}{\rho+d}}\, dx \\
&\leq \big\{2(n-1)^{-(1-\alpha)}\big\}^{\frac{\rho(1-\tilde{\alpha})}{\rho+d}} (1+M_0)^{1-\frac{\rho(1-\tilde{\alpha})}{\rho+d}} \\
&\quad \times \bigg\{\int_{\mathbb{R}^d} \frac{1}{(1+\|x\|^\rho)^{(\rho+d)/\{\rho(1-\tilde{\alpha})\}-1}}\, dx\bigg\}^{\frac{\rho(1-\tilde{\alpha})}{\rho+d}}.
\end{aligned} \tag{31}$$

Thus, if $\rho > 4$, then we can choose $\alpha \in ((1+d/4)\beta, d(\rho-4)/\{\rho(d+4)\})$ and $\tilde{\alpha} < 1 - 4(\rho+d)/\{\rho(1-\alpha)(d+4)\}$ in (31) to conclude that

$$\sup_{P\in\mathcal{P}_{d,\theta}} P_X(\mathcal{R}_n^c) \leq \sup_{P\in\mathcal{P}_{d,\theta}} \mathbb{P}\big\{\bar{f}(X) \leq 2(n-1)^{-(1-\alpha)}\big\} = o\big(n^{-4/(d+4)}\big).$$

Moreover, writing

$$B_{3,n} := \int_{\mathcal{S}_n} \frac{\bar{f}(x_0)^{d/(d+4)}}{\|\dot{\eta}(x_0)\|} \Big\{\frac{1}{4B} + B^{4/d} a(x_0)^2\Big\}\, d\,\mathrm{Vol}^{d-1}(x_0),$$

by very similar arguments to those given in the proof of Theorem 1, $B_{3,n} \to B_3$ and $\gamma_n(k_{\mathrm{O}}) = O(n^{-4/(d+4)})$ as $n \to \infty$, both uniformly for $P \in \mathcal{P}_{d,\theta}$ and $B \in [B_*, B^*]$. The proof of part (i) therefore follows from Theorem 5.

On the other hand, if $\rho \leq 4$, then choosing both $\tilde{\alpha} > 0$ and $\alpha > (1+d/4)\beta$ to be sufficiently small, we find from (31) that

$$B_{3,n} n^{-4/(d+4)} + \gamma_n(k_{\mathrm{O}}) + P_X\big((\partial \mathcal{R}_n)^{\epsilon_n} \cap \mathcal{S}^{\epsilon_n}\big) + P_X\big(\mathcal{R}_n^c\big) = o\big(n^{-\frac{\rho}{\rho+d}+\beta+\epsilon}\big),$$

for every $\epsilon > 0$, uniformly for $P \in \mathcal{P}_{d,\theta}$ and $B \in [B_*, B^*]$. After another application of Theorem 5, this proves part (ii). $\square$

PROOF OF THEOREM 3. We prove parts (i) and (ii) of the theorem simultaneously, by appealing to the corresponding arguments in the proof of Theorem 2. First, as in the proof of Theorem 2, for $\alpha \in ((1+d/4)\beta, 1)$, we define $\mathcal{R}_n = \{x \in \mathbb{R}^d : \bar{f}(x) > (n-1)^{-(1-\alpha)}\} \cap \mathcal{X}_{\bar{f}}$ and introduce the following class of functions: for $\tau > 0$, let

$$\mathcal{F}_{n,\tau} := \left\{ \tilde{f} : \mathbb{R}^d \to \mathbb{R} : \tilde{f} \text{ continuous}, \sup_{x \in \mathcal{R}_n} \left| \frac{\bar{f}(x)}{\tilde{f}(x)} - 1 \right| \leq \tau \right\}.$$

Let $\tau = \tau_n := 2(n-1)^{-\alpha/2}$. We first show that $\hat{f}_m \in \mathcal{F}_{n,\tau}$ with high probability. For $x \in \mathcal{R}_n$,

$$\left| \frac{\hat{f}_m(x)}{\bar{f}(x)} - 1 \right| \leq (n-1)^{1-\alpha} |\hat{f}_m(x) - \bar{f}(x)| \leq (n-1)^{1-\alpha} \|\hat{f}_m - \bar{f}\|_\infty.$$

Now

$$\|\hat{f}_m - \bar{f}\|_\infty \leq \|\hat{f}_m - \mathbb{E}\hat{f}_m\|_\infty + \|\mathbb{E}\hat{f}_m - \bar{f}\|_\infty. \tag{32}$$

To bound the first term in (32), by Giné and Guillou (2002, Corollary 2.2), there exist $C, L > 0$, such that

$$\sup_{P \in \mathcal{P}_{d,\theta} \cap \mathcal{Q}_{d,\gamma,\lambda}} \mathbb{P}\left(\|\hat{f}_m - \mathbb{E}\hat{f}_m\|_\infty \geq \frac{s}{m^{\gamma/(d+2\gamma)}} \right) \leq L \left(\frac{4L}{4L+C} \right)^{\frac{A^d s^2}{LC\lambda R(K)}}, \tag{33}$$

for all $s \in [\frac{C\|\bar{f}\|_\infty^{1/2} R(K)^{1/2}}{A^{d/2}} \log^{1/2}(\frac{\|K\|_\infty m^{d/(2(d+2\gamma))}}{\|\bar{f}\|_\infty^{1/2} A^{d/2} R(K)^{1/2}}), \frac{C\|\bar{f}\|_\infty R(K) m^{\gamma/(d+2\gamma)}}{\|K\|_\infty}]$ and $A \in [A_*, A^*]$.

Recall that for $P \in \mathcal{P}_{d,\theta}$, we have $\|\bar{f}\|_\infty \leq \lambda$ and $\|\bar{f}\|_\infty$ also satisfies the lower bound in (27). Hence, by applying the bound in (33) with $s = s_0 := m^{\gamma/(d+2\gamma)}/(n-1)^{1-\alpha/2}$, since $m \geq m_0(n-1)^{d/\gamma+2}$, we have that there exists $n_* \in \mathbb{N}$, not depending on $P \in \mathcal{P}_{d,\theta}$ or $A \in [A_*, A^*]$ such that for $n \geq n_*$,

$$\begin{aligned}
&\sup_{P \in \mathcal{P}_{d,\theta} \cap \mathcal{Q}_{d,\gamma,\lambda}} \mathbb{P}\left\{ \|\hat{f}_m - \mathbb{E}\hat{f}_m\|_\infty \geq \frac{1}{(n-1)^{1-\alpha/2}} \right\} \\
&\quad = \sup_{P \in \mathcal{P}_{d,\theta} \cap \mathcal{Q}_{d,\gamma,\lambda}} \mathbb{P}\{ \|\hat{f}_m - \mathbb{E}\hat{f}_m\|_\infty \geq s_0 m^{-\gamma/(d+2\gamma)} \} \\
&\quad \leq L \left(\frac{4L}{4L+C} \right)^{\frac{A^d (n-1)^\alpha m_0^{2\gamma/(d+2\gamma)}}{LC\lambda R(K)}} \\
&\quad = O\big(n^{-M}\big),
\end{aligned}$$

for all $M > 0$, uniformly for $A \in [A_*, A^*]$. For the second term in (32), by a Taylor expansion, we have that for all $P \in \mathcal{P}_{d,\theta} \cap \mathcal{Q}_{d,\gamma,\lambda}$ and $A \in [A_*, A^*]$,

$$\begin{aligned}\|\mathbb{E}\hat{f}_m - \bar{f}\|_\infty &\le \lambda A^\gamma m^{-\gamma/(d+2\gamma)} \int_{\mathbb{R}^d} \|z\|^\gamma |K(z)|\,dz \\ &\le \frac{\lambda A^\gamma m_0^{-\gamma/(d+2\gamma)}}{n-1} \int_{\mathbb{R}^d} \|z\|^\gamma |K(z)|\,dz.\end{aligned}$$

It follows that, writing $\tau_0 := 2(n-1)^{-\alpha/2}$, we have

$$\sup_{P \in \mathcal{P}_{d,\theta} \cap \mathcal{Q}_{d,\gamma,\lambda}} \sup_{A \in [A_*, A^*]} \mathbb{P}(\hat{f}_m \notin \mathcal{F}_{n,\tau_0}) = O(n^{-M})$$

for all $M > 0$.

Now, for $\tilde{f} \in \mathcal{F}_{n,\tau_0}$, let

$$k_{\tilde{f}}(x) := \max\bigl[\lceil (n-1)^\beta \rceil, \min\bigl\{\lfloor B\{\tilde{f}(x)(n-1)\}^{4/(d+4)} \rfloor, \lfloor (n-1)^{1-\beta} \rfloor\bigr\}\bigr].$$

Let $c_n := \sup_{x_0 \in \mathcal{S}: \bar{f}(x_0) \ge k_{\tilde{f}}(x_0)/(n-1)} \ell(\bar{f}(x_0))$, and let

$$\delta_{n,\tilde{f}}(x) := \frac{k_{\tilde{f}}(x)}{n-1} c_n^d \log^d\biggl(\frac{n-1}{k_{\tilde{f}}(x)}\biggr).$$

Then there exists $n_0 \in \mathbb{N}$ such that for $n \ge n_0$ and $\tilde{f} \in \mathcal{F}_{n,\tau_0}$, we have $\mathcal{R}_n \subseteq \{x \in \mathbb{R}^d : \bar{f}(x) \ge \delta_{n,\tilde{f}}(x)\}$ and $k_{\tilde{f}} \in K_{\beta,\tau_0}$. We can therefore apply Theorem 5 (similar to the application in the proof of Theorem 2) to conclude that for every $\epsilon > 0$,

$$R\bigl(\hat{C}_n^{k_{\tilde{f}}\mathrm{nn}}\bigr) - R\bigl(C^{\mathrm{Bayes}}\bigr) = B_{3,n} n^{-4/(d+4)} + o\bigl(n^{-4/(d+4)} + n^{-\frac{\rho}{\rho+d}+\beta+\epsilon}\bigr)$$

uniformly for $P \in \mathcal{P}_{d,\theta} \cap \mathcal{Q}_{d,\gamma,\lambda}$ and $\tilde{f} \in \mathcal{F}_{n,\tau_0}$, where $B_{3,n}$ was defined in the proof of Theorem 2. The proof of both parts (i) and (ii) is now completed by following the relevant steps in the proof of Theorem 2. □

Acknowledgements. The authors are grateful to the anonymous reviewers, whose constructive comments helped to improve the paper. We would also like to thank the Isaac Newton Institute for Mathematical Sciences for support and hospitality during the programme "Statistical Scalability" when work on this paper was undertaken.

This work was supported by EPSRC grant number EP/R014604/1.

The second author was supported by an Engineering and Physical Sciences Research Council (EPSRC) programme grant.

The third author was supported by an EPSRC Fellowship and programme grant, as well as a grant from the Leverhulme Trust.

SUPPLEMENTARY MATERIAL

Online supplement "Local nearest neighbour classification with applications to semi-supervised learning" (DOI: 10.1214/19-AOS1868SUPP; .pdf). We present our remaining theoretical arguments and a simulation study. Further, we provide an introduction to differential geometry, tubular neighbourhoods and integration on manifolds.

REFERENCES

ABRAMSON, I. S. (1982). On bandwidth variation in kernel estimates—a square root law. *Ann. Statist.* **10** 1217–1223. MR0673656

AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** 608–633. MR2336861 https://doi.org/10.1214/009053606000001217

BERRETT, T. B. and SAMWORTH, R. J. (2019a). Efficient two-sample functional estimation and the super-oracle phenomenon. https://arxiv.org/abs/1904.09347.

BERRETT, T. B. and SAMWORTH, R. J. (2019b). Nonparametric independence testing via mutual information. *Biometrika* **106** 547–566. MR3992389 https://doi.org/10.1093/biomet/asz024

BERRETT, T. B., SAMWORTH, R. J. and YUAN, M. (2019). Efficient multivariate entropy estimation via k-nearest neighbour distances. *Ann. Statist.* **47** 288–318. MR3909934 https://doi.org/10.1214/18-AOS1688

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

BIAU, G. and DEVROYE, L. (2015). *Lectures on the Nearest Neighbor Method. Springer Series in the Data Sciences*. Springer, Cham. MR3445317 https://doi.org/10.1007/978-3-319-25388-6

BOUCHERON, S., BOUSQUET, O. and LUGOSI, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM Probab. Stat.* **9** 323–375. MR2182250 https://doi.org/10.1051/ps:2005018

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities*. Oxford Univ. Press, Oxford. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

BREIMAN, L., MEISEL, W. and PURCELL, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics* **19** 135–144.

CANNINGS, T. I, BERRETT, T. B and SAMWORTH, R. J (2020). Supplement to "Local nearest neighbour classification with applications to semi-supervised learning." https://doi.org/10.1214/19-AOS1868SUPP.

CELISSE, A. and MARY-HUARD, T. (2018). Theoretical analysis of cross-validation for estimating the risk of the k-nearest neighbor classifier. *J. Mach. Learn. Res.* **19** 58. MR3899760

CHAPELLE, O., ZIEN, A. and SCHÖLKOPF, B., eds. (2006). *Semi-Supervised Learning*. MIT Press, Cambridge MA.

CHAUDHURI, K. and DASGUPTA, S. (2014). Rates of convergence for nearest neighbor classification. *Adv. Neural Inf. Process. Syst.* **27** 3437–3445.

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 https://doi.org/10.1007/978-1-4612-0711-5

FIX, E. and HODGES, J. L. (1951). Discriminatory analysis—nonparametric discrimination: Consistency properties Technical Report number 4, USAF School of Aviation Medicine, Randolph Field, TX.

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

GADAT, S., KLEIN, T. and MARTEAU, C. (2016). Classification in general finite dimensional spaces with the k-nearest neighbor rule. *Ann. Statist.* **44** 982–1009. MR3485951 https://doi.org/10.1214/15-AOS1395

GINÉ, E. and GUILLOU, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. Henri Poincaré Probab. Stat.* **38** 907–921. MR1955344 https://doi.org/10.1016/S0246-0203(02)01128-7

GINÉ, E. and SANG, H. (2010). Uniform asymptotics for kernel density estimators with variable bandwidths. *J. Nonparametr. Stat.* **22** 773–795. MR2682221 https://doi.org/10.1080/10485250903483331

HALL, P. and KANG, K.-H. (2005). Bandwidth choice for nonparametric classification. *Ann. Statist.* **33** 284–306. MR2157804 https://doi.org/10.1214/009053604000000959

HALL, P., PARK, B. U. and SAMWORTH, R. J. (2008). Choice of neighbor order in nearest-neighbor classification. *Ann. Statist.* **36** 2135–2152. MR2458182 https://doi.org/10.1214/07-AOS537

HECKEL, R. and BÖLCSKEI, H. (2015). Robust subspace clustering via thresholding. *IEEE Trans. Inform. Theory* **61** 6320–6342. MR3418967 https://doi.org/10.1109/TIT.2015.2472520

KOZACHENKO, L. F. and LEONENKO, N. N. (1987). A statistical estimate for the entropy of a random vector. *Problemy Peredachi Informatsii* **23** 9–16. MR0908626

KULKARNI, S. R. and POSNER, S. E. (1995). Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Trans. Inform. Theory* **41** 1028–1039. MR1366756 https://doi.org/10.1109/18.391248

LOFTSGAARDEN, D. O. and QUESENBERRY, C. P. (1965). A nonparametric estimate of a multivariate density function. *Ann. Math. Stat.* **36** 1049–1051. MR0176567 https://doi.org/10.1214/aoms/1177700079

MACK, Y. P. (1983). Rate of strong uniform convergence of k-NN density estimates. *J. Statist. Plann. Inference* **8** 185–192. MR0720150 https://doi.org/10.1016/0378-3758(83)90037-X

MACK, Y. P. and ROSENBLATT, M. (1979). Multivariate k-nearest neighbor density estimates. *J. Multivariate Anal.* **9** 1–15. MR0530638 https://doi.org/10.1016/0047-259X(79)90065-4

MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. MR1765618 https://doi.org/10.1214/aos/1017939240

SAMWORTH, R. J. (2012). Optimal weighted nearest neighbour classifiers. *Ann. Statist.* **40** 2733–2763. MR3097618 https://doi.org/10.1214/12-AOS1049

SCHILLING, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.* **81** 799–806. MR0860514

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

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