Modern k-Nearest Neighbour Methods in Entropy Estimation, Independence Testing and Classification

Thomas Benjamin Berrett

Gonville and Caius College

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

July 2017

Abstract

Nearest neighbour methods are a classical approach in nonparametric statistics. The k-nearest neighbour classifier can be traced back to the seminal work of Fix and Hodges (1951) and they also enjoy popularity in many other problems including density estimation and regression. In this thesis we study their use in three different situations, providing new theoretical results on the performance of commonly-used nearest neighbour methods and proposing new procedures that are shown to outperform these existing methods in certain settings.

The first problem we discuss is that of entropy estimation. Many statistical procedures, including goodness-of-fit tests and methods for independent component analysis, rely critically on the estimation of the entropy of a distribution. In this chapter, we seek entropy estimators that are efficient and achieve the local asymptotic minimax lower bound with respect to squared error loss. To this end, we study weighted averages of the estimators originally proposed by Kozachenko and Leonenko (1987), based on the k-nearest neighbour distances of a sample. A careful choice of weights enables us to obtain an efficient estimator in arbitrary dimensions, given sufficient smoothness, while the original unweighted estimator is typically only efficient in up to three dimensions.

A related topic of study is the estimation of the mutual information between two random vectors, and its application to testing for independence. We propose tests for the two different situations of the marginal distributions being known or unknown and analyse their performance.

Finally, we study the classical k-nearest neighbour classifier of Fix and Hodges (1951) and provide a new asymptotic expansion for its excess risk. We also show that, in certain situations, a new modification of the classifier that allows k to vary with the location of the test point can provide improvements. This has applications to the field of semi-supervised learning, where, in addition to labelled training data, we also have access to a large sample of unlabelled data.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared below. It is not substantially the same as any that I have submitted, or is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution.

Chapter 2 is joint work with Ming Yuan (University of Wisconsin–Madison) and Richard Samworth. Chapter 3 is joint work with Richard Samworth. Chapter 4 is joint work with Timothy Cannings (University of Southern California) and Richard Samworth. Chapters 2 and 4 have been submitted for publication as Berrett, Samworth and Yuan (2017) and Cannings, Berrett and Samworth (2017). Chapter 3 constitutes part of work I intend to submit for publication in the near future.

Thomas Benjamin Berrett
Cambridge, July 2017

Acknowledgements

There are many people without whose help and support over the past three years I would not have been able to complete this thesis. First and foremost I would like to acknowledge the debt I owe to my supervisor Professor Richard Samworth. While working on his Part III essay I made the decision to pursue an academic career, starting here in Cambridge. I have been lucky enough to collaborate with him extensively and have benefited tremendously from his hard work, expertise and general guidance. My other collaborators Tim Cannings and Ming Yuan have also been inspiring and very helpful.

I am very grateful to the Sims Fund for awarding me a scholarship and providing funding without which it would not have been possible to work towards a PhD.

I would also like to thank everybody who has been in the Statistical Laboratory during my PhD who has made it such a friendly place to work, including John Aston, Tim Cannings, Yining Chen, Oliver Feng, Milana Gataric, Arlene Kim, Danning Li, Susan Pitts, Rajen Shah, Shahin Tavakoli, Jenny Wadsworth, Tengyao Wang and Yi Yu. Thank you also to my friends in college, in the department and elsewhere in the university, including Lawrence Barrott, Benjamin Barrett, James Bell, Nigel Burke, Nicolas Dupré, Pierre Haas, Fiona Hamey, Amit Hazi, Alex Jeffreys, James Kilbane, Chris King, James Munro, Jack Smith and Benjamin Stokell, for making my time in Cambridge so enjoyable. Finally, I would like to thank Jo Evans for her kindness and support and everything else.

Contents

Chapter 1

Introduction to nearest neighbour methods

Nearest neighbour methods are a family of techniques whose wide-ranging influence can be felt in many areas of data science. Their use in statistics dates back at least as far as Fix and Hodges (1951) in which the search for a fully nonparametric classification rule led the authors to propose a k-nearest neighbour classifier and density estimator. A large part of their popularity is undoubtedly due to their simplicity and analytic tractability, which make efficient practical implementation possible and open up the challenge of providing a thorough theoretical understanding.

Perhaps the context in which nearest neighbour methods are most popular is classification and the closely related context of regression. Stone (1977) proved that the k-nearest neighbour approach to classification and regression has the remarkable property of universal consistency in finite dimensions. Indeed, whenever the feature vector X takes values in $\mathbb{R}^d$ for some $d \in \mathbb{N}$ the k-nearest neighbour approach is consistent, in that k can be chosen so that the classifier's asymptotic risk is the same as that of the Bayes classifier, regardless of the distribution of X. Since then a large literature on the subject has developed, though important questions remain unanswered. In Chapter 4 we derive new theoretical results on the k-nearest neighbour classifier and propose a new variant in which the value of k is allowed to depend on the location of the test point.

The success of nearest neighbour methods in density estimation (e.g. Mack and Rosenblatt, 1979) naturally suggests their use in density functional estimation, and in more recent years there have been many works on this topic; see Dasgupta and Kpotufe (2014) for an example of mode estimation and Duong et al. (2016) for an example of density derivative estimation. An important class of density functionals is the class of integral functionals, that is those of the form

$$T(f) = \int \phi(x, f(x))\, dx$$

for some function ϕ; see for example Leonenko, Pronzato and Savani (2008), Evans, Jones and Schmidt (2002), Sricharan, Raich and Hero (2012) and Baryshnikov, Penrose and Yukich (2009). Many of the functionals considered in these works are related to notions of entropy such as Rényi entropy or Shannon entropy. The estimation of Shannon entropy in particular has received a lot of attention in the statistics and machine learning communities where it naturally arises in many applications. In Chapter 2 we study a popular nearest neighbour estimator of Shannon entropy and propose a new estimator and in Chapter 3 we use these estimators in the context of testing for independence through the estimation of mutual information.

Quite apart from their use in classification and density estimation, the versatility of nearest neighbour methods has resulted in their use in disparate settings. Other classical areas of statistics and machine learning where they have been applied include two-sample testing problems (Schilling, 1986) and nonparametric clustering (Heckel and Bölcskei, 2015). In nonlinear dimensionality reduction and manifold learning (Roweis and Saul, 2000; Costa and Hero, 2004; Law and Jain, 2006) they are used for data visualisation and for estimating the intrinsic dimension of large-scale data. They are also a very popular solution to the important practical problems of missing data (Chen and Shao, 2000) and outlier detection (Zhao and Saligrama, 2009; Chandola, Banerjee and Kumar, 2009).

We now give formal definitions. Let $X_1, \ldots, X_n$ be (labelled or unlabelled) random variables taking values in $\mathbb{R}^d$ and, given $x \in \mathbb{R}^d$, define $X_{(1)}(x), \ldots, X_{(n)}(x)$ to be the permutation of $X_1, \ldots, X_n$ such that

$$\|X_{(1)}(x) - x\| \leq \|X_{(2)}(x) - x\| \leq \ldots \leq \|X_{(n)}(x) - x\|.$$

Given $k \in \{1, \ldots, n\}$, we say that $X_{(1)}(x), \ldots, X_{(k)}(x)$ are the k-nearest neighbours of x and define the kth nearest neighbour distance of x to be $\rho_{(k)}(x) = \|X_{(k)}(x) - x\|$. The standard k-nearest neighbour classifier would then assign the test point x to the class which is most represented among the k nearest neighbours of x. When $X_1, \ldots, X_n$ are independent and identically distributed with density function f, the basis of nearest neighbour methods in estimation problems is the approximation

$$\frac{k}{n} \approx V_d \rho_{(k)}^d(x) f(x),$$

valid when $\rho_{(k)}$ is small by the Lebesgue differentiation theorem, where V_d is the volume of the unit ball in $\mathbb{R}^d$. In this thesis we will use $\|\cdot\|$ to represent the Euclidean norm, though any other norm may also be used and one may also consider a more general metric. A link to kernel density estimation can be established by noting that the k-nearest neighbour density estimator may be written as

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^n K\Big(\frac{x - X_i}{h}\Big)$$

where $h = h(x) = \rho_{(k)}(x)$ and $K(x) = V_d^{-1} \mathbb{1}_{\{\|x\| \leq 1\}}$. Here the bandwidth h adapts to the location of the test point, with smaller bandwidths used in areas of high density.

In the i.i.d. setting when X_1 has density f, the density function of $X_{(k)}(x) - x$ at $u \in \mathbb{R}^d$ may be written explicitly as

$$nf(x+u)\binom{n-1}{k-1} p_x(\|u\|)^{k-1}\{1 - p_x(\|u\|)\}^{n-k},$$

where $p_x(r) := \mathbb{P}(\|X_1 - x\| \leq r)$. Similarly, when $p_x(r)$ is differentiable, the density function of $\rho_{(k)}(x)$ may be written as

$$n\binom{n-1}{k-1} p_x(r)^{k-1}\{1 - p_x(r)\}^{n-k} \frac{\partial}{\partial r} p_x(r).$$

The previous expression reveals a connection between nearest neighbour distances and order statistics. Indeed, writing $U_{(1)}, \ldots, U_{(n)}$ for the order statistics of a sample of size n from the uniform distribution on $[0,1]$, we have that

$$\big(p_x(\rho_{(1)}), \ldots, p_x(\rho_{(n)})\big) \overset{d}{=} (U_{(1)}, \ldots, U_{(n)}),$$

and in particular $p_x(\rho_{(k)}) \sim \mathrm{Beta}(k, n-k+1)$. In the analysis of nearest neighbour methods, concentration and moment properties of the Beta distribution are often helpful. Also revealed by the above expressions is the fact that the function $p_x(\cdot)$ plays a crucial role in the analysis of many nearest neighbour methods; assumptions on the smoothness of the density function f are often made to facilitate expansions of $p_x(r)$ for small values of r. Controlling the relative error in these expansions is made considerably easier by assuming that f is bounded below on its support and this is an assumption that is commonly made in previous works; see, for example, Samworth (2012) and Singh and Póczos (2016). In this thesis we do not make this assumption and instead place additional restrictions on the smoothness of f in areas of low density.

As with many nonparametric techniques there is a tuning parameter, in this case k, whose value may affect the performance of the procedures significantly. Heuristically speaking, in many applications k can be seen as controlling a bias–variance trade-off where larger values of k result in larger bias and smaller values of k result in larger variance. In the classification setting it is the bias and variance in estimating the regression function that is balanced through k. Often, as in Chapter 4 here, asymptotic results provide some knowledge of the relationship between n and the optimal choice of k, and allow one to achieve the best rate of convergence, though estimating the precise value of the optimal k is a difficult problem and the value will often depend on the underlying distribution of the data; see, for example, Hall et al. (2008). In practice k is usually chosen heuristically or empirically, often by cross-validation as for our numerical results in Section 4.5; see also Chapter 26 of Devroye et al. (1996) for an overview of some empirical methods. Interestingly, in some situations, such as classification and density estimation, one requires $k \to \infty$ as $n \to \infty$ for consistency whereas in other situations, such as entropy estimation, consistency can be achieved with a fixed value of k.

There are many notable modifications of standard k-nearest neighbour methods. One modern topic of research focuses on using a data-driven metric on the feature space to weight the features differently and improve the performance of nearest neighbour methods; see for example Weinberger and Saul (2009). Another modern modification of the standard nearest neighbour methods is to allow the choice of k to vary with x, often in order to better balance bias and variance by choosing k to be smaller in areas of low density (e.g. Wettschereck and Dietterich, 1994). This is the approach we take in Chapter 4. There are also potential improvements to be made over the standard methods by using weighted nearest neighbour methods. In classification problems this amounts to assigning the test point x to the class C that maximises

$$\sum_{k=1}^{n} w_k \mathbb{1}_{\{X_{(k)}(x) \text{ belongs to class } C\}}$$

for some weight vector w; see for example Hall and Samworth (2005) and Samworth (2012). In estimation problems one can consider a weighted average of k-nearest neighbour estimates over k. This is the approach we take in Chapter 2; see also, for example, Moon et al. (2016) and Sricharan, Wei and Hero (2013).

In modern applications the practicality of a statistical procedure is very important; with large datasets the computational complexity of algorithms must be considered. Due to their simplicity nearest neighbour methods can often be efficiently implemented, and there is a large literature on finding nearest neighbours in a streamlined way. For finding the k-nearest neighbours of a single query point in a sample of size n, for example to classify a test point, methods based on k-d trees achieve an average complexity of $O(\log n)$; see Friedman, Bentley and Finkel (1977). To find all the k-nearest neighbours of a whole sample of size n, such as is required in estimation problems, the complexity is bounded by $O(kn \log n)$; see Vaidya (1989). There has also been extensive research into the approximate nearest neighbours problem, in which neighbours are found whose distance to the test point is close to the nearest neighbour distance, up to some specified threshold. These algorithms can achieve reductions in run time at the expense of accuracy. For an overview of exact and approximate nearest neighbour search algorithms see Muja and Lowe (2014).

The remainder of this thesis is organised as follows. In Chapter 2 we study the problem of entropy estimation and use the estimator of Kozachenko and Leonenko (1987) as a starting point. Proposed in that chapter is a generalisation of this estimator that can be written as the weighted sum of Kozachenko–Leonenko estimators with different values of the tuning parameter. We focus on efficient estimation, in the sense of van der Vaart (1998), and on achieving the local asymptotic minimax lower bound, and find conditions in any fixed number of dimensions under which our estimator is efficient. Our results also show that the original Kozachenko–Leonenko estimator is efficient in up to 3 dimensions, under regularity conditions, but in general cannot be efficient in higher dimensions due to a non-trivial bias. Chapter 3 concerns independence testing, and we propose tests based on our entropy estimator of Chapter 2 when either the marginal distributions are known or unknown. We carry out a local power analysis of the test in the case of known marginals and prove the consistency of the test in the case of unknown marginals. In Chapter 4 we shift our attention to classification. We provide a new asymptotic expansion of the excess risk of the standard k-nearest neighbour estimator and use this to motivate a new k-nearest neighbour classifier for the semi-supervised setting in which k is allowed to depend on the test point. We provide theoretical and empirical arguments to show that this classifier can outperform the standard classifier in many settings.

Chapter 2

Efficient multivariate entropy estimation via k-nearest neighbour distances

2.1 Introduction

The concept of entropy plays a central role in information theory, and has found a wide array of uses in other disciplines, including statistics, probability and combinatorics. The *(differential) entropy* of a random vector X with density function f is defined as

$$H = H(X) = H(f) := -\mathbb{E}\{\log f(X)\} = -\int_{\mathcal{X}} f(x) \log f(x)\, dx$$

where $\mathcal{X} := \{x : f(x) > 0\}$. Introduced simultaneously in the highly influential Shannon (1948) and Wiener (1948), it represents the average information content of an observation, and is usually thought of as a measure of unpredictability. For an overview of its properties see, for example, Cover and Thomas (2012) or Wang, Kulkarni and Verdú (2008). Importantly, given constraints on certain moments and a support set one can find the distribution that maximises H. This leads to the principle of maximum entropy, which has found applications in areas such as selection of a prior distribution in Bayesian statistics (Jaynes, 1968) and density estimation (Buchen and Kelly, 1996).

In statistical contexts, it is often the estimation of entropy that is of primary interest, for instance in goodness-of-fit tests of normality (Vasicek, 1976) or uniformity (Cressie, 1976), tests of independence (Goria et al., 2005), independent component analysis (Miller and Fisher, 2003) and feature selection in classification (Kwak and Choi, 2002; Peng, Long and Ding, 2005). See, for example, Beirlant et al. (1997), Paninski (2003) and Wang, Kulkarni and Verdú (2008) for other applications and an overview of nonparametric techniques, which include methods based on sample spacings in the univariate case (e.g. El Haje Hussein and Golubev , 2009), histograms (Hall and Morton, 1993) and kernel density estimates (Paninski and Yajima, 2008; Sricharan, Wei and Hero, 2013), among others. The estimator of Kozachenko and Leonenko (1987) is particularly attractive as a starting point, both because it generalises easily to multivariate cases, and because, since it only relies on the evaluation of kth-nearest neighbour distances, it is straightforward to compute.

To introduce this estimator, for $n \geq 2$, let $X_1, \ldots, X_n$ be independent random vectors with density f on $\mathbb{R}^d$. Write $\|\cdot\|$ for the Euclidean norm on $\mathbb{R}^d$, and for $i = 1, \ldots, n$, let $X_{(1),i}, \ldots, X_{(n-1),i}$ denote a permutation of $\{X_1, \ldots, X_n\} \setminus \{X_i\}$ such that $\|X_{(1),i} - X_i\| \leq \ldots \leq \|X_{(n-1),i} - X_i\|$. For conciseness, we let

$$\rho_{(k),i} := \|X_{(k),i} - X_i\|$$

denote the distance between X_i and the kth nearest neighbour of X_i. The Kozachenko–Leonenko estimator of the entropy H is given by

$$\hat{H}_n = \hat{H}_n(X_1, \ldots, X_n) := \frac{1}{n} \sum_{i=1}^{n} \log\biggl(\frac{\rho_{(k),i}^d V_d (n-1)}{e^{\Psi(k)}} \biggr), \tag{2.1}$$

where $V_d := \pi^{d/2}/\Gamma(1 + d/2)$ denotes the volume of the unit d-dimensional Euclidean ball and where Ψ denotes the digamma function. In fact, this is a generalisation of the estimator originally proposed by Kozachenko and Leonenko (1987), which was defined for $k = 1$. For integers k we have $\Psi(k) = -\gamma + \sum_{j=1}^{k-1} 1/j$ where $\gamma := 0.577216\ldots$ is the Euler–Mascheroni constant, so that $e^{\Psi(k)}/k \to 1$ as $k \to \infty$. This estimator can be regarded as an attempt to mimic the 'oracle' estimator $H_n^* := -n^{-1} \sum_{i=1}^{n} \log f(X_i)$, based on a k-nearest neighbour density estimate that relies on the approximation

$$\frac{k}{n-1} \approx V_d \rho_{(k),1}^d f(X_1).$$

It turns out that, when $d \leq 3$ and other regularity conditions hold, the estimator $\hat{H}_n$ in (2.1) has the same asymptotic behaviour as H_n^*, in that

$$n^{1/2}(\hat{H}_n - H) \xrightarrow{d} N\bigl(0, \operatorname{Var} \log f(X_1)\bigr).$$

We will see that in such settings, this estimator is asymptotically efficient, in the sense of, e.g., van der Vaart (1998, p. 367). However, when $d \geq 4$, a non-trivial bias typically precludes its efficiency. Our main object of interest, therefore, will be a generalisation of the estimator (2.1), formed as a weighted average of Kozachenko–Leonenko estimators for different values of k, where the weights are chosen to try to cancel the dominant bias terms. More precisely, for a weight vector $w = (w_1, \ldots, w_k)^T \in \mathbb{R}^k$ with $\sum_{j=1}^{k} w_j = 1$, we consider the estimator

$$\hat{H}_n^w := \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} w_j \log \xi_{(j),i},$$

where $\xi_{(j),i} := e^{-\Psi(j)} V_d (n-1) \rho_{(j),i}^d$. Weighted estimators of this general type have been considered recently (e.g. Sricharan, Wei and Hero, 2013; Moon et al., 2016), though our construction of the weights and our analysis is new. In particular, we show that under stronger smoothness assumptions, and with a suitable choice of weights, the weighted Kozachenko–Leonenko estimator is efficient in arbitrary dimensions.

There have been several previous studies of the (unweighted) Kozachenko–Leonenko estimator, but results on the rate of convergence have until now confined either to the case $k = 1$ or (very recently) the case where k is fixed as n diverges. The original Kozachenko and Leonenko (1987) paper proved consistency of the estimator under mild conditions in the case $k = 1$. Tsybakov and Van der Meulen (1996) proved that the mean squared error of a truncated version of the estimator is $O(n^{-1})$ when $k = 1$ and $d = 1$ under a condition that is almost equivalent to an exponential tail; Biau and Devroye (2015) showed that the bias vanishes asymptotically while the variance is $O(n^{-1})$ when $k = 1$ and f is compactly supported and bounded away from zero on its support. Very recently, in independent work and under regularity conditions, Delattre and Fournier (2017) derived the asymptotic normality of the estimator when $k = 1$, confirming the suboptimal asymptotic variance in this case. Previous works on the general k case include Singh et al. (2003), where heuristic arguments were presented to suggest the estimator is consistent for general d and general fixed k and has variance $O(n^{-1})$ for $d = 1$ and general fixed k. Gao, Oh and Viswanath (2016) obtain a mean squared error bound of $O(n^{-1})$ up to polylogarithmic factors for fixed k and $d \leq 2$, though the only densities which the authors can show satisfy their tail condition have bounded support. Singh and Póczos (2016) obtain a similar bound (without the polylogarithmic factors, but explicitly assuming bounded support) for fixed k and $d \leq 4$. Mnatsakanov et al. (2008) allow k to diverge with n, and show that the estimator is consistent for general d.

Plug-in kernel methods are also popular for entropy estimation. Paninski and Yajima (2008), for example, show that a smaller bandwidth than would be required for a consistent density estimator can still yield a consistent entropy estimator. A k-nearest neighbour density estimate can be regarded as a kernel estimator with a bandwidth that depends both on the data and on the point at which the estimate is required. Sricharan, Wei and Hero (2013) obtain the parametric rate of convergence for a plug-in kernel method, assuming bounded support and at least d derivatives in the interior of the support.

Importantly, the class of densities considered in our results allows the support of the density to be unbounded; for instance, it may be the whole of $\mathbb{R}^d$. Such settings present significant new challenges and lead to different behaviour compared with more commonly-studied situations where the underlying density is compactly supported and bounded away from zero on its support. To gain intuition, consider the following second-order Taylor expansion of $H(f)$ around a density estimator $\hat{f}$:

$$H(f) \approx -\int_{\mathbb{R}^d} f(x) \log \hat{f}(x)\,dx - \frac{1}{2}\biggl(\int_{\mathbb{R}^d} \frac{f^2(x)}{\hat{f}(x)}\,dx - 1\biggr).$$

When f is bounded away from zero on its support, one can estimate the (smaller order) second term on the right-hand side, thereby obtaining efficient estimators of entropy in higher dimensions (Laurent, 1996); however, when f is not bounded away from zero on its support such procedures are no longer effective. To the best of our knowledge, therefore, this is the first time that a nonparametric entropy estimator has been shown to be efficient in multivariate settings for densities having unbounded support. (We remark that when $d = 1$, the histogram estimator of Hall and Morton (1993) is known to be efficient under fairly strong tail conditions.)

The outline of the rest of the chapter is as follows. In Section 2.2, we give our main results on the mean squared error and asymptotic normality of weighted Kozachenko–Leonenko estimators, and discuss confidence interval construction. These main results arise from asymptotic expansions for the bias and variance, which are stated in Section 2.3. Here, we also give examples to illustrate densities satisfying our conditions, discuss how they may be weakened, and address the fixed k case. Corresponding lower bounds are presented in Section 2.4. Proofs of main results are presented in Section 2.5 with auxiliary material and detailed bounds for various error terms deferred to Section 2.6.

We conclude the introduction with some notation used throughout this chapter. For $x \in \mathbb{R}^d$ and $r > 0$, let $B_x(r)$ be the closed Euclidean ball of radius r about x, and let $B_x^\circ(r) := B_x(r) \setminus \{x\}$ denote the corresponding punctured ball. We write $\|A\|_{\mathrm{op}}$ and $|A|$ for the operator norm and determinant, respectively, of $A \in \mathbb{R}^{d \times d}$, and let $\|A\|$ denote the vectorised Euclidean norm of a vector, matrix or array. For a smooth function $f : \mathbb{R}^d \to [0, \infty)$, we write $\dot{f}(x)$, $\ddot{f}(x)$ and $f^{(m)}(x)$ respectively for the gradient vector of f at x, Hessian matrix of f at x and the array with $(j_1, \ldots, j_m)$th entry $\frac{\partial^m f(x)}{\partial x_{j_1} \ldots \partial x_{j_m}}$. We also write $\Delta f(x) := \sum_{j=1}^d \frac{\partial^2 f}{\partial x_j^2}(x)$ for its Laplacian, and $\|f\|_\infty := \sup_{x \in \mathbb{R}^d} f(x)$ for its uniform norm.

2.2 Main results

We begin by introducing the class of densities over which our results will hold. Let $\mathcal{F}_d$ denote the class of all density functions with respect to Lebesgue measure on $\mathbb{R}^d$. For $f \in \mathcal{F}_d$ and $\alpha > 0$, let

$$\mu_\alpha(f) := \int_{\mathbb{R}^d} \|x\|^\alpha f(x) \, dx.$$

Now let $\mathcal{A}$ denote the class of decreasing functions $a : (0, \infty) \to [1, \infty)$ satisfying $a(\delta) = o(\delta^{-\epsilon})$ as $\delta \searrow 0$, for every $\epsilon > 0$. If $a \in \mathcal{A}$, $\beta > 0$ and $f \in \mathcal{F}_d$ is $m := \lceil \beta \rceil - 1$-times differentiable and $x \in \mathcal{X}$, we define $r_a(x) := \{8d^{1/2} a(f(x))\}^{-1/(\beta \wedge 1)}$ and

$$M_{f,a,\beta}(x) := \max\biggl\{ \max_{t=1,\ldots,m} \frac{\|f^{(t)}(x)\|}{f(x)}, \sup_{y \in B_x^\circ(r_a(x))} \frac{\|f^{(m)}(y) - f^{(m)}(x)\|}{f(x)\|y - x\|^{\beta - m}} \biggr\}.$$

The quantity $M_{f,a,\beta}(x)$ measures the smoothness of derivatives of f in neighbourhoods of x, relative to $f(x)$ itself. Note that these neighbourhoods of x are allowed to become smaller when $f(x)$ is small. Finally, for $\Theta := (0, \infty)^4 \times \mathcal{A}$, and $\theta = (\alpha, \beta, \nu, \gamma, a) \in \Theta$, let

$$\mathcal{F}_{d,\theta} := \biggl\{ f \in \mathcal{F}_d : \mu_\alpha(f) \leq \nu, \|f\|_\infty \leq \gamma, \sup_{x : f(x) \geq \delta} M_{f,a,\beta}(x) \leq a(\delta) \ \forall \delta > 0 \biggr\}.$$

We note here that Lemma 2.12 in the Section 2.6.2 can be used to derive a nestedness property of the classes with respect to the smoothness parameter, namely that if $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$, $\beta' \in (0, \beta)$ and $a'(\delta) = 15 d^{\lceil \beta \rceil / 2} a(\delta)$, then $\mathcal{F}_{d,\theta} \subseteq \mathcal{F}_{d,\theta'}$, where $\theta' = (\alpha, \beta', \gamma, \nu, a') \in \Theta$. In Section 2.3.2 below, we discuss the requirements of the class $\mathcal{F}_{d,\theta}$ in greater detail, and give several examples, including Gaussian and multivariate-t densities, which belong to $\mathcal{F}_{d,\theta}$ for suitable θ.

We now introduce the class of weights $w = (w_1, \ldots, w_k)^T$ that we consider. For $k \in \mathbb{N}$, let

$$\begin{aligned} \mathcal{W}^{(k)} := \biggl\{ w \in \mathbb{R}^k : & \sum_{j=1}^k w_j \frac{\Gamma(j + 2\ell/d)}{\Gamma(j)} = 0 \quad \text{for } \ell = 1, \ldots, \lfloor d/4 \rfloor \\ & \sum_{j=1}^k w_j = 1 \, \text{and} \ w_j = 0 \ \text{if} \, j \notin \{\lfloor k/d \rfloor, \lfloor 2k/d \rfloor, \ldots, k\} \biggr\}. \end{aligned} \tag{2.2}$$

Our main result below shows that for appropriately chosen weight vectors in $\mathcal{W}^{(k)}$, the normalised risk of the weighted Kozachenko–Leonenko estimator $\hat{H}_n^w$ converges in a uniform sense to that of the oracle estimator $H_n^* := -n^{-1} \sum_{i=1}^n \log f(X_i)$. Theorem 2.8 in Section 2.4 shows that this limiting risk is optimal.

Theorem 2.1. *Fix $d \in \mathbb{N}$ and $\theta = (\alpha, \beta, \nu, \gamma, a) \in \Theta$ with $\alpha > d$ and with $\beta > d/2$. Let $k_0^* = k_{0,n}^*$ and $k_1^* = k_{1,n}^*$ denote any two deterministic sequences of positive integers with $k_0^* \leq k_1^*$, with $k_0^*/\log^5 n \to \infty$ and with $k_1^* = O(n^{\tau_1})$ and $k_1^* = o(n^{\tau_2})$, where*

$$\tau_1 < \min\biggl(\frac{2\alpha}{5\alpha + 3d}, \frac{\alpha - d}{2\alpha}, \frac{4\beta^*}{4\beta^* + 3d}\biggr), \tau_2 := \min\biggl(1 - \frac{d/4}{1 + \lfloor d/4 \rfloor}, 1 - \frac{d}{2\beta}\biggr)$$

and $\beta^* := \beta \wedge 1$. There exists $k_d \in \mathbb{N}$, depending only on d, such that for each $k \geq k_d$, we can find $w = w^{(k)} \in \mathcal{W}^{(k)}$ with $\sup_{k \geq k_d} \|w^{(k)}\| < \infty$. For such w,

$$\sup_{k \in \{k_0^*, \ldots, k_1^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} n\mathbb{E}_f\bigl\{(\hat{H}_n^w - H_n^*)^2\bigr\} \to 0 \tag{2.3}$$

as $n \to \infty$. In particular,

$$\sup_{k \in \{k_0^*, \ldots, k_1^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} \bigl|n\mathbb{E}_f\{(\hat{H}_n^w - H(f))^2\} - V(f)\bigr| \to 0,$$

where $V(f) := \mathrm{Var}_f \log f(X_1) = \int_{\mathcal{X}} f \log^2 f - H(f)^2$.

We remark that the level of smoothness we require for efficiency in Theorem 2.1, namely $\beta > d/2$ is more than is needed for the two-stage estimator of Laurent (1996) in the case where f is compactly supported and bounded away from zero on its support, where $\beta > d/4$ suffices. As alluded to in the introduction, the fact that the function $x \mapsto -x \log x$ is non-differentiable at $x = 0$ means that the entropy functional is no longer smooth when f has full support, so the arguments of Laurent (1996) can no longer be applied and very different behaviour may occur (Lepski, Nemirovski and Spokoiny, 1999; Cai and Low, 2011).

It is also useful, e.g. for the purposes of constructing confidence intervals for the entropy, to understand the asymptotic normality of the estimator. To this end, let $\mathcal{H}$ denote the class of functions $h : \mathbb{R} \to \mathbb{R}$ with $\|h\|_\infty \leq 1$ and $|h(x) - h(y)| \leq |x - y|$ for all $x, y \in \mathbb{R}$. For probability measures P, Q on $\mathbb{R}$, we write

$$d_{\mathrm{BL}}(P, Q) := \sup_{h \in \mathcal{H}} \biggl| \int_{-\infty}^{\infty} h \, d(P - Q) \biggr|$$

for the bounded Lipschitz distance between P and Q. Recall that d_{BL} metrises weak convergence. The asymptotic variance $V(f)$ can be estimated analogously to $H(f)$ by $\hat{V}_n^w := \max(\tilde{V}_n^w, 0)$, where

$$\tilde{V}_n^w := \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} w_j \log^2 \xi_{(j),i} - (\hat{H}_n^w)^2.$$

Fixing $q \in (0, 1)$, this suggests that a natural asymptotic $(1 - q)$-level confidence interval for $H(f)$ is given by

$$I_{n,q} := \bigl[\hat{H}_n^w - n^{-1/2} z_{q/2} (\hat{V}_n^w)^{1/2}, \hat{H}_n^w + n^{-1/2} z_{q/2} (\hat{V}_n^w)^{1/2}\bigr],$$

where z_q is the $(1-q)$th quantile of the standard normal distribution; see also Delattre and Fournier (2017). Write $\mathcal{L}(Z)$ for the distribution of a random variable Z.

Theorem 2.2. *Under the conditions of Theorem 2.1, we have*

$$\sup_{k\in\{k_0^*,\ldots,k_1^*\}} \sup_{f\in\mathcal{F}_{d,\theta}} d_{\mathrm{BL}}\Bigl(\mathcal{L}\bigl(n^{1/2}(\hat{H}_n^w - H(f))\bigr), N\bigl(0, V(f)\bigr)\Bigr) \to 0$$

as $n \to \infty$. Consequently,

$$\sup_{q\in(0,1)} \sup_{k\in\{k_0^*,\ldots,k_1^*\}} \sup_{f\in\mathcal{F}_{d,\theta}} \Bigl|\mathbb{P}_f\bigl(I_{n,q} \ni H(f)\bigr) - (1-q)\Bigr| \to 0.$$

We remark that the choice $k = k_n = \lceil \log^6 n \rceil$ with $w = w^{(k)} \in \mathcal{W}^{(k)}$ satisfying $\sup_{k\geq k_d} \|w^{(k)}\| < \infty$ for the weighted Kozachenko–Leonenko estimator satisfies the conditions for efficiency in Theorem 2.1 whenever $f \in \mathcal{F}_{d,\theta}$ with $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$ satisfying $\alpha > d$ and $\beta > d/2$; knowledge of the precise values of α and β is not required. Moreover, the uniformity of the asymptotics in k means that if $\hat{k}_n = \hat{k}_n(X_1, \ldots, X_n)$ is a data-driven choice of k, the conclusions Theorem 2.2 remain valid provided that $\mathbb{P}(\hat{k}_n < k_0^*) + \mathbb{P}(\hat{k}_n > k_1^*) \to 0$.

2.3 Bias and variance expansions for Kozachenko–Leonenko estimators

2.3.1 Bias

The proof of (2.3) is derived from separate expansions for the bias and variance of the weighted Kozachenko–Leonenko estimator, and we treat the bias in this subsection. To gain intuition, we initially focus for simplicity of exposition on the unweighted estimator

$$\hat{H}_n = \frac{1}{n}\sum_{i=1}^n \log \xi_i,$$

where we have written ξ_i as shorthand for $\xi_{(k),i}$. For $x \in \mathbb{R}^d$ and $u \in [0, \infty)$, we introduce the sequence of distribution functions

$$F_{n,x}(u) := \mathbb{P}(\xi_i \leq u | X_i = x) = \sum_{j=k}^{n-1} \binom{n-1}{j} p_{n,x,u}^j (1 - p_{n,x,u})^{n-1-j},$$

where

$$p_{n,x,u} := \int_{B_x(r_{n,u})} f(y)\, dy \qquad \text{and} \qquad r_{n,u} := \biggl\{\frac{e^{\Psi(k)}u}{V_d(n-1)}\biggr\}^{1/d}.$$

Further, for $u \in [0, \infty)$, define the limiting (Gamma) distribution function

$$F_x(u) := \exp\{-uf(x)e^{\Psi(k)}\} \sum_{j=k}^{\infty} \frac{1}{j!}\{uf(x)e^{\Psi(k)}\}^j = e^{-\lambda_{x,u}} \sum_{j=k}^{\infty} \frac{\lambda_{x,u}^j}{j!},$$

where $\lambda_{x,u} := uf(x)e^{\Psi(k)}$. That this is the limit distribution for each fixed k follows from a Poisson approximation to the Binomial distribution and the Lebesgue differentiation theorem. We therefore expect that

$$\begin{aligned}\mathbb{E}(\hat{H}_n) &= \int_{\mathcal{X}} f(x) \int_0^\infty \log u \, dF_{n,x}(u) \, dx \approx \int_{\mathcal{X}} f(x) \int_0^\infty \log u \, dF_x(u) \, dx \\ &= \int_{\mathcal{X}} f(x) \int_0^\infty \log\Bigl(\frac{te^{-\Psi(k)}}{f(x)}\Bigr) e^{-t} \frac{t^{k-1}}{(k-1)!} \, dt \, dx = H.\end{aligned}$$

Although we do not explicitly use this approximation in our asymptotic analysis of the bias, it motivates much of our development. It also explains the reason for using $e^{\Psi(k)}$ in the definition of $\xi_{(k),i}$, rather than simply k. Lemma 2.3 below gives an expression for the asymptotic bias of the unweighted Kozachenko–Leonenko estimator.

Lemma 2.3. *Fix $d \in \mathbb{N}$ and $\theta = (\alpha, \beta, \nu, \gamma, a) \in \Theta$. Let $k^* = k_n^*$ denote any deterministic sequence of positive integers with $k^* = O(n^{1-\epsilon})$ as $n \to \infty$ for some $\epsilon > 0$. Then there exist $\lambda_1, \ldots, \lambda_{\lceil \beta/2 \rceil - 1} \in \mathbb{R}$, depending only on f and d, such that $\sup_{f \in \mathcal{F}_{d,\theta}} \max_{l=1,\ldots,\lceil \beta/2 \rceil - 1} |\lambda_l| < \infty$ and for each $\epsilon > 0$,*

$$\sup_{f \in \mathcal{F}_{d,\theta}} \biggl| \mathbb{E}_f(\hat{H}_n) - H - \sum_{l=1}^{\lceil \beta/2 \rceil - 1} \frac{\Gamma(k + 2l/d)\Gamma(n)}{\Gamma(k)\Gamma(n + 2l/d)} \lambda_l \biggr| = O\biggl(\max\Bigl\{ \frac{k^{\frac{\alpha}{\alpha+d} - \epsilon}}{n^{\frac{\alpha}{\alpha+d} - \epsilon}}, \frac{k^{\frac{\beta}{d}}}{n^{\frac{\beta}{d}}} \Bigr\} \biggr)$$

as $n \to \infty$, uniformly for $k \in \{1, \ldots, k^*\}$, where $\lambda_l = 0$ if $2l \geq d\alpha/(\alpha + d)$.

When $d \geq 3$, $\alpha > 2d/(d-2)$ and $\beta > 2$, we have

$$\lambda_1 = -\frac{1}{2(d+2)V_d^{2/d}} \int_{\mathcal{X}} \frac{\Delta f(x)}{f(x)^{2/d}} \, dx,$$

which is finite under these assumptions; cf. the second part of Proposition 2.9 in Section 2.5.1. Moreover, since, for each $l > 0$, we have $\frac{\Gamma(n)}{\Gamma(n+2l/d)} = n^{-2l/d}\{1 + O(n^{-1})\}$, we deduce from Lemma 2.3 that in this setting,

$$\sup_{f \in \mathcal{F}_{d,\theta}} \biggl| \mathbb{E}_f(\hat{H}_n) - H + \frac{\Gamma(k + 2/d)}{2(d+2)V_d^{2/d}\Gamma(k)n^{2/d}} \int_{\mathcal{X}} \frac{\Delta f(x)}{f(x)^{2/d}} \, dx \biggr| = o\Bigl(\frac{k^{2/d}}{n^{2/d}}\Bigr).$$

In particular, when $d \geq 4$ and $\int_{\mathcal{X}} \frac{\Delta f(x)}{f(x)^{2/d}} \, dx \neq 0$, the bias of the unweighted Kozachenko–Leonenko estimator precludes its efficiency.

On the other hand, Lemma 2.3 motivates the definition of the class of weight vectors $\mathcal{W}^{(k)}$ in (2.2), and facilitates the expansion for the bias of the weighted Kozachenko–Leonenko estimator in Corollary 2.4 below. In particular, since $2(\lfloor d/4 \rfloor + 1)/d > 1/2$, we see that this result provides conditions under which the bias is $o(n^{-1/2})$ for suitably chosen k. This explains why we let ℓ take values in the range $\{1, \ldots, \lfloor d/4 \rfloor\}$ in (2.2).

Corollary 2.4. *Assume the conditions of Lemma 2.3. If $w = w^{(k)} \in \mathcal{W}^{(k)}$ for $k \geq k_d$ and $\sup_{k \geq k_d} \|w^{(k)}\| < \infty$, then for every $\epsilon > 0$,*

$$\sup_{f \in \mathcal{F}_{d,\theta}} \bigl| \mathbb{E}_f(\hat{H}_n^w) - H(f) \bigr| = O\biggl(\max\Bigl\{ \frac{k^{\frac{\alpha}{\alpha+d} - \epsilon}}{n^{\frac{\alpha}{\alpha+d} - \epsilon}}, \frac{k^{\frac{2(\lfloor d/4 \rfloor + 1)}{d}}}{n^{\frac{2(\lfloor d/4 \rfloor + 1)}{d}}}, \frac{k^{\frac{\beta}{d}}}{n^{\frac{\beta}{d}}} \Bigr\} \biggr),$$

uniformly for $k \in \{1, \ldots, k^*\}$.

The proof of Lemma 2.3 is given in Section 2.5.1, but we present here some of the main ideas that are particularly relevant for the case $d \geq 3$, $\alpha > 2d/(d-2)$ and $\beta \in (2,4]$. First, note that

$$\frac{dF_{n,x}(u)}{du} = \mathrm{B}_{k,n-k}(p_{n,x,u})\frac{\partial p_{n,x,u}}{\partial u}, \tag{2.4}$$

where $\mathrm{B}_{a,b}(s) := \mathrm{B}_{a,b}^{-1}s^{a-1}(1-s)^{b-1}$ denotes the density of a Beta(a,b) random variable at $s \in (0,1)$, with $\mathrm{B}_{a,b} := \Gamma(a)\Gamma(b)/\Gamma(a+b)$. For $x \in \mathcal{X}$ and $r > 0$, define $h_x(r) := \int_{B_x(r)} f(y)\,dy$. Since $h_x(r)$ is a continuous, non-decreasing function of r, we can define a left-continuous inverse for $s \in (0,1)$ by

$$h_x^{-1}(s) := \inf\{r > 0 : h_x(r) \geq s\} = \inf\{r > 0 : h_x(r) = s\}, \tag{2.5}$$

so that $h_x(r) \geq s$ if and only if $r \geq h_x^{-1}(s)$. We use the approximation

$$V_d f(x) h_x^{-1}(s)^d \approx s - \frac{s^{1+2/d}\Delta f(x)}{2(d+2)V_d^{2/d}f(x)^{1+2/d}}$$

for small $s > 0$, which is formalised in Lemma 2.10(ii) in Section 2.5.1. In the case $d \geq 3$, $\alpha > 2d/(d-2)$ and $\beta \in (2,4]$, the proof of Lemma 2.3 can be seen as justifying the use of the above approximation in the following:

$$\begin{aligned}
\mathbb{E}(\hat{H}_n) &= \int_{\mathcal{X}} f(x) \int_0^\infty \log u \, dF_{n,x}(u)\,dx = \int_{\mathcal{X}} f(x)\int_0^1 \log\Big(\frac{V_d(n-1)h_x^{-1}(s)^d}{e^{\Psi(k)}}\Big)\mathrm{B}_{k,n-k}(s)\,ds\,dx \\
&\approx \int_{\mathcal{X}} f(x)\int_0^1 \Big\{\log\Big(\frac{(n-1)s}{e^{\Psi(k)}f(x)}\Big) - \frac{V_d^{-2/d}s^{2/d}\Delta f(x)}{2(d+2)f(x)^{1+2/d}}\Big\}\mathrm{B}_{k,n-k}(s)\,ds\,dx \\
&= \log(n-1) - \Psi(n) + H - \frac{V_d^{-2/d}\Gamma(k+2/d)\Gamma(n)}{2(d+2)\Gamma(k)\Gamma(n+2/d)}\int_{\mathcal{X}} \frac{\Delta f(x)}{f(x)^{2/d}}\,dx.
\end{aligned}$$

Note that $\log(n-1) - \Psi(n) = -1/(2n) + o(1/n)$, which leads to the given bias expression. The proof in other cases proceeds along similar lines. These heuristics make clear that the function $h_x^{-1}(\cdot)$ plays a key role in understanding the bias. This function is in general complicated, though some understanding can be gained from the following uniform density example, where it can be evaluated explicitly. This leads to an exact expression for the bias, even though the discontinuities mean that the density does not belong to $\mathcal{F}_{1,\theta}$ for any $\theta \in \Theta$.

Example 2.1. Consider the uniform distribution, $U[0,1]$. For $x \leq 1/2$, we have

$$h_x^{-1}(s) = \begin{cases} s/2, & \text{if } s \leq 2x \\ s - x, & \text{if } 2x < s \leq 1. \end{cases}$$

It therefore follows that

$$\begin{aligned}
\mathbb{E}(\hat{H}_n) - H &= 2\int_0^{1/2}\int_0^\infty \log u\,dF_{n,x}(u)\,dx = 2\int_0^{1/2}\int_0^1 \log\Big(\frac{2(n-1)h_x^{-1}(s)}{e^{\Psi(k)}}\Big)\mathrm{B}_{k,n-k}(s)\,ds\,dx \\
&= 2\int_0^1 \mathrm{B}_{k,n-k}(s)\Big\{\int_0^{s/2}\log(2(s-x))\,dx + \int_{s/2}^{1/2}\log s\,dx\Big\}\,ds + \log\Big(\frac{n-1}{e^{\Psi(k)}}\Big) \\
&= \frac{k}{n}(\log 4 - 1) + \log(n-1) - \Psi(n).
\end{aligned}$$

2.3.2 Discussion of conditions and weakening of conditions

Recall the definitions of the quantity $M_{f,a,\beta}(x)$ and $\mathcal{A}$ from Section 2.2. In addition to standard moment and boundedness assumptions, the condition $f \in \mathcal{F}_{d,\theta}$ requires that

$$\sup_{x: f(x) \geq \delta} M_{f,a,\beta}(x) \leq a(\delta) \quad \text{for all } \delta > 0 \text{ and some } a \in \mathcal{A}. \tag{2.6}$$

In this subsection, we explore the condition (2.6) further, with the aid of several examples.

The condition (2.6) is reminiscent of more standard Hölder smoothness assumptions, though we also require that the partial derivatives of the density vary less where f is small. On the other hand, we also allow the neighbourhoods of x in the definition of $M_{f,a,\beta}(x)$ to shrink where $f(x)$ is small. Roughly speaking, the condition requires that the partial derivatives of the density decay nearly as fast as the density itself in the tails of the distribution. As a simple stability property, if (2.6) holds for a density f_0, then it also holds for any density from the location-scale family:

$$\{f_\Sigma(\cdot) = |\Sigma|^{-1/2} f_0\bigl(\Sigma^{-1/2}(\cdot - \mu)\bigr) : \mu \in \mathbb{R}^d, \Sigma = \Sigma^T \in \mathbb{R}^{d \times d} \text{ positive definite}\}.$$

This observation allows us to consider canonical representatives of location-scale families in the examples below.

Proposition 2.5. *For each of the following densities f, and for each $d \in \mathbb{N}$, there exists $\theta \in \Theta$ such that $f \in \mathcal{F}_{d,\theta}$:*

(i) $f(x) = f(x_1, \ldots, x_d) = (2\pi)^{-d/2} e^{-\|x\|^2/2}$*, the standard normal density;*

(ii) $f(x) = f(x_1, \ldots, x_d) \propto (1 + \|x\|^2/\rho)^{-\frac{d+\rho}{2}}$*, the multivariate-t distribution with $\rho > 0$ degrees of freedom.*

Moreover, the following univariate density f also belongs to $\mathcal{F}_{1,\theta}$ for suitable $\theta \in \Theta$:

$$f(x) \propto \exp\Bigl(-\frac{1}{1 - x^2}\Bigr) \mathbb{1}_{\{x \in (-1,1)\}}.$$

The final part of Proposition 2.5 is included because it provides an example of a density f that belongs to $\mathcal{F}_{1,\theta}$ for suitable $\theta \in \Theta$, even though there exist points $x_0 \in \mathbb{R}$ with $f(x_0) = 0$.

On the other hand, there are also examples, such as Example 2.2 below, where the behaviour of f near a point x_0 with $f(x_0) = 0$ precludes f belonging to $\mathcal{F}_{d,\theta}$ for any $\theta \in \Theta$. To provide some guarantees in such settings, we now give a very general condition under which our approach to studying the bias can be applied.

Proposition 2.6. *Assume that f is bounded, that $\mu_\alpha(f) < \infty$ for some $\alpha > 0$, and let k^* be as in Lemma 2.3. Let $a_n := 3(k+1)\log(n-1)$, let $r_x := \bigl\{\frac{2a_n}{V_d(n-1)f(x)}\bigr\}^{1/d}$, and assume further that there exists $\beta > 0$ such that the function on $\mathcal{X}$ given by*

$$C_{n,\beta}(x) := \begin{cases} \sup_{y \in B_x^\circ(r_x)} |f(y) - f(x)|/\|y - x\|^\beta & \textit{if } \beta \leq 1, \\ \sup_{y \in B_x^\circ(r_x)} \|\dot{f}(y) - \dot{f}(x)\|/\|y - x\|^{\beta - 1} & \textit{if } \beta > 1, \end{cases}$$

is real-valued. Suppose that $\mathcal{X}_n \subseteq \mathcal{X}$ is such that

$$\sup_{x \in \mathcal{X}_n} \Bigl(\frac{a_n}{n-1}\Bigr)^{\tilde{\beta}/d} \frac{C_{n,\tilde{\beta}}(x)}{f(x)^{1+\tilde{\beta}/d}} \to 0 \tag{2.7}$$

as $n \to \infty$, where $\tilde{\beta} := \beta \wedge 2$. Then writing $q_n := \int_{\mathcal{X}_n^c} f$, we have for every $\epsilon > 0$ that

$$\mathbb{E}_f(\hat{H}_n) - H = O\biggl(\max\biggl\{\frac{k^{\tilde{\beta}/d}}{n^{\tilde{\beta}/d}} \int_{\mathcal{X}_n} \frac{C_{n,\tilde{\beta}}(x)}{f(x)^{\tilde{\beta}/d}}\,dx\,,\, q_n^{1-\epsilon}\,,\, q_n \log n\,,\, \frac{1}{n}\biggr\}\biggr), \tag{2.8}$$

uniformly for $k \in \{1, \ldots, k^*\}$.

To aid interpretation of Proposition 2.6, we first remark that if $f \in \mathcal{F}_{d,\theta}$ for some $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$, then (2.7) holds, with $\mathcal{X}_n := \{x \in \mathcal{X} : f(x) \geq \delta_n\}$, where δ_n is defined in (2.12) below. On the other hand, if $f \notin \mathcal{F}_{d,\theta}$, we may still be able to obtain explicit bounds on the terms in (2.8) on a case-by-case basis, as in the following example.

Example 2.2. For $a > 1$, consider $f(x) = \Gamma(a)^{-1} x^{a-1} e^{-x} \mathbb{1}_{\{x>0\}}$, the density of the $\Gamma(a, 1)$ distribution. Then for any $\tau \in (0, 1)$ small enough, we may take

$$\mathcal{X}_n = \biggl[\Bigl(\frac{k}{n}\Bigr)^{\frac{1}{a}-\tau}, (1-\tau)\log\frac{n}{k}\biggr]$$

to deduce from Proposition 2.6 that for every $\epsilon > 0$,

$$\mathbb{E}_f(\hat{H}_n) - H = o\Bigl(\frac{k^{1-\epsilon}}{n^{1-\epsilon}}\Bigr),$$

uniformly for $k \in \{1, \ldots, k^*\}$.

Similar calculations show that the bias is of the same order for $\mathrm{Beta}(a, b)$ distributions with $a, b > 1$.

2.3.3 Asymptotic variance and normality

We now study the asymptotic variance of Kozachenko–Leonenko estimators under the assumption that the tuning parameter k is diverging with n; the fixed k case is deferred to the next subsection.

Lemma 2.7. *Let $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$ with $\alpha > d$ and $\beta > 0$. Let $k_0^* = k_{0,n}^*$ and $k_1^* = k_{1,n}^*$ denote any two deterministic sequences of positive integers with $k_0^* \leq k_1^*$, with $k_0^*/\log^5 n \to \infty$ and with $k_1^* = O(n^{\tau_1})$, where τ_1 satisfies the condition in Theorem 2.1. Then for any $w = w^{(k)} \in \mathcal{W}^{(k)}$ with $\sup_{k \geq k_d} \|w^{(k)}\| < \infty$, we have*

$$\sup_{k \in \{k_0^*, \ldots, k_1^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} \bigl|n \mathrm{Var}_f \hat{H}_n^w - V(f)\bigr| \to 0$$

as $n \to \infty$.

The proof of this lemma is lengthy, and involves many delicate error bounds, so we outline the main ideas in the unweighted case here. First, we argue that

$$\begin{aligned} \mathrm{Var}\,\hat{H}_n &= n^{-1}\,\mathrm{Var}\log\xi_1 + (1 - n^{-1})\,\mathrm{Cov}(\log\xi_1, \log\xi_2) \\ &= n^{-1}V(f) + \mathrm{Cov}\bigl(\log(\xi_1 f(X_1)), \log(\xi_2 f(X_2))\bigr) + o(n^{-1}), \end{aligned}$$

where we hope to exploit the fact that $\xi_1 f(X_1) \xrightarrow{p} 1$. The main difficulties in the argument are caused by the fact that handling the covariance above requires us to study the joint distribution of (ξ_1, ξ_2), and this is complicated by the fact that X_2 may be one of the k nearest neighbours of

X_1 or vice versa, and more generally, X_1 and X_2 may have some of their k nearest neighbours in common. Dealing carefully with the different possible events requires us to consider separately the cases where $f(X_1)$ is small and large, as well as the proximity of X_2 to X_1. Finally, however, we can apply a normal approximation to the relevant multinomial distribution (which requires that $k \to \infty$) to deduce the result. We remark that under stronger conditions on k, it should also be possible to derive the same conclusion about the asymptotic variance of $\hat{H}_n$ while only assuming similar conditions on the density to those required in Proposition 2.6, but we do not pursue this here.

2.3.4 Fixed k

A crucial step in the proof of Lemma 2.7 is the normal approximation to a certain multinomial distribution (cf. the bound on the term W_4). This normal approximation is only valid when $k \to \infty$ as $n \to \infty$. In this subsection, we present evidence to suggest that, when k is fixed (i.e. not depending on n), then Kozachenko–Leonenko estimators are inefficient. For simplicity, we focus on the unweighted version of estimator.

Define the functions

$$\alpha_r(s,t) := \frac{1}{V_d}\mu_d\bigl(B_0(s^{1/d}) \cap B_{r^{1/d}e_1}(t^{1/d})\bigr),$$

where $e_1 = (1, 0, \dots, 0)$ is the first element of the standard basis for $\mathbb{R}^d$ and μ_d denotes Lebesgue measure on $\mathbb{R}^d$. Also define the functions T_k on $[0,\infty)^3$ by

$$T_k(r,s,t) := e^{\alpha_r(s,t)} \sum_{\ell=0}^{L(r,s,t)} \sum_{i=0}^{I(r,s)-\ell} \sum_{j=0}^{J(r,t)-\ell} \frac{\{s-\alpha_r(s,t)\}^i\{t-\alpha_r(s,t)\}^j\alpha_r^\ell(s,t)}{i!j!\ell!} - \sum_{i=0}^{I(r,s)} \sum_{j=0}^{J(r,t)} \frac{s^it^j}{i!j!},$$

where $L(r,s,t) := k-1-\mathbb{1}_{\{r<\max(s,t)\}}$, $I(r,s) := k-1-\mathbb{1}_{\{r<s\}}$, $J(r,t) := k-1-\mathbb{1}_{\{r<t\}}$.

In the case $k = 1$, this function appears in Delattre and Fournier (2017), where the authors show that, under certain regularity conditions,

$$\lim_{n\to\infty} n \operatorname{Var} \hat{H}_n - V(f) = \Psi'(1) + \int_{[0,\infty)^3} e^{-s-t}\frac{T_1(r,s,t)}{st}\,dr\,ds\,dt - 1 + 2\log 2.$$

More generally, Poisson approximation to the same multinomial distribution mentioned above, together with analysis similar to the proof of Lemma 2.7, suggests that for (fixed) $k \geq 2$,

$$\begin{aligned}\lim_{n\to\infty} n \operatorname{Var} \hat{H}_n - V(f) =\ & \Psi'(k) + \int_{[0,\infty)^3} e^{-s-t}\frac{T_k(r,s,t)}{st}\,dr\,ds\,dt - 1 \\ & + 2^{-(2k-2)}\binom{2k-2}{k-1}\{\Psi(2k-1) - \Psi(k) - \log 2\} \\ & + \frac{1}{k-1}\sum_{j=0}^{k-2} 2^{-k-j}\binom{k+j-1}{j}[1-(k-j)\{\Psi(k+j) - \log 2 - \Psi(k)\}]. \qquad (2.9)\end{aligned}$$

Here, the $\Psi'(k)$ term arises as in (2.18), the integral term arises from the Poisson approximation, the -1 arises as in (2.27), and the remaining terms come from the fact that X_1 can be one of the k nearest neighbours of X_2, or vice-versa, which induces a singular component into the joint distribution function $F_{n,x,y}$ of (ξ_1, ξ_2) given $(X_1, X_2) = (x, y)$. It is interesting to observe that this

$d\backslash k$	1	2	3	4	5
1	2.14	0.97	0.64	0.48	0.39
2	2.29	1.01	0.64	0.47	0.38
3	2.42	1.03	0.64	0.47	0.37
5	2.61	1.05	0.65	0.47	0.37
10	2.85	1.10	0.68	0.50	0.40

Table 2.1: Asymptotic variance inflation (2.9) of the Kozachenko–Leonenko estimator for fixed k.

asymptotic inflation of the variance is distribution-free; in Table 2.1, we tabulate numerical values for (2.9) for a few values of d and k. These agree with those obtained by Delattre and Fournier (2017) for the case $k = 1$.

2.4 Lower bounds

In this section, we address the optimality in a local asymptotic minimax sense of the limiting normalised risk $V(f)$ given in Theorem 2.1 using ideas of semiparametric efficiency (e.g. van der Vaart, 1998, Chapter 25). For $f \in \mathcal{F}_{d,\theta}$, $t \geq 0$ and a Borel measurable function $g : \mathbb{R}^d \to \mathbb{R}$, define $f_{t,g} : \mathbb{R}^d \to [0, \infty)$ by

$$f_{t,g}(x) := \frac{2c(t)}{1 + e^{-2tg(x)}} f(x), \tag{2.10}$$

where $c(t) := \bigl(\int_{\mathbb{R}^d} \frac{2}{1+e^{-2tg(x)}} f(x)\, dx\bigr)^{-1}$. This definition ensures that $\{f_{t,g} : t \geq 0\}$ is differentiable in quadratic mean at $t = 0$ with score function g (e.g. van der Vaart, 1998, Example 25.16). We say $(\tilde{H}_n)$ is an estimator sequence if $\tilde{H}_n : (\mathbb{R}^d)^{\times n} \to \mathbb{R}$ is a measurable function for each $n \in \mathbb{N}$.

Theorem 2.8. *Fix $d \in \mathbb{N}$, $\theta = (\alpha, \beta, \gamma, \mu, a) \in \Theta$ and $f \in \mathcal{F}_{d,\theta}$. For $\lambda \in \mathbb{R}$, let $g_\lambda := \lambda\{\log f + H(f)\}$. Then, writing $\mathcal{I}$ for the set of finite subsets of $\mathbb{R}$, we have for any estimator sequence $(\tilde{H}_n)$ that*

$$\sup_{I \in \mathcal{I}} \liminf_{n \to \infty} \max_{\lambda \in I} n\mathbb{E}_{f_{n^{-1/2},g_\lambda}} \bigl[\{\tilde{H}_n - H(f_{n^{-1/2},g_\lambda})\}^2\bigr] \geq V(f). \tag{2.11}$$

Moreover, whenever $t|\lambda| \leq \min(1, \{144V(f)\}^{-1/2})$, with $\tilde{\theta} := \bigl(\alpha, \beta, 4\gamma, 4\mu, \tilde{a}\bigr) \in \Theta$ and $\tilde{a} \in \mathcal{A}$ defined in (2.89) in Section 2.6.6 we have that $f_{t,g_\lambda} \in \mathcal{F}_{d,\tilde{\theta}}$.

The proof of Theorem 2.8 reveals that, at every $f \in \mathcal{F}_{d,\theta}$, the entropy functional H is differentiable relative to the tangent set $\{g_\lambda : \lambda \in \mathbb{R}\}$ with efficient influence function

$$\tilde{\psi}_f := -\log f - H(f).$$

This observation, together with Theorem 2.1, confirms that under the assumptions on θ, w and k in that result, the weighted Kozachenko–Leonenko estimator $\hat{H}_n^w$ is (asymptotically) efficient at $f \in \mathcal{F}_{d,\theta}$ in the sense that

$$n^{1/2}\{\hat{H}_n^w - H(f)\} = \frac{1}{n^{1/2}} \sum_{i=1}^n \tilde{\psi}_f(X_i) + o_p(1)$$

(cf. van der Vaart, 1998, p. 367). Moreover, the second part of Theorem 2.8 and Theorem 2.1 imply in particular that, under these same conditions on θ, w and k, the estimator $\hat{H}_n^w$ attains the local asymptotic minimax lower bound, in the sense that

$$\sup_{I\in\mathcal{I}}\lim_{n\to\infty}\max_{\lambda\in I} n\mathbb{E}_{f_{n^{-1/2},g_\lambda}}\bigl[\bigl\{\hat{H}_n^w - H(f_{n^{-1/2},g_\lambda})\bigr\}^2\bigr] = V(f).$$

2.5 Proofs of main results

2.5.1 Auxiliary results and proofs of Lemma 2.3 and Corollary 2.4

Throughout the proofs, we write $a \lesssim b$ to mean that there exists $C > 0$, depending only on $d \in \mathbb{N}$ and $\theta \in \Theta$, such that $a \leq Cb$. The proof of Lemma 2.3 relies on the following two auxiliary results, whose proofs are given in Section 2.6.1.

Proposition 2.9. *Let $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$, $d \in \mathbb{N}$ and $\tau \in \bigl(\frac{d}{\alpha+d}, 1\bigr]$. Then*

$$\sup_{f\in\mathcal{F}_{d,\theta}}\int_{\{x:f(x)<\delta\}} a\bigl(f(x)\bigr)f(x)^\tau\,dx \to 0$$

as $\delta \searrow 0$. Moreover, for every $\rho > 0$,

$$\sup_{f\in\mathcal{F}_{d,\theta}}\int_{\mathcal{X}} a\bigl(f(x)\bigr)^\rho f(x)^\tau < \infty.$$

Recall the definition of $h_x^{-1}(\cdot)$ in (2.5). The first part of Lemma 2.10 below provides crude but general bounds; the second gives much sharper bounds in a more restricted region.

Lemma 2.10. *(i) Let $f \in \mathcal{F}_d$ and let $\alpha > 0$. Then for every $s \in (0,1)$ and $x \in \mathbb{R}^d$,*

$$\Bigl(\frac{s}{V_d\|f\|_\infty}\Bigr)^{1/d} \leq h_x^{-1}(s) \leq \|x\| + \Bigl(\frac{\mu_\alpha(f)}{1-s}\Bigr)^{1/\alpha}.$$

(ii) Fix $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$, and let $\mathcal{S}_n \subseteq (0,1)$, $\mathcal{X}_n \subseteq \mathbb{R}^d$ be such that

$$C_n := \sup_{f\in\mathcal{F}_{d,\theta}}\sup_{s\in\mathcal{S}_n}\sup_{x\in\mathcal{X}_n}\frac{a(f(x))^{d/(1\wedge\beta)}s}{f(x)} \to 0.$$

Then there exists $n_* = n_*(d,\theta) \in \mathbb{N}$ such that for all $n \geq n_*$, $s \in \mathcal{S}_n$, $x \in \mathcal{X}_n$ and $f \in \mathcal{F}_{d,\theta}$, we have

$$\biggl|V_d f(x)h_x^{-1}(s)^d - \sum_{l=0}^{\lceil\beta/2\rceil-1} b_l(x)s^{1+2l/d}\biggr| \lesssim s\biggl\{\frac{a(f(x))^{d/(2\wedge\beta)}s}{f(x)}\biggr\}^{\beta/d},$$

where $b_0(x) = 1$ and $|b_l(x)| \lesssim a(f(x))^l f(x)^{-2l/d}$ for $l \geq 1$. Moreover, if $\beta > 2$, then

$$b_1(x) = -\frac{\Delta f(x)}{2(d+2)V_d^{2/d}f(x)^{1+2/d}}.$$

We are now in a position to prove Lemma 2.3.

Proof of Lemma 2.3. (i) We initially prove the result in the case $d \geq 3$, $\alpha > 2d/(d-2)$ and

$\beta \in (2,4]$, where it suffices to show that

$$\sup_{f \in \mathcal{F}_{d,\theta}} \biggl| \mathbb{E}_f(\hat{H}_n) - H + \frac{\Gamma(k+2/d)\Gamma(n)}{2(d+2)V_d^{2/d}\Gamma(k)\Gamma(n+2/d)} \int_{\mathcal{X}} \frac{\Delta f(x)}{f(x)^{2/d}}\,dx \biggr| = O\biggl(\max\biggl\{ \frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}, \frac{k^{\frac{\beta}{d}}}{n^{\frac{\beta}{d}}} \biggr\}\biggr)$$

as $n \to \infty$, uniformly for $k \in \{1,\ldots,k^*\}$. Fix $f \in \mathcal{F}_{d,\theta}$. Define $c_n := a(k/(n-1))^{1/(1\wedge\beta)}$, let

$$\delta_n := kc_n^d \log^2(n-1)/(n-1) \tag{2.12}$$

and let $\mathcal{X}_n := \{x : f(x) \geq \delta_n\}$. Recall that $a_n := 3(k+1)\log(n-1)$ and let

$$u_{x,s} := \frac{V_d(n-1)h_x^{-1}(s)^d}{e^{\Psi(k)}}.$$

The proof is based on (2.4) and Lemma 2.10(ii), which allow us to make the transformation $s = p_{n,x,u} = h_x(r_{n,u})$. Writing R_i, $i = 1,\ldots,5$ for remainder terms to be bounded at the end of the proof, we can write

$$\begin{aligned}
&\mathbb{E}(\hat{H}_n) = \int_{\mathcal{X}} f(x) \int_0^\infty \log u \, dF_{n,x}(u)\,dx \\
&= \int_{\mathcal{X}_n} f(x) \int_0^1 \mathrm{B}_{k,n-k}(s) \log u_{x,s}\,ds\,dx + R_1 \\
&= \int_{\mathcal{X}_n} f(x) \int_0^{\frac{a_n}{n-1}} \mathrm{B}_{k,n-k}(s) \log u_{x,s}\,ds\,dx + R_1 + R_2 \\
&= \int_{\mathcal{X}_n} f(x) \int_0^{\frac{a_n}{n-1}} \Bigl\{ \log\Bigl(\frac{(n-1)s}{e^{\Psi(k)}f(x)} \Bigr) - \frac{V_d^{-2/d}s^{2/d}\Delta f(x)}{2(d+2)f(x)^{1+2/d}} \Bigr\} \mathrm{B}_{k,n-k}(s)\,ds\,dx + \sum_{i=1}^3 R_i \\
&= \int_{\mathcal{X}_n} f(x) \Bigl\{ \log\Bigl(\frac{n-1}{f(x)}\Bigr) - \Psi(n) - \frac{V_d^{-2/d}\mathrm{B}_{k+2/d,n-k}\,\Delta f(x)}{2(d+2)\mathrm{B}_{k,n-k}\,f(x)^{1+2/d}} \Bigr\}\,dx + \sum_{i=1}^4 R_i \\
&= H + \log(n-1) - \Psi(n) - \frac{V_d^{-2/d}\Gamma(k+2/d)\Gamma(n)}{2(d+2)\Gamma(k)\Gamma(n+2/d)} \int_{\mathcal{X}_n} \frac{\Delta f(x)}{f(x)^{2/d}}\,dx + \sum_{i=1}^5 R_i.
\end{aligned}$$

After multiplying the integrand by an appropriate positive power of $\delta_n/f(x)$, the first part of Proposition 2.9 tells us that for every $\epsilon > 0$,

$$\sup_{k \in \{1,\ldots,k^*\}} \frac{k^{2/d}}{n^{2/d}} \sup_{f \in \mathcal{F}_{d,\theta}} \int_{\mathcal{X}_n^c} \frac{\Delta f(x)}{f(x)^{2/d}}\,dx = O\biggl(\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}} \biggr)$$

as $n \to \infty$. Since $\log(n-1) - \Psi(n) = O(1/n)$, it now remains to bound $R_1,\ldots,R_5$. Henceforth, to save repetition, we adopt without further mention the convention that whenever an error term inside $O(\cdot)$ or $o(\cdot)$ depends on k, this error is uniform for $k \in \{1,\ldots,k^*\}$; thus $g(n,k) = h(n,k) + o(1)$ as $n \to \infty$ means $\sup_{k \in \{1,\ldots,k^*\}} |g(n,k) - h(n,k)| \to 0$ as $n \to \infty$.

To bound R_1. By Lemma 2.10(i), we have $V_d^\alpha \mu_\alpha(f)^d \|f\|_\infty^\alpha \geq \alpha^\alpha d^d/(\alpha+d)^{\alpha+d}$. Hence

$$\begin{aligned}
|\log u_{x,s}| &\leq \log(n-1) + |\Psi(k)| - \log s + |\log \|f\|_\infty| + |\log V_d| \\
&\quad + \frac{d}{\alpha}|\log \mu_\alpha(f)| - \frac{d}{\alpha}\log(1-s) + d\log\biggl(1 + \frac{\|x\|}{\mu_\alpha^{1/\alpha}(f)}\biggr) \\
&\leq \log(n-1) + |\Psi(k)| - \log s + \max\biggl\{\log\gamma\,,\, \frac{1}{\alpha}\log\biggl(\frac{V_d^\alpha \nu^d(\alpha+d)^{\alpha+d}}{\alpha^\alpha d^d}\biggr)\biggr\} \\
&\quad + |\log V_d| + \frac{d}{\alpha}\max\biggl\{\log\nu\,,\, \frac{1}{d}\log\biggl(\frac{V_d^\alpha \gamma^\alpha(\alpha+d)^{\alpha+d}}{\alpha^\alpha d^d}\biggr)\biggr\} - \frac{d}{\alpha}\log(1-s) \\
&\quad + d\log\biggl(1 + \frac{\|x\|(\alpha+d)^{\frac{1}{\alpha}+\frac{1}{d}} V_d^{1/d}\gamma^{1/d}}{\alpha^{1/d}d^{1/\alpha}}\biggr). \qquad (2.13)
\end{aligned}$$

Moreover, for any $C_0, C_1 \geq 0, \epsilon \in (0,\alpha)$ and $\epsilon' \in (0,\epsilon)$, we have by Hölder's inequality that

$$\begin{aligned}
\sup_{f\in\mathcal{F}_{d,\theta}} \int_{\mathcal{X}_n^c} f(x)\bigl\{C_0 + \log(1+C_1\|x\|)\bigr\}\,dx &\leq \delta_n^{\frac{\alpha-\epsilon'}{\alpha+d}} \sup_{f\in\mathcal{F}_{d,\theta}} \int_{\mathcal{X}} f(x)^{\frac{d+\epsilon'}{\alpha+d}}\bigl\{C_0 + \log(1+C_1\|x\|)\bigr\}\,dx \\
\leq \delta_n^{\frac{\alpha-\epsilon'}{\alpha+d}}(1+\nu)^{\frac{d+\epsilon'}{\alpha+d}} \biggl[\int_{\mathbb{R}^d} \frac{\bigl\{C_0 + \log(1+C_1\|x\|)\bigr\}^{\frac{\alpha+d}{\alpha-\epsilon'}}}{(1+\|x\|^\alpha)^{\frac{d+\epsilon'}{\alpha-\epsilon'}}}\,dx\biggr]^{\frac{\alpha-\epsilon'}{\alpha+d}} &= o\biggl(\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}\biggr).
\end{aligned}$$

Since $|\mathbb{E}(\log \mathrm{B})| = \Psi(a+b) - \Psi(a)$ when $\mathrm{B} \sim \mathrm{Beta}(a,b)$, we deduce that for each $\epsilon > 0$,

$$R_1 = \int_{\mathcal{X}_n^c} f(x)\int_0^1 \mathrm{B}_{k,n-k}(s)\log u_{x,s}\,ds\,dx = o\biggl(\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}\biggr)$$

as $n \to \infty$, uniformly for $f \in \mathcal{F}_{d,\theta}$.

To bound R_2. For random variables $\mathrm{B}_1 \sim \mathrm{Beta}(k, n-k)$ and $B_2 \sim \mathrm{Bin}\bigl(n-1, a_n/(n-1)\bigr)$ we have that for every $\epsilon > 0$,

$$\mathbb{P}\bigl(\mathrm{B}_1 \geq a_n/(n-1)\bigr) = \mathbb{P}(B_2 \leq k-1) \leq \exp\biggl(-\frac{(a_n-k+1)^2}{2a_n}\biggr) = o(n^{-(3-\epsilon)}), \qquad (2.14)$$

where the inequality follows from standard bounds on the left-hand tail of the binomial distribution (see, e.g. Shorack and Wellner (2009), Equation (6), page 440). Now, for any $C_1 > 0$, we have $\alpha\log(1+C_1\|x\|) \leq (1+C_1\|x\|)^\alpha - 1$, so that $\sup_{f\in\mathcal{F}_{d,\theta}}\int_{\mathcal{X}} f(x)\log(1+C_1\|x\|)\,dx < \infty$. Moreover,

$$-\int_{\frac{a_n}{n-1}}^1 \log(1-s)\mathrm{B}_{k,n-k}(s)\,ds \leq \frac{n-1}{n-k-1}\int_{\frac{a_n}{n-1}}^1 \mathrm{B}_{k,n-k-1}(s)\,ds = o(n^{-(3-\epsilon)}),$$

for every $\epsilon > 0$, by a virtually identical argument to (2.14). We therefore deduce from these facts and (2.13) that for each $\epsilon > 0$,

$$R_2 = \int_{\mathcal{X}_n} f(x)\int_{\frac{a_n}{n-1}}^1 \mathrm{B}_{k,n-k}(s)\log u_{x,s}\,ds\,dx = o(n^{-(3-\epsilon)}), \qquad (2.15)$$

which again holds uniformly in $f \in \mathcal{F}_{d,\theta}$.

To bound R_3. We can write

$$\begin{aligned}R_3 = \int_{\mathcal{X}_n} f(x) \int_0^{\frac{a_n}{n-1}} \biggl[\biggl\{\log\biggl(\frac{V_d f(x) h_x^{-1}(s)^d}{s}\biggr) - \frac{V_d f(x) h_x^{-1}(s)^d - s}{s}\biggr\} \\ + \biggl\{\frac{V_d f(x) h_x^{-1}(s)^d - s}{s} + \frac{V_d^{-2/d} s^{2/d} \Delta f(x)}{2(d+2) f(x)^{1+2/d}}\biggr\}\biggr] \mathrm{B}_{k,n-k}(s)\, ds\, dx \\ =: R_{31} + R_{32},\end{aligned}$$

say. Now, note that

$$\sup_{k\in\{1,\ldots,k^*\}} \sup_{f\in\mathcal{F}_{d,\theta}} \sup_{s\in(0,a_n/(n-1)]} \sup_{x\in\mathcal{X}_n} \frac{a(f(x))^d s}{f(x)} \leq \frac{6}{\log(n-1)} \to 0.$$

It follows by Lemma 2.10(ii) that there exist a constant $C = C(d,\theta) > 0$ and $n_1 = n_1(d,\theta) \in \mathbb{N}$ such that for $n \geq n_1$, $k \in \{1,\ldots,k^*\}$, $s \leq a_n/(n-1)$ and $x \in \mathcal{X}_n$,

$$\biggl|\frac{V_d f(x) h_x^{-1}(s)^d - s}{s} + \frac{s^{2/d}\Delta f(x)}{2(d+2)V_d^{2/d} f(x)^{1+2/d}}\biggr| \leq C\biggl\{\frac{sa(f(x))^{d/2}}{f(x)}\biggr\}^{\beta/d},$$

and

$$\biggl|\frac{V_d f(x) h_x^{-1}(s)^d - s}{s}\biggr| \leq \frac{d^{1/2} V_d^{-2/d} s^{2/d} a(f(x))}{2(d+2) f(x)^{2/d}} + C\biggl\{\frac{sa(f(x))^{d/2}}{f(x)}\biggr\}^{\beta/d} \leq \frac{1}{2}.$$

Thus, for $n \geq n_1$ and $k \in \{1,\ldots,k^*\}$, using the fact that $|\log(1+z) - z| \leq z^2$ for $|z| \leq 1/2$,

$$\begin{aligned}|R_{31}| &\leq 2\int_{\mathcal{X}_n} f(x) \int_0^1 \biggl[\biggl\{\frac{dV_d^{-4/d} s^{4/d} a(f(x))^2}{4(d+2)^2 f(x)^{4/d}} + C^2\biggl\{\frac{sa(f(x))^{d/2}}{f(x)}\biggr\}^{2\beta/d}\biggr] \mathrm{B}_{k,n-k}(s)\, ds\, dx \\ &\leq \frac{dV_d^{-4/d}\Gamma(k+4/d)\Gamma(n)}{2(d+2)^2\Gamma(k)\Gamma(n+4/d)} \int_{\mathcal{X}_n} a(f(x))^2 f(x)^{1-4/d}\, dx \\ &\quad + \frac{2C^2\Gamma(k+2\beta/d)\Gamma(n)}{\Gamma(k)\Gamma(n+2\beta/d)} \int_{\mathcal{X}_n} a(f(x))^\beta f(x)^{1-2\beta/d}\, dx.\end{aligned}$$

On the other hand, we also have for $n \geq n_1$ and $k \in \{1,\ldots,k^*\}$ that

$$\begin{aligned}|R_{32}| &\leq C\int_{\mathcal{X}_n} f(x) \int_0^1 \biggl\{\frac{sa(f(x))^{d/2}}{f(x)}\biggr\}^{\beta/d} \mathrm{B}_{k,n-k}(s)\, ds\, dx \\ &\leq \frac{C\Gamma(k+\beta/d)\Gamma(n)}{\Gamma(k)\Gamma(n+\beta/d)} \int_{\mathcal{X}_n} a(f(x))^{\beta/2} f(x)^{1-\beta/d}\, dx.\end{aligned}$$

Multiplying each of the integrals by $f(x)/\delta_n$ to an appropriate positive power if necessary and by the second part of Proposition 2.9, for every $\epsilon > 0$,

$$\max(|R_{31}|, |R_{32}|) = O\biggl(\max\biggl\{\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}, \frac{k^{\frac{\beta}{d}}}{n^{\frac{\beta}{d}}}\biggr\}\biggr),$$

uniformly for $f \in \mathcal{F}_{d,\theta}$.

To bound R_4. We have

$$R_4 = \int_{\mathcal{X}_n} f(x) \int_{\frac{a_n}{n-1}}^1 \biggl\{\log\biggl(\frac{(n-1)s}{e^{\Psi(k)} f(x)}\biggr) - \frac{V_d^{-2/d} s^{2/d} \Delta f(x)}{2(d+2) f(x)^{1+2/d}}\biggr\} \mathrm{B}_{k,n-k}(s)\, ds\, dx.$$

Consider the random variable $\mathrm{B}_1 \sim \mathrm{Beta}(k, n-k)$. Then, using (2.14) and the fact that $(n-1)s/e^{\Psi(k)} \geq 1$ for $s \geq a_n/(n-1)$ and $n \geq 3$, we conclude that for every $\epsilon > 0$ and $n \geq 3$,

$$|R_4| \leq \biggl\{\log\Bigl(\frac{n-1}{e^{\Psi(k)}}\Bigr) + \int_{\mathcal{X}_n} f(x)\biggl(|\log f(x)| + \frac{a(f(x))}{f(x)^{\frac{2}{d}} V_d^{\frac{2}{d}}}\biggr)\,dx\biggr\}\mathbb{P}\Bigl(\mathrm{B}_1 \geq \frac{a_n}{n-1}\Bigr) = o(n^{-(3-\epsilon)}),$$

uniformly for $f \in \mathcal{F}_{d,\theta}$, since we have $\sup_{f\in\mathcal{F}_{d,\theta}} \int_{\mathcal{X}_n} f(x)|\log f(x)|\,dx < \infty$ by Lemma 2.11(i) in Section 2.6.2.

To bound R_5. We use the fact that for $f \in \mathcal{F}_{d,\theta}$, $x \in \mathcal{X}$ and $\epsilon' > 0$,

$$\begin{aligned}|\log f(x)| &\leq \bigl|\log \|f\|_\infty\bigr| + \log\Bigl(\frac{\|f\|_\infty}{f(x)}\Bigr)\\ &\leq \max\biggl\{\log\gamma\,,\, \log V_d + \frac{1}{\alpha}\log\biggl(\frac{\nu^d(\alpha+d)^{\alpha+d}}{\alpha^\alpha d^d}\biggr)\biggr\} + \frac{1}{\epsilon'}\Bigl(\frac{\gamma}{f(x)}\Bigr)^{\epsilon'}.\end{aligned}$$

It follows from the first part of Proposition 2.9 (having replaced $a(\delta)$ with $\max\{a(\delta), |\log\delta|\}$ if necessary) that for each $\epsilon > 0$,

$$R_5 = \int_{\mathcal{X}_n^c} f(x)\{\log(n-1) - \Psi(n) - \log f(x)\}\,dx = o\biggl(\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}\biggr)$$

uniformly in $f \in \mathcal{F}_{d,\theta}$. The claim follows when $d \geq 3$, $\alpha > 2d/(d-2)$ and $\beta \in (2,4]$.

We now consider the case where either $d \leq 2$ or $\alpha \leq 2d/(d-2)$ or $\beta \in (0,2]$, for which we need only show that

$$\sup_{f\in\mathcal{F}_{d,\theta}} |\mathbb{E}_f(\hat{H}_n) - H| = O\biggl(\max\biggl\{\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}\,,\,\frac{k^{\frac{\beta}{d}}}{n^{\frac{\beta}{d}}}\biggr\}\biggr).$$

The calculation here is very similar, but we approximate $\log u_{x,s}$ simply by $\log\bigl(\frac{(n-1)s}{e^{\Psi(k)}f(x)}\bigr)$. Writing $R_1', \ldots, R_5'$ for the modified error terms, we obtain

$$\mathbb{E}_f(\hat{H}_n) = H + \log(n-1) - \Psi(n) + \sum_{i=1}^5 R_i'.$$

Here, $R_1' = R_1 = o\bigl\{\bigl(\frac{k}{n}\bigr)^{\alpha/(\alpha+d)-\epsilon}\bigr\}$, and $R_2' = R_2 = o(n^{-(3-\epsilon)})$, for every $\epsilon > 0$ in both cases. On the other hand,

$$R_3' = \int_{\mathcal{X}_n} f(x)\int_0^{\frac{a_n}{n-1}} \log\biggl(\frac{V_d f(x) h_x^{-1}(s)^d}{s}\biggr)\mathrm{B}_{k,n-k}(s)\,ds\,dx = O\biggl(\max\biggl\{\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}\,,\,\frac{k^{\beta/d}}{n^{\beta/d}}\biggr\}\biggr)$$

for every $\epsilon > 0$, by Lemma 2.10(ii). Similarly, for every $\epsilon > 0$,

$$R_4' = \int_{\mathcal{X}_n} f(x)\int_{\frac{a_n}{n-1}}^1 \log\biggl(\frac{(n-1)s}{e^{\Psi(k)}f(x)}\biggr)\,\mathrm{B}_{k,n-k}(s)\,ds\,dx = o(n^{-(3-\epsilon)}),$$

and $R_5' = R_5 = o\bigl\{\bigl(\frac{k}{n}\bigr)^{\alpha/(\alpha+d)-\epsilon}\bigr\}$. All of these bounds hold uniformly in $f \in \mathcal{F}_{d,\theta}$, so the claim is established for this setting.

Finally, consider now the case $d \geq 3$, $\alpha > 2d/(d-2)$ and $\beta > 4$. Again the calculation is very similar to the earlier cases, with the main difference being that in bounding the error corresponding to R_3, we require a higher-order Taylor expansion of

$$\log\biggl(1+\frac{V_d f(x)h_x^{-1}(s)^d - s}{s}\biggr).$$

This can be done using Lemma 2.10(ii); we omit the details for brevity. □

Proof of Corollary 2.4. It is convenient to write $d' := \lfloor d/4 \rfloor + 1$ and $\beta' := \lceil \beta/2 \rceil - 1$. We have

$$\begin{aligned}|\mathbb{E}_f(\hat{H}_n^w) - H| &= \biggl|\sum_{j=1}^k w_j \biggl\{\mathbb{E}_f(\log \xi_{(j),1}) - H - \sum_{l=1}^{\lfloor d/4 \rfloor} \frac{\Gamma(j+2l/d)\Gamma(n)}{\Gamma(j)\Gamma(n+2l/d)}\lambda_l\biggr\}\biggr| \\ &\leq \biggl|\sum_{j=1}^k w_j \biggl\{\mathbb{E}_f(\log \xi_{(j),1}) - H - \sum_{l=1}^{\beta'} \frac{\Gamma(j+2l/d)\Gamma(n)}{\Gamma(j)\Gamma(n+2l/d)}\lambda_l\biggr\}\biggr| \\ &\qquad + \biggl|\sum_{j=1}^k w_j \sum_{l=d'}^{\beta'} \frac{\Gamma(j+2l/d)\Gamma(n)}{\Gamma(j)\Gamma(n+2l/d)}\lambda_l\biggr\}\biggr|.\end{aligned}$$

The first term can be bounded, uniformly for $f \in \mathcal{F}_{d,\theta}$ and $k \in \{1,\ldots,k^*\}$, using Lemma 2.3. For the second term, we can use monotonicity properties of ratios of gamma functions to write

$$\begin{aligned}\biggl|\sum_{j=1}^k w_j \sum_{l=d'}^{\beta'} \frac{\Gamma(j+2l/d)\Gamma(n)}{\Gamma(j)\Gamma(n+2l/d)}\lambda_l\biggr\}\biggr| &\leq \max_{d' \leq \ell \leq \beta'} |\lambda_\ell| \sum_{j=1}^k |w_j| \sum_{l=d'}^{\beta'} \frac{\Gamma(k+2l/d)\Gamma(n)}{\Gamma(k)\Gamma(n+2l/d)} \\ &\leq d^{1/2}\|w\|\bigl(\beta' - d' + 1\bigr)\frac{\Gamma(k+2d'/d)\Gamma(n)}{\Gamma(k)\Gamma(n+2d'/d)} \max_{d' \leq l \leq \beta'} |\lambda_l| = O\Bigl(\frac{k^{2d'/d}}{n^{2d'/d}}\Bigr),\end{aligned}$$

uniformly for $f \in \mathcal{F}_{d,\theta}$. The result follows. □

2.5.2 Proof of Lemma 2.7

Since this proof is long, we focus here on the main argument, and defer proofs of bounds on the many error terms to Section 2.6.5.

Proof of Lemma 2.7. We employ the same notation as in the proof of Lemma 2.3, except that we redefine δ_n so that $\delta_n := kc_n^d \log^3(n-1)/(n-1)$. We write $\mathcal{X}_n := \{x : f(x) \geq \delta_n\}$ for this newly-defined δ_n. Similar to the proof of Lemma 2.3, all error terms inside $O(\cdot)$ and $o(\cdot)$ that depend on k are uniform for $k \in \{k_0^*,\ldots,k_1^*\}$, and we now adopt the additional convention that, where relevant, these error terms are also uniform for $f \in \mathcal{F}_{d,\theta}$. By the nested properties of the classes $\mathcal{F}_{d,\theta}$ with respect to the smoothness parameter β, we may assume without loss of generality that $\beta \in (0,1]$. We first deal with the variance of the unweighted estimator $\hat{H}_n$, and note that

$$\begin{aligned}\operatorname{Var}\hat{H}_n &= n^{-1}\operatorname{Var}\log\xi_1 + (1-n^{-1})\operatorname{Cov}(\log\xi_1,\log\xi_2) \\ &= n^{-1}\operatorname{Var}\log\xi_1 + (1-n^{-1})\bigl\{\operatorname{Cov}\bigl(\log(\xi_1 f(X_1)),\log(\xi_2 f(X_2))\bigr) \\ &\qquad - 2\operatorname{Cov}\bigl(\log(\xi_1 f(X_1)),\log f(X_2)\bigr)\bigr\}. \qquad (2.16)\end{aligned}$$

We claim that for every $\epsilon > 0$,

$$\operatorname{Var}\log\xi_1 = V(f) + \frac{1}{k}\{1+o(1)\} + O\biggl\{\max\biggl(\frac{k^{\beta/d}}{n^{\beta/d}}\log n\,,\,\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}\biggr)\biggr\} \qquad (2.17)$$

as $n \to \infty$. The proof of this claim uses similar methods to those in the proof of Lemma 2.3. In particular, writing $S_1, \ldots, S_5$ for remainder terms to be bounded later, we have

$$\begin{aligned}
\mathbb{E}(\log^2 \xi_1) &= \int_{\mathcal{X}} f(x) \int_0^\infty \log^2 u \, dF_{n,x}(u)\, dx \\
&= \int_{\mathcal{X}_n} f(x) \int_0^1 \mathrm{B}_{k,n-k}(s) \log^2 u_{x,s}\, ds\, dx + S_1 \\
&= \int_{\mathcal{X}_n} f(x) \int_0^{\frac{a_n}{n-1}} \mathrm{B}_{k,n-k}(s) \log^2 u_{x,s}\, ds\, dx + S_1 + S_2 \\
&= \int_{\mathcal{X}_n} f(x) \int_0^{\frac{a_n}{n-1}} \log^2\biggl(\frac{(n-1)s}{e^{\Psi(k)} f(x)}\biggr) \mathrm{B}_{k,n-k}(s)\, ds\, dx + S_1 + S_2 + S_3 \\
&= \int_{\mathcal{X}_n} f(x) \bigl[\log^2 f(x) - 2\{\log(n-1) - \Psi(n)\} \log f(x) \\
&\qquad\qquad + \Psi'(k) - \Psi'(n) + \{\log(n-1) - \Psi(n)\}^2\bigr]\, dx + \sum_{i=1}^4 S_i \\
&= \int_{\mathcal{X}} f(x) \log^2 f(x)\, dx + \sum_{i=1}^5 S_i + \frac{1}{k}\{1 + o(1)\},
\end{aligned} \tag{2.18}$$

as $n \to \infty$. In Section 2.6.5, we show that for every $\epsilon > 0$,

$$\sum_{i=1}^5 |S_i| = O\biggl\{\max\biggl(\frac{k^{\beta/d}}{n^{\beta/d}} \log n\,,\ \frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}\biggr)\biggr\} \tag{2.19}$$

as $n \to \infty$. Combining (2.18) with (2.19) and Lemma 2.3, we deduce that (2.17) holds.

The next step of our proof consists of showing that for every $\epsilon > 0$,

$$\mathrm{Cov}\bigl(\log(\xi_1 f(X_1)), \log f(X_2)\bigr) = O\biggl(\max\biggl\{\frac{k^{-\frac{1}{2}+\frac{2\alpha-\epsilon}{\alpha+d}}}{n^{\frac{2\alpha-\epsilon}{\alpha+d}}}\,,\ \frac{k^{\frac{1}{2}+\frac{\beta}{d}}}{n^{1+\frac{\beta}{d}}} \log^{2+\beta/d} n\biggr\}\biggr) \tag{2.20}$$

as $n \to \infty$. Define

$$\begin{aligned}
F^-_{n,x}(u) &:= \sum_{j=k}^{n-2} \binom{n-2}{j} p^j_{n,x,u}(1 - p_{n,x,u})^{n-2-j}, \\
\tilde{F}_{n,x}(u) &:= \sum_{j=k-1}^{n-2} \binom{n-2}{j} p^j_{n,x,u}(1 - p_{n,x,u})^{n-2-j},
\end{aligned}$$

so that

$$\mathbb{P}(\xi_1 \leq u | X_1 = x, X_2 = y) = \begin{cases} F^-_{n,x}(u) & \text{if } \|x - y\| > r_{n,u} \\ \tilde{F}_{n,x}(u) & \text{if } \|x - y\| \leq r_{n,u}. \end{cases}$$

Writing $\tilde{u}_{n,x,y} := V_d(n-1)\|x - y\|^d e^{-\Psi(k)}$, we therefore have that

$$\begin{aligned}
&\mathrm{Cov}\bigl(\log(\xi_1 f(X_1)), \log f(X_2)\bigr) \\
&\quad= \int_{\mathcal{X}\times\mathcal{X}} f(x) f(y) \log f(y) \int_{\tilde{u}_{n,x,y}}^\infty \log\bigl(uf(x)\bigr)\, d(\tilde{F}_{n,x} - F^-_{n,x})(u)\, dx\, dy \\
&\qquad\qquad - H(f) \int_{\mathcal{X}} f(x) \int_0^\infty \log\bigl(uf(x)\bigr)\, d(F^-_{n,x} - F_{n,x})(u)\, dx.
\end{aligned} \tag{2.21}$$

To deal with the first term in (2.21), we make the substitution

$$y = y_{x,z} := x + \frac{r_{n,1}}{f(x)^{1/d}} z, \tag{2.22}$$

and let $d_n := (24 \log n)^{1/d}$. Writing T_1, T_2, T_3 for remainder terms to be bounded later, for every $\epsilon > 0$ and for $k \geq 2$,

$$\begin{aligned}
&\int_{\mathcal{X}\times\mathcal{X}} f(x) f(y) \log f(y) \int_{\tilde{u}_{n,x,y}}^{\infty} \log(uf(x))\, d(\tilde{F}_{n,x} - F_{n,x}^{-})(u)\, dy\, dx \\
&= r_{n,1}^d \int_{\mathcal{X}_n} \int_{B_0(d_n)} f(y_{x,z}) \log f(y_{x,z}) \int_{\frac{\|z\|^d}{f(x)}}^{\infty} \log(uf(x))\, d(\tilde{F}_{n,x} - F_{n,x}^{-})(u) dz dx + T_1 \\
&= r_{n,1}^d \int_{\mathcal{X}_n} f(x) \log f(x) \int_{B_0(d_n)} \int_{\frac{\|z\|^d}{f(x)}}^{\infty} \log(uf(x))\, d(\tilde{F}_{n,x} - F_{n,x}^{-})(u) dz dx + T_1 + T_2 \\
&= \frac{k-1}{n-k-1} \int_{\mathcal{X}_n} f(x) \log f(x) dx \int_0^{\frac{a_n}{n-1}} \log\Bigl(\frac{(n-1)s}{e^{\Psi(k)}}\Bigr) \mathrm{B}_{k,n-k-1}(s) \Bigl(1 - \frac{(n-2)s}{k-1}\Bigr) ds + \sum_{i=1}^{3} T_i \\
&= \frac{H(f)}{n} + O(n^{-2}) + o\biggl(\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{1+\frac{\alpha}{\alpha+d}-\epsilon}}\biggr) + \sum_{i=1}^{3} T_i.
\end{aligned} \tag{2.23}$$

In Section 2.6.5, we show that for every $\epsilon > 0$,

$$\sum_{i=1}^{3} |T_i| = O\biggl(\max\biggl\{\frac{k^{-\frac{1}{2}+\frac{2\alpha}{\alpha+d}-\epsilon}}{n^{\frac{2\alpha}{\alpha+d}-\epsilon}}, \frac{k^{\frac{1}{2}+\frac{\beta}{d}}}{n^{1+\frac{\beta}{d}}} \log^{2+\beta/d} n\biggr\}\biggr) \tag{2.24}$$

as $n \to \infty$. We now deal with the second term in (2.21). Writing U_1, U_2 for remainder terms to be bounded later, for every $\epsilon > 0$,

$$\begin{aligned}
&\int_{\mathcal{X}} f(x) \int_0^{\infty} \log\bigl(uf(x)\bigr)\, d(F_{n,x}^{-} - F_{n,x})(u)\, dx \\
&\quad = \int_{\mathcal{X}_n} f(x) \int_0^{\frac{a_n}{n-1}} \log(u_{x,s} f(x)) \mathrm{B}_{k,n-k-1}(s) \Bigl\{\frac{(n-1)s-k}{n-k-1}\Bigr\}\, ds\, dx + U_1 \\
&\quad = \int_{\mathcal{X}_n} f(x) \int_0^{1} \log\Bigl(\frac{(n-1)s}{e^{\Psi(k)}}\Bigr) \mathrm{B}_{k,n-k-1}(s) \Bigl\{\frac{(n-1)s-k}{n-k-1}\Bigr\}\, ds\, dx + U_1 + U_2 \\
&\quad = \frac{1}{n-1} + U_1 + U_2 + o\biggl(\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{1+\frac{\alpha}{\alpha+d}-\epsilon}}\biggr).
\end{aligned} \tag{2.25}$$

In Section 2.6.5, we show that for every $\epsilon > 0$,

$$|U_1| + |U_2| = O\biggl(\frac{k^{1/2}}{n} \max\biggl\{\frac{k^{\beta/d}}{n^{\beta/d}}, \frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}\biggr\}\biggr). \tag{2.26}$$

From (2.21), (2.23), (2.24), (2.25) and (2.26), we conclude that (2.20) holds.

By (2.16), it remains to consider $\mathrm{Cov}\bigl(\log(\xi_1 f(X_1)), \log(\xi_2 f(X_2))\bigr)$. We require some further notation. Let $F_{n,x,y}$ denote the conditional distribution function of (ξ_1, ξ_2) given $X_1 = x, X_2 = y$. Let $a_n^- := (k - 3k^{1/2} \log^{1/2} n) \vee 0$, $a_n^+ := (k + 3k^{1/2} \log^{1/2} n) \wedge (n-1)$, and let

$$v_x := \inf\{u \geq 0 : (n-1)p_{n,x,u} = a_n^+\}, \quad l_x := \inf\{u \geq 0 : (n-1)p_{n,x,u} = a_n^-\},$$

so that $\mathbb{P}\{\xi_1 \leq l_{X_1}\} = o(n^{-(9/2-\epsilon)})$ and $\mathbb{P}\{\xi_1 \geq v_{X_1}\} = o(n^{-(9/2-\epsilon)})$ for every $\epsilon > 0$. For pairs (u, v) with $u \leq v_x$ and $v \leq v_y$, let $(M_1, M_2, M_3) \sim \mathrm{Multi}(n-2; p_{n,x,u}, p_{n,y,v}, 1 - p_{n,x,u} - p_{n,y,v})$, and write

$$G_{n,x,y}(u, v) := \mathbb{P}(M_1 \geq k, M_2 \geq k),$$

so that $F_{n,x,y}(u, v) = G_{n,x,y}(u, v)$ for $\|x - y\| > r_{n,u} + r_{n,v}$. Write

$$\Sigma := \begin{pmatrix} 1 & \alpha_z \\ \alpha_z & 1 \end{pmatrix}$$

with $\alpha_z := V_d^{-1}\mu_d\bigl(B_0(1) \cap B_z(1)\bigr)$ for $z \in \mathbb{R}^d$, let $\Phi_\Sigma(s, t)$ denote the distribution function of a $N_2(0, \Sigma)$ random vector at (s, t), and let Φ denote the standard univariate normal distribution function. Writing W_i for remainder terms to be bounded later, and writing $h(u, v) := \log(uf(x)) \log(vf(y))$ as shorthand, we have

$$\begin{aligned}
&\mathrm{Cov}(\log(\xi_1 f(X_1)), \log(\xi_2 f(X_2))) \\
&= \int_{\mathcal{X}\times\mathcal{X}} f(x)f(y) \int_0^\infty \int_0^\infty h(u, v)\, d(F_{n,x,y} - F_{n,x}F_{n,y})(u, v)\, dx\, dy \\
&= \int_{\mathcal{X}\times\mathcal{X}} f(x)f(y) \int_{[l_x,v_x]\times[l_y,v_y]} h(u, v)\, d(F_{n,x,y} - F_{n,x}F_{n,y})(u, v)\, dx\, dy + W_1 \\
&= \int_{\mathcal{X}\times\mathcal{X}} f(x)f(y) \int_{[l_x,v_x]\times[l_y,v_y]} h(u, v)\, d(F_{n,x,y} - G_{n,x,y})(u, v) dx dy - \frac{1}{n} + \sum_{i=1}^{2} W_i \\
&= \int_{\mathcal{X}_n\times\mathcal{X}} f(x)f(y) \int_{l_x}^{v_x}\int_{l_y}^{v_y} \frac{(F_{n,x,y} - G_{n,x,y})(u, v)}{uv}\, du\, dv\, dx\, dy - \frac{1}{n} + \sum_{i=1}^{3} W_i \\
&= \frac{r_{n,1}^d}{k} \int_{B_0(2)} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{\Phi_\Sigma(s, t) - \Phi(s)\Phi(t)\}\, ds\, dt\, dz - \frac{1}{n} + \sum_{i=1}^{4} W_i \\
&= \frac{e^{\Psi(k)}}{k(n-1)} - \frac{1}{n} + \sum_{i=1}^{4} W_i = O\Bigl(\frac{1}{nk}\Bigr) + \sum_{i=1}^{4} W_i. \qquad (2.27)
\end{aligned}$$

The proof in the unweighted case is completed by showing in Section 2.6.5 that for every $\epsilon > 0$,

$$\sum_{i=1}^{4} |W_i| = O\biggl(\max\biggl\{\frac{\log^{\frac{5}{2}} n}{nk^{\frac{1}{2}}}, \frac{k^{\frac{3}{2}+\frac{\alpha-\epsilon}{\alpha+d}}}{n^{1+\frac{\alpha-\epsilon}{\alpha+d}}}, \frac{k^{\frac{3}{2}+\frac{2\beta}{d}}}{n^{1+\frac{2\beta}{d}}}, \frac{k^{(1+\frac{d}{2\beta})\frac{\alpha-\epsilon}{\alpha+d}}}{n^{1+\frac{\alpha-\epsilon}{\alpha+d}}}, \frac{k^{\frac{1}{2}+\frac{\beta}{d}} \log n}{n^{1+\frac{\beta}{d}}}, \frac{k^{\frac{2\alpha-\epsilon}{\alpha+d}}}{n^{\frac{2\alpha-\epsilon}{\alpha+d}}}\biggr\}\biggr)$$

as $n \to \infty$.

The proof in the weighted case uses similar arguments; details are deferred to Section 2.6.5. □

2.5.3 Proofs of Theorems 2.1 and 2.2

Proof of Theorem 2.1. Writing $j_t := \lfloor tk/d \rfloor$ for $t = 1, \ldots, d$ and $d' := \lfloor d/4 \rfloor + 1$ for convenience, a sufficient condition for $\mathcal{W}^{(k)} \neq \emptyset$ is that the matrix $A^{(k)} \in \mathbb{R}^{d'\times d'}$ with $(l, t)^{th}$ entry

$$A_{lt}^{(k)} = \Gamma(j_t)^{-1}\Gamma(j_t + 2(l-1)/d)k^{-2(l-1)/d},$$

is invertible. This follows because, writing $e_1 := (1, 0, \ldots, 0)^T \in \mathbb{R}^{d'}$ we can then define $w = w^{(k)} \in \mathcal{W}^{(k)}$ by setting

$$(w_{j_t})_{t=1}^{\lfloor d/4 \rfloor + 1} := (A^{(k)})^{-1} e_1$$

and setting all other entries of w to be zero. Now define $A \in \mathbb{R}^{d' \times d'}$ to have $(l, t)^{th}$ entry $A_{lt} := (t/d)^{2(l-1)/d}$. Since $x^{-a}\Gamma(x)^{-1}\Gamma(x + a) \to 1$ as $x \to \infty$ for $a \in \mathbb{R}$, we have $\|A^{(k)} - A\| \to 0$ as $k \to \infty$. Now, A is a Vandermonde matrix (depending only on d) and as such has determinant

$$|A| = \prod_{1 \leq t_1 < t_2 \leq d'} d^{-2/d} (t_2^{2/d} - t_1^{2/d}) > 0.$$

Hence, by the continuity of the determinant and eigenvalues of a matrix, we have that there exists $k_d > 0$ such that, for $k \geq k_d$, the matrix $A^{(k)}$ is invertible and

$$\|(A^{(k)})^{-1} e_1\| \leq |\lambda_{\min}(A^{(k)})|^{-1} \leq 2|\lambda_{\min}(A)|^{-1},$$

where $\lambda_{\min}(\cdot)$ denotes the eigenvalue of a matrix with the smallest absolute value. It follows that, for each $k \geq k_d$, there exists $w^{(k)} \in \mathcal{W}^{(k)}$ satisfying $\sup_{k \geq k_d} \|w^{(k)}\| < \infty$, as required.

Now, by Corollary 2.4 and the fact that $w \in \mathcal{W}^{(k)}$, we have for $\epsilon > 0$ sufficiently small,

$$\mathbb{E}_f(\hat{H}_n^w) - H(f) = O\biggl(\max\biggl\{\frac{k^{\frac{\alpha}{\alpha+d} - \epsilon}}{n^{\frac{\alpha}{\alpha+d} - \epsilon}}, \frac{k^{\frac{2d'}{d}}}{n^{\frac{2d'}{d}}}, \frac{k^{\frac{\beta}{d}}}{n^{\frac{\beta}{d}}}\biggr\}\biggr) = o(n^{-1/2}),$$

uniformly for $f \in \mathcal{F}_{d,\theta}$, under our conditions on k_1^*, α and β. By Lemma 2.7 we have $\operatorname{Var} \hat{H}_n^w = n^{-1}V(f) + o(n^{-1})$ uniformly for $f \in \mathcal{F}_{d,\theta}$. Note that by Cauchy–Schwarz, very similar arguments to those used at (2.18) and Lemma 2.11 in Section 2.6.2 we have that, for $j \in \operatorname{supp}(w)$,

$$\Bigl|\operatorname{Cov}_f\Bigl(\log\bigl(\xi_{(j),1} f(X_1)\bigr), \log f(X_1)\Bigr)\Bigr| \leq \bigl\{V(f)\mathbb{E}_f\bigl[\log^2\bigl(\xi_{(j),1} f(X_1)\bigr)\bigr]\bigr\}^{1/2} \to 0$$

uniformly for $f \in \mathcal{F}_{d,\theta}$. Therefore, also using (2.20), we have that

$$\begin{aligned}
\operatorname{Var}_f(\hat{H}_n^w - H_n^*) &= \operatorname{Var}_f \hat{H}_n^w + 2\operatorname{Cov}_f(\hat{H}_n^w, \log f(X_1)) + n^{-1}V(f) \\
&= \operatorname{Var}_f \hat{H}_n^w - n^{-1}V(f) + \frac{2}{n}\sum_{j=1}^k w_j \operatorname{Cov}_f\Bigl(\log\bigl(\xi_{(j),1} f(X_1)\bigr), \log f(X_1)\Bigr) \\
&\qquad + 2(1 - n^{-1})\sum_{j=1}^k w_j \operatorname{Cov}\Bigl(\log\bigl(\xi_{(j),2} f(X_2)\bigr), \log f(X_1)\Bigr) = o(n^{-1})
\end{aligned}$$

as $n \to \infty$, uniformly for $f \in \mathcal{F}_{d,\theta}$. The conclusion (2.3) follows on writing

$$\mathbb{E}_f\bigl\{(\hat{H}_n^w - H_n^*)^2\bigr\} = \operatorname{Var}_f(\hat{H}_n^w - H_n^*) + (\mathbb{E}_f \hat{H}_n^w - H(f))^2,$$

and the final conclusion is then immediate. □

Proof of Theorem 2.2. We have

$$\begin{aligned} d_{\mathrm{BL}}&\Bigl(\mathcal{L}\bigl(n^{1/2}\{\hat{H}_n^w - H(f)\}\bigr), \mathcal{L}\bigl(n^{1/2}\{H_n^* - H(f)\}\bigr)\Bigr) \\ &\leq \sup_{h \in \mathcal{H}} \mathbb{E}_f \bigl|h\bigl(n^{1/2}\{\hat{H}_n^w - H(f)\}\bigr) - h\bigl(n^{1/2}\{H_n^* - H(f)\}\bigr)\bigr| \\ &\leq n^{1/2}\mathbb{E}_f|\hat{H}_n^w - H_n^*| \leq n^{1/2}\bigl[\mathbb{E}_f\bigl\{(\hat{H}_n^w - H_n^*)^2\bigr\}\bigr]^{1/2}. \qquad (2.28) \end{aligned}$$

Now write $\mathcal{H}^*$ for the class of functions $h : \mathbb{R} \to \mathbb{R}$ having Lipschitz constant at most 1, and let $Z \sim N\bigl(0, V(f)\bigr)$. Then by standard properties of the Wasserstein distance (e.g. Gibbs and Su, 2002, p. 424) and the non-uniform version of the Berry–Esseen theorem (e.g. Paditz, 1989, Theorem 1),

$$\begin{aligned} d_{\mathrm{BL}}&\Bigl(\mathcal{L}\bigl(n^{1/2}\{H_n^* - H(f)\}\bigr), N\bigl(0, V(f)\bigr)\Bigr) \\ &\leq \sup_{h \in \mathcal{H}^*} \bigl|\mathbb{E}_f h\bigl(n^{1/2}\{H_n^* - H(f)\}\bigr) - \mathbb{E}h(Z)\bigr| \\ &= \int_{-\infty}^{\infty} \Bigl|\mathbb{P}_f\bigl(n^{1/2}\{H_n^* - H(f)\} \leq x\bigr) - \mathbb{P}(Z \leq x)\Bigr|\,dx \leq \frac{78\beta_3(f)}{n^{1/2}V(f)}, \qquad (2.29) \end{aligned}$$

where

$$\beta_3(f) := \mathbb{E}_f\bigl\{\bigl|\log f(X_1) + H(f)\bigr|^3\bigr\} = \int_{\mathcal{X}} f(x)|\log f(x) + H(f)|^3\,dx.$$

We conclude from (2.28) and (2.29), together with Theorem 2.1 and Lemma 2.11 in Section 2.6.2, that

$$\sup_{k \in \{k_0^*, \ldots, k_1^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} d_{\mathrm{BL}}\Bigl(\mathcal{L}\bigl(n^{1/2}(\hat{H}_n^w - H(f))\bigr), N\bigl(0, V(f)\bigr)\Bigr) \to 0$$

as $n \to \infty$, as required.

For the second part of the theorem, set

$$\epsilon_n = \epsilon_n^w(d,\theta) := \frac{\sup_{k \in \{1,\ldots,k^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} \Bigl(2\mathbb{E}_f\bigl[\{\tilde{V}_n^w - V(f)\}^2\bigr]\Bigr)^{1/3}}{\inf_{f \in \mathcal{F}_{d,\theta}} V(f)^{2/3}},$$

so that $\epsilon_n \to 0$, by Lemmas 2.11(ii) and 2.13 in Section 2.6.2. Then, by two applications of Markov's inequality, for n large enough that $\epsilon_n \leq 1$,

$$\begin{aligned} \mathbb{P}_f\biggl(\biggl|\frac{(\hat{V}_n^w)^{1/2}}{V^{1/2}(f)} - 1\biggr| \geq \epsilon_n\biggr) &\leq \mathbb{P}_f\biggl(\biggl|\frac{\tilde{V}_n^w}{V(f)} - 1\biggr| \geq \epsilon_n\biggr) + \mathbb{P}_f(\tilde{V}_n^w \leq 0) \\ &\leq \frac{\mathbb{E}_f\bigl[\{\tilde{V}_n^w - V(f)\}^2\bigr]}{V(f)^2}\biggl(\frac{1}{\epsilon_n^2} + 1\biggr) \leq \epsilon_n. \end{aligned}$$

For $n \in \mathbb{N}$ and $L \geq 1$, define $h_{n,L} : \mathbb{R} \to [0,1]$ by

$$h_{n,L}(x) := \begin{cases} 0 & \text{if } |x| > z_{q/2}(1+\epsilon_n) + 1/L \\ L\{z_{q/2}(1+\epsilon_n) + 1/L - |x|\} & \text{if } 0 < |x| - z_{q/2}(1+\epsilon_n) \leq 1/L \\ 1 & \text{if } |x| \leq z_{q/2}(1+\epsilon_n). \end{cases}$$

Thus $h_{n,L}$ has Lipschitz constant L and $h_{n,L}(x) \geq \mathbb{1}_{\{|x| \leq z_{q/2}(1+\epsilon_n)\}}$. Then, with $Z \sim N(0,1)$,

$$\begin{aligned}
&\mathbb{P}_f\bigl(I_{n,q} \ni H(f)\bigr) \\
&\quad \leq \mathbb{P}_f\biggl(\frac{n^{1/2}|\hat{H}_n^w - H(f)|}{V^{1/2}(f)} \leq z_{q/2}(1+\epsilon_n)\biggr) + \mathbb{P}_f\biggl(\frac{V^{1/2}(f)}{(\hat{V}_n^w)^{1/2}} \leq \frac{1}{1+\epsilon_n}\biggr) \\
&\quad \leq \mathbb{E}_f h_{n,L}\biggl(\frac{n^{1/2}\{\hat{H}_n^w - H(f)\}}{V^{1/2}(f)}\biggr) + \epsilon_n \\
&\quad \leq \mathbb{E}_f h_{n,L}(Z) + \epsilon_n + L d_{\mathrm{BL}}\biggl(\mathcal{L}\biggl(\frac{n^{1/2}\{\hat{H}_n^w - H(f)\}}{V^{1/2}(f)}\biggr), \mathcal{L}(Z)\biggr) \\
&\quad \leq \mathbb{P}\bigl(|Z| \leq z_{q/2}(1+\epsilon_n) + L^{-1}\bigr) + \epsilon_n \\
&\qquad + L\max\bigl((1, V^{-1/2}(f)\bigr)d_{\mathrm{BL}}\Bigl(\mathcal{L}\bigl(n^{1/2}(\hat{H}_n^w - H(f))\bigr), N\bigl(0, V(f)\bigr)\Bigr).
\end{aligned}$$

Since $L \geq 1$ was arbitrary, we deduce from the first part of the theorem and Lemma 2.11 in Section 2.6.2 that

$$\limsup_{n\to\infty} \sup_{q\in(0,1)} \sup_{k\in\{k_0^*,\ldots,k_1^*\}} \sup_{f\in\mathcal{F}_{d,\theta}} \mathbb{P}_f\bigl(I_{n,q} \ni H(f)\bigr) - (1-q) \leq \inf_{L\geq 1} \frac{2}{L(2\pi)^{1/2}} = 0.$$

The lower bound is obtained by a similar argument, omitted for brevity. □

2.6 Appendix

2.6.1 Proofs of auxiliary results

Proof of Proposition 2.9. Fix $\tau \in \bigl(\frac{d}{\alpha+d}, 1\bigr]$. We first claim that given any $\epsilon > 0$, there exists $A_\epsilon > 0$ such that $a(\delta) \leq A_\epsilon \delta^{-\epsilon}$ for all $\delta \in (0, \gamma]$. To see this, observe that there exists $\delta_0 \in (0,\gamma]$ such that $a(\delta) \leq \delta^{-\epsilon}$ for $\delta \leq \delta_0$. But then

$$\sup_{\delta\in(0,\gamma]} \delta^\epsilon a(\delta) \leq \max\bigl\{1, \gamma^\epsilon a(\delta_0)\bigr\} \leq \gamma^\epsilon \delta_0^{-\epsilon},$$

which establishes the claim, with $A_\epsilon := \gamma^\epsilon \delta_0^{-\epsilon}$. Now choose $\epsilon = \frac{1}{3}\bigl(\tau - \frac{d}{\alpha+d}\bigr)$ and let $\tau' := \frac{\tau}{3} + \frac{2d}{3(\alpha+d)} \in \bigl(\frac{d}{\alpha+d}, 1\bigr)$. Then, by Hölder's inequality, and since $\alpha\tau'/(1-\tau') > d$,

$$\begin{aligned}
\sup_{f\in\mathcal{F}_{d,\theta}} \int_{\{x:f(x)<\delta\}} a\bigl(f(x)\bigr)f(x)^\tau \, dx &\leq A_\epsilon \delta^\epsilon \sup_{f\in\mathcal{F}_{d,\theta}} \int_{\{x:f(x)<\delta\}} f(x)^{\tau'} \, dx \\
&\leq A_\epsilon \delta^\epsilon (1+\nu)^{\tau'} \biggl\{\int_{\mathbb{R}^d} (1+\|x\|^\alpha)^{-\frac{\tau'}{1-\tau'}} \, dx\biggr\}^{1-\tau'} \to 0
\end{aligned}$$

as $\delta \searrow 0$, as required.

For the second part, fix $\rho > 0$, set $\epsilon := \frac{1}{2}\bigl(\tau - \frac{d}{\alpha+d}\bigr)$ and $\tau' := \frac{\tau}{2} + \frac{d}{2(\alpha+d)} \in \bigl(\frac{d}{\alpha+d}, 1\bigr)$. Then, by Hölder's inequality again,

$$\begin{aligned}
\sup_{f\in\mathcal{F}_{d,\theta}} \int_{\mathcal{X}} a\bigl(f(x)\bigr)^\rho f(x)^\tau \, dx &\leq A_{\epsilon/\rho} \sup_{f\in\mathcal{F}_{d,\theta}} \int_{\mathcal{X}} f(x)^{\tau'} \, dx \\
&\leq A_{\epsilon/\rho}(1+\nu)^{\tau'} \biggl\{\int_{\mathbb{R}^d} (1+\|x\|^\alpha)^{-\frac{\tau'}{1-\tau'}} \, dx\biggr\}^{1-\tau'} < \infty,
\end{aligned}$$

as required. □

Proof of Lemma 2.10. (i) The lower bound is immediate from the fact that $h_x(r) \leq V_d \|f\|_\infty r^d$ for any $r > 0$. For the upper bound, observe that by Markov's inequality, for any $r > 0$,

$$h_x(\|x\| + r) = \int_{B_x(\|x\|+r)} f(y)\,dy \geq \int_{B_0(r)} f(y)\,dy \geq 1 - \frac{\mu_\alpha(f)}{r^\alpha}.$$

The result follows on substituting $r = \bigl(\frac{\mu_\alpha(f)}{1-s}\bigr)^{1/\alpha}$ for $s \in (0,1)$.

(ii) We first prove this result in the case $\beta \in (2,4]$, giving the stated form of $b_1(\cdot)$. Let $C := 4dV_d^{-\beta/d}/(d+\beta)$, and let $y := Ca(f(x))^{\beta/2} s\{s/f(x)\}^{\beta/d}$. Now, by the mean value theorem, we have for $r \leq r_a(x)$ that

$$\biggl| h_x(r) - V_d r^d f(x) - \frac{V_d}{2(d+2)} r^{d+2} \Delta f(x) \biggr| \leq a(f(x)) f(x) \frac{dV_d}{2(d+\beta)} r^{d+\beta}.$$

It is convenient to write

$$s_{x,y} := s - \frac{s^{1+2/d} \Delta f(x)}{2(d+2) V_d^{2/d} f(x)^{1+2/d}} + y.$$

Then, provided $s_{x,y} \in (0, V_d r_a^d(x) f(x)]$, we have

$$\begin{aligned} h_x\biggl(\frac{s_{x,y}^{1/d}}{\{V_d f(x)\}^{1/d}}\biggr) \\ &\geq s_{x,y} + \frac{V_d^{-2/d} \Delta f(x)}{2(d+2) f(x)^{1+2/d}} s_{x,y}^{1+2/d} - \frac{a(f(x)) d V_d^{-\beta/d}}{2(d+\beta) f(x)^{\beta/d}} s_{x,y}^{1+\beta/d}. \end{aligned}$$

Now, by our hypothesis, we know that

$$\sup_{f \in \mathcal{F}_{d,\theta}} \sup_{s \in \mathcal{S}_n} \sup_{x \in \mathcal{X}_n} \max\biggl\{ \frac{V_d^{-2/d} s^{2/d} |\Delta f(x)|}{2(d+2) f(x)^{1+2/d}}, \frac{y}{s} \biggr\} \leq \max\biggl\{ \frac{d^{1/2} V_d^{-2/d} C_n^{2/d}}{2(d+2)}, C C_n^{\beta/d} \biggr\} \to 0$$

as $n \to \infty$. Hence there exists $n_1 = n_1(d,\theta) \in \mathbb{N}$ such that for all $n \geq n_1$, all $f \in \mathcal{F}_{d,\theta}$, $s \in \mathcal{S}_n$ and $x \in \mathcal{X}_n$, we have

$$\frac{1}{2(d+2)} (s_{x,y}^{1+2/d} - s^{1+2/d}) \geq -\frac{s^{1+2/d}}{2d} \biggl\{ \frac{d^{1/2} V_d^{-2/d} a(f(x)) s^{2/d}}{2(d+2) f(x)^{2/d}} + \frac{y}{s} \biggr\}.$$

Moreover, there exists $n_2 = n_2(d,\theta) \in \mathbb{N}$ such that for all $n \geq n_2$, all $s \in \mathcal{S}_n$, $x \in \mathcal{X}_n$ and $f \in \mathcal{F}_{d,\theta}$ we have

$$|s_{x,y}|^{1+\beta/d} \leq 2s^{1+\beta/d}.$$

Finally, we can choose $n_3 = n_3(d,\theta) \in \mathbb{N}$ such that

$$\max\biggl\{ \frac{C_n^{(4-\beta)/d}}{4(d+2) V_d^{(4-\beta)/d}}, \frac{2d^{1/2} C_n^{2/d}}{(d+\beta) V_d^{2/d}}, \frac{d^{3/2} C_n^{2/d}}{2(d+2)(d+\beta) V_d^{2/d}} \biggr\} \leq \frac{d}{d+\beta}$$

and such that $C_n \leq (8d^{1/2})^{-d} V_d/2$ for $n \geq n_3$. It follows that for $n \geq \max(n_1, n_2, n_3) =: n_*$, for

$f \in \mathcal{F}_{d,\theta}$, $s \in \mathcal{S}_n$ and for $x \in \mathcal{X}_n$, we have that $s_{x,y} \in (0, V_d r_a^d(x) f(x)]$ and

$$\begin{aligned}
&h_x\biggl(\frac{s_{x,y}^{1/d}}{\{V_d f(x)\}^{1/d}}\biggr) - s \\
&\geq y - \frac{a(f(x))s^{1+2/d}}{2d^{1/2}V_d^{2/d}f(x)^{2/d}}\biggl\{\frac{d^{1/2}V_d^{-2/d}a(f(x))s^{2/d}}{2(d+2)f(x)^{2/d}} + \frac{y}{s}\biggr\} - \frac{da(f(x))s^{1+\beta/d}}{(d+\beta)V_d^{\frac{\beta}{d}}f(x)^{\frac{\beta}{d}}} \\
&\geq \frac{a(f(x))^{\beta/2}s^{1+\beta/d}}{f(x)^{\beta/d}}\biggl[C - \frac{a(f(x))^{2-\beta/2}}{4(d+2)V_d^{4/d}}\biggl\{\frac{s}{f(x)}\biggr\}^{(4-\beta)/d} - \frac{Ca(f(x))}{2d^{1/2}V_d^{2/d}}\biggl\{\frac{s}{f(x)}\biggr\}^{2/d} - \frac{dV_d^{-\beta/d}}{d+\beta}\biggr] \geq 0.
\end{aligned}$$

The lower bound is proved by very similar calculations, and the result for the case $\beta \in (2,4]$ follows. The general case can be proved using very similar arguments, and is omitted for brevity. □

2.6.2 Auxiliary results for the proof of Theorem 2.2

Recall the definition of $V(f)$ given in the statement of Theorem 2.1.

Lemma 2.11. *For each $d \in \mathbb{N}$ and $\theta \in \Theta$ and $m \in \mathbb{N}$, we have*

(i) $\sup_{f \in \mathcal{F}_{d,\theta}} \int_{\mathcal{X}} f(x)|\log^m f(x)|\,dx < \infty$*;*

(ii) $\inf_{f \in \mathcal{F}_{d,\theta}} V(f) > 0$*;*

Proof of Lemma 2.11. Fix $d \in \mathbb{N}$ and $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$.

(i) For $\epsilon \in (0,1)$ and $t \in (0,1]$, we have

$$\log\frac{1}{t} \leq \frac{1}{\epsilon}t^{-\epsilon}.$$

Let $\epsilon = \frac{\alpha}{m(\alpha+2d)}$, so that $\frac{\alpha(1-m\epsilon)}{m\epsilon} = 2d$. Then, by Hölder's inequality, for any $f \in \mathcal{F}_{d,\theta}$,

$$\begin{aligned}
\int_{\mathcal{X}} f(x)|\log^m f(x)|\,dx &\leq 2^{m-1}\int_{\mathcal{X}} f(x)\log^m\Bigl(\frac{\|f\|_\infty}{f(x)}\Bigr)\,dx + 2^{m-1}|\log^m \|f\|_\infty| \\
&\leq \frac{2^{m-1}\|f\|_\infty^{m\epsilon}}{\epsilon^m}\int_{\mathcal{X}} f(x)^{1-m\epsilon}\,dx + 2^{m-1}|\log^m \|f\|_\infty| \\
&\leq \frac{2^{m-1}\gamma^{m\epsilon}}{\epsilon^m}(1+\nu)^{1-m\epsilon}\Bigl\{\int_{\mathcal{X}}(1+\|x\|^\alpha)^{-\frac{1-m\epsilon}{m\epsilon}}\,dx\Bigr\}^{m\epsilon} \\
&\quad + 2^{m-1}\max\biggl\{\log^m\gamma\,,\,\frac{1}{\alpha^m}\log^m\biggl(\frac{V_d^\alpha\nu^d(\alpha+d)^{\alpha+d}}{\alpha^\alpha d^d}\biggr)\biggr\},
\end{aligned}$$

where the bound on $\bigl|\log^m \|f\|_\infty\bigr|$ comes from (2.13).

(ii) Now define

$$A_{d,\theta} := \max\biggl\{\sup_{f \in \mathcal{F}_{d,\theta}}|H(f)|\,,\,-\frac{1}{2}\log\inf_{f \in \mathcal{F}_{d,\theta}}\|f\|_\infty\,,\,1\biggr\}$$

and the set $S_{d,\theta} := \{x \in \mathcal{X} : e^{-4A_{d,\theta}} \leq f(x) \leq e^{-2A_{d,\theta}}\}$. For $f \in \mathcal{F}_{d,\theta}, x \in S_{d,\theta}$ and $y \in B_x(\{8d^{1/2}a(e^{-4A_{d,\theta}})\}^{-1/(\beta\wedge 1)})$ we have by Lemma 2.12 below that

$$|f(y) - f(x)| \leq \frac{15d^{1/2}}{7}a(e^{-4A_{d,\theta}})e^{-2A_{d,\theta}}\|y-x\|^{\beta\wedge 1}. \tag{2.30}$$

By the continuity of f, there exists $x_0 \in S_{d,\theta}$ such that $f(x_0) = \frac{1}{2}e^{-2A_{d,\theta}}(1+e^{-2A_{d,\theta}})$. Thus, by (2.30), we have that $B_{x_0}(r_{d,\theta}) \subseteq S_{d,\theta}$, where

$$r_{d,\theta} := \biggl\{\frac{7(1-e^{-2A_{d,\theta}})}{30d^{1/2}a(e^{-4A_{d,\theta}})}\biggr\}^{1/(\beta\wedge 1)} \wedge \frac{1}{8d^{1/2}a(e^{-4A_{d,\theta}})\}^{1/(\beta\wedge 1)}}.$$

Hence

$$V(f) = \mathbb{E}_f[\{\log f(X_1) + H(f)\}^2] \geq A_{d,\theta}^2 \mathbb{P}_f(X_1 \in S_{d,\theta}) \geq A_{d,\theta}^2 e^{-4A_{d,\theta}} V_d r_{d,\theta}^d,$$

as required. □

The following auxiliary result provides control on deviations of the density arising from the smoothness condition of our $\mathcal{F}_{d,\theta}$ classes.

Lemma 2.12. *For $\theta = (\alpha,\beta,\gamma,\nu,a) \in \Theta$, $m := \lceil \beta \rceil - 1$, $f \in \mathcal{F}_{d,\theta}$ and $y \in B_x\bigl(r_a(x)\bigr)$, we have, for multi-indices t with $|t| \leq m$, that*

$$\biggl|\frac{\partial f^t(y)}{\partial x^t} - \frac{\partial f^t(x)}{\partial x^t}\biggr| \leq \frac{15d^{1/2}}{7} a\bigl(f(x)\bigr) f(x) \|y-x\|^{\min(\beta-|t|,1)}.$$

Proof of Lemma 2.12. If $|t| = m$ then the result follows immediately from the definition of $\mathcal{F}_{d,\theta}$. Henceforth, therefore, assume that $m \geq 1$ and $|t| \leq m-1$. Writing $|||\cdot|||$ here for the largest absolute entry of an array, we have for $y \in B_x\bigl(r_a(x)\bigr)$ that

$$\begin{aligned}
&\biggl|\frac{\partial f^t(y)}{\partial x^t} - \frac{\partial f^t(x)}{\partial x^t}\biggr| \leq \|y-x\| \sup_{z \in B_x(\|y-x\|)} \biggl\|\nabla \frac{\partial f^t(z)}{\partial x^t}\biggr\| \\
&\leq \|y-x\|\|f^{(|t|+1)}(x)\| + d^{1/2}\|y-x\| \sup_{z \in B_x(\|y-x\|)} \Bigl|\Bigl|\Bigl| f^{(|t|+1)}(z) - f^{(|t|+1)}(x) \Bigr|\Bigr|\Bigr| \\
&\leq \sum_{\ell=1}^{m-|t|} d^{(\ell-1)/2}\|y-x\|^\ell \|f^{(|t|+\ell)}(x)\| + d^{m/2}\|y-x\|^m \sup_{z \in B_x(\|y-x\|)} \Bigl|\Bigl|\Bigl| f^{(m)}(z) - f^{(m)}(x) \Bigr|\Bigr|\Bigr| \\
&\leq a(f(x))f(x)\|y-x\| \biggl\{\frac{1}{1-d^{1/2}\|y-x\|} + d^{m/2}\|y-x\|^{\beta-1}\biggr\} \leq \frac{15d^{1/2}}{7} a(f(x))f(x)\|y-x\|,
\end{aligned}$$

as required. □

Lemma 2.13. *Under the conditions of Theorem 2.1 we have that*

$$\sup_{k \in \{k_0^*,\ldots,k_1^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} \mathbb{E}_f[\{\tilde{V}_n^w - V(f)\}^2] \to 0.$$

Proof of Lemma 2.13. For $w = (w_1,\ldots,w_k)^T \in \mathcal{W}^{(k)}$, write $\mathrm{supp}(w) := \{j : w_j \neq 0\}$. Then

$$\begin{aligned}
|\mathbb{E}_f \tilde{V}_n^w - V(f)| &\leq \biggl|\sum_{j=1}^k w_j \mathbb{E}_f \log^2 \xi_{(j),1} - \int_{\mathcal{X}} f \log^2 f\biggr| + \bigl|\mathbb{E}_f\{(\hat{H}_n^w)^2\} - H(f)^2\bigr| \\
&\leq \|w\|_1 \max_{j \in \mathrm{supp}(w)} \biggl|\mathbb{E}_f \log^2 \xi_{(j),1} - \int_{\mathcal{X}} f \log^2 f\biggr| + \mathrm{Var}_f \hat{H}_n^w + |(\mathbb{E}_f \hat{H}_n^w)^2 - H(f)^2|.
\end{aligned}$$

Thus, by Theorem 2.1, (2.18) in the proof of that result and Lemma 2.11(i), we have that

$\sup_{k\in\{k_0^*,\ldots,k_1^*\}} \sup_{f\in\mathcal{F}_{d,\theta}} |\mathbb{E}_f \tilde{V}_n^w - V(f)| \to 0$. Now,

$$\mathrm{Var}_f \tilde{V}_n^w \leq \frac{\|w\|_1^2}{n} \max_{j\in\mathrm{supp}(w)} \mathrm{Var}_f \log^2 \xi_{(j),1} + \|w\|_1^2 \max_{j,\ell\in\mathrm{supp}(w)} \bigl|\mathrm{Cov}_f(\log^2 \xi_{(j),1}, \log^2 \xi_{(\ell),2})\bigr|. \tag{2.31}$$

Let $a_{n,j}^- := (j - 3j^{1/2}\log^{1/2} n) \vee 0$ and $a_{n,j}^+ := (j + 3j^{1/2}\log^{1/2} n) \wedge (n-1)$. Mimicking arguments in the proof of Theorem 2.1, for any $m \in \mathbb{N}$, $j \in \mathrm{supp}(w)$ and $\epsilon > 0$,

$$\begin{aligned}\mathbb{E}_f\bigl\{\log^m(\xi_{(j),1} f(X_1))\bigr\} &= \int_{\mathcal{X}} f(x) \int_0^\infty \log^m\biggl(\frac{V_d(n-1)f(x)h_x^{-1}(s)^d}{e^{\Psi(j)}}\biggr) \mathrm{B}_{j,n-j}(s)\,ds\,dx \\ &= \int_{\frac{a_{n,j}^-}{n-1}}^{\frac{a_{n,j}^+}{n-1}} \log^m\biggl(\frac{(n-1)s}{e^{\Psi(j)}}\biggr)\mathrm{B}_{j,n-j}(s)\,ds + O\biggl(\max\biggl\{\frac{k^{\beta/d}}{n^{\beta/d}}\log^{m-1} n\,,\, \frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}\biggr\}\biggr) \to 0,\end{aligned}$$

uniformly for $j \in \mathrm{supp}(w)$, $k \in \{k_0^*, \ldots, k_1^*\}$ and $f \in \mathcal{F}_{d,\theta}$. Moreover, by Cauchy–Schwarz, we can now show, for example, that

$$\mathbb{E}_f \log^4 \xi_{(j),1} = \mathbb{E}_f[\{\log(\xi_{(j),1} f(X_1)) - \log f(X_1)\}^4] \to \mathbb{E}_f \log^4 f(X_1)$$

uniformly for $j \in \mathrm{supp}(w)$, $k \in \{k_0^*, \ldots, k_1^*\}$ and $f \in \mathcal{F}_{d,\theta}$. Using a similar approach for the covariance term in (2.31) we see that $\sup_{k\in\{k_0^*,\ldots,k_1^*\}} \sup_{f\in\mathcal{F}_{d,\theta}} \mathrm{Var}\, \tilde{V}_n^w \to 0$ and the result follows. □

2.6.3 Proof of Proposition 2.5

Proof of Proposition 2.5. In each of the three examples, we provide $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$ such that $f \in \mathcal{F}_{d,\theta}$. In fact, $\beta > 0$ may be chosen arbitrarily in each case.

(i) We may choose any $\alpha > 0$, and then set $\nu = d2^{\alpha/2-1}\Gamma\bigl(\frac{\alpha}{2} + \frac{d}{2}\bigr)/\Gamma\bigl(1 + \frac{d}{2}\bigr)$. We may also set $\gamma = (2\pi)^{-d/2}$. It remains to find $a \in \mathcal{A}$ such that (2.6) holds. Write $H_r(y) := (-1)^r e^{y^2/2}\frac{d^r}{dy^r}e^{-y^2/2}$ for the rth Hermite polynomial, and note that $|H_r(y)| \leq p_r(|y|)$, where p_r is a polynomial of degree r with non-negative coefficients. Using multi-index notation for partial derivatives, if $t = (t_1, \ldots, t_d) \in \{0, 1, \ldots, \}^d$ with $|t| := t_1 + \ldots + t_d$, we have

$$\biggl|\frac{\partial f^t(x)}{\partial x^t}\biggr| = f(x)\prod_{j=1}^d |H_{t_j}(x_j)| \leq f(x)\prod_{j=1}^d p_{t_j}(\|x\|) \leq f(x) q_{|t|}(\|x\|),$$

for some polynomial q_r of degree r, with non-negative coefficients. It follows that if $y \in B_x^\circ(1)$, then for any $\beta > 0$ with $m = \lceil\beta\rceil - 1$,

$$\begin{aligned}\frac{\|f^{(m)}(x) - f^{(m)}(y)\|}{f(x)\|y-x\|^{\beta-m}} &\leq \frac{d^{m/2}}{f(x)\|y-x\|^{\beta-m}} \max_{t:|t|=m} \biggl|\frac{\partial f^t(x)}{\partial x^t} - \frac{\partial f^t(y)}{\partial x^t}\biggr| \\ &\leq \frac{d^{(m+1)/2}}{f(x)} \max_{t:|t|=m+1} \sup_{w\in B_0(1)} \biggl|\frac{\partial f^t(x+w)}{\partial x^t}\biggr| \\ &\leq d^{(m+1)/2} \sup_{w\in B_0(1)} \frac{f(x+w)q_{m+1}(\|x+w\|)}{f(x)} \\ &\leq d^{(m+1)/2} e^{\|x\|} q_{m+1}(\|x\| + 1).\end{aligned}$$

Similarly,

$$\max_{r=1,\ldots,m} \frac{\|f^{(r)}(x)\|}{f(x)} \leq d^{m/2} \max_{r=1,\ldots,m} q_r(\|x\|).$$

Write $g(\delta) := \bigl\{-2\log\bigl(\delta(2\pi)^{d/2}\bigr)\bigr\}^{1/2}$ and define $a \in \mathcal{A}$ by setting $a(\delta) := \max\{1, \tilde{a}(\delta)\}$, where

$$\begin{aligned}\tilde{a}(\delta) &:= d^{m/2} \sup_{x:\|x\|\leq g(\delta)} \max\biggl\{\max_{r=1,\ldots,m} q_r(\|x\|)\,,\, d^{1/2}e^{\|x\|}q_{m+1}(\|x\|+1)\biggr\} \\ &= d^{m/2}\max\biggl\{\max_{r=1,\ldots,m} q_r\bigl(g(\delta)\bigr)\,,\, d^{1/2}e^{g(\delta)}q_{m+1}\bigl(g(\delta)+1\bigr)\biggr\}.\end{aligned}$$

Then $\sup_{x:f(x)\geq\delta} M_{f,a,\beta}(x) \leq a(\delta)$ and $a(\delta) = o(\delta^{-\epsilon})$ for every $\epsilon > 0$, so (2.6) holds.

(ii) We may choose any $\alpha < \rho$, and set

$$\nu = d2^{\alpha/2-1}\frac{\Gamma\bigl(\frac{\alpha}{2}+\frac{d}{2}\bigr)}{\Gamma\bigl(1+\frac{d}{2}\bigr)}\frac{(\rho/2)^{\alpha/2}\Gamma\bigl(\frac{\rho-\alpha}{2}\bigr)}{\Gamma\bigl(\frac{\rho}{2}\bigr)}.$$

We may also set $\gamma = \frac{\Gamma\left(\frac{\rho}{2}+\frac{d}{2}\right)}{\Gamma(\rho/2)\rho^{\alpha/2}\pi^{d/2}}$. To verify (2.6) for suitable $a \in \mathcal{A}$, we note by induction, that if $t = (t_1, \ldots, t_d) \in \{0, 1, \ldots, \}^d$ with $|t| := t_1 + \ldots + t_d$, then

$$\left|\frac{\partial f^t(x)}{\partial x^t}\right| \leq \frac{f(x)q_{|t|}(\|x\|)}{(1+\|x\|^2/\rho)^{|t|}},$$

where q_r is a polynomial of degree r with non-negative coefficients. Thus, similarly to the Gaussian example, for any $\beta > 0$ with $m = \lceil\beta\rceil - 1$,

$$\sup_{x\in\mathbb{R}^d}\sup_{y\in B_x^\circ(1)} \frac{\|f^{(m)}(x)-f^{(m)}(y)\|}{f(x)\|y-x\|^{\beta-m}} \leq d^{(m+1)/2}\sup_{x\in\mathbb{R}^d}\sup_{w\in B_0(1)} \frac{f(x+w)q_{m+1}(\|x+w\|)}{f(x)(1+\|x\|^2/\rho)^{m+1}} =: A^{(1)}_{d,m,\rho},$$

say, where $A^{(1)}_{d,m,\rho} \in [0,\infty)$. Similarly,

$$\sup_{x\in\mathbb{R}^d}\max_{r=1,\ldots,m}\frac{\|f^{(r)}(x)\|}{f(x)} \leq d^{m/2}\sup_{x\in\mathbb{R}^d}\max_{r=1,\ldots,m}\frac{q_r(\|x\|)}{(1+\|x\|^2/\rho)^r} =: A^{(2)}_{d,m,\rho},$$

say, where $A^{(2)}_{d,m,\rho} \in [0,\infty)$. Now defining $a \in \mathcal{A}$ to be the constant function

$$a(\delta) := \max\{1, A^{(1)}_{d,m,\rho}, A^{(2)}_{d,m,\rho}\},$$

we again have that $\sup_{x:f(x)\geq\delta} M_{f,a,\beta}(x) \leq a(\delta)$, so (2.6) holds.

(iii) We may take any $\alpha > 0$ and $\nu = 1$, $\gamma = 3$. To verify (2.6), fix $\beta > 0$, set $m := \lceil\beta\rceil - 1$, and define $a \in \mathcal{A}$ by

$$a(\delta) := A_m \max\bigl\{1\,,\, \log^{2(m+1)}(1/\delta)\bigr\},$$

for some $A_m \geq 1$ depending only on m. Then, by induction, we find that for some constants

$A'_m, B'_m > 0$ depending only on m, and $x \in (-1,1)$

$$\begin{aligned}M_{f,a,\beta}(x) &\leq \max\biggl\{\max_{r=1,\ldots,m} \frac{A'_r}{(1-x^2)^{2r}}\,,\ \sup_{y:0<|y-x|\leq r_a(x)} \frac{A'_{m+1}f(y)}{(1-y^2)^{2(m+1)}f(x)}\biggr\} \\ &\leq \frac{B'_{m+1}}{(1-x^2)^{2(m+1)}} \leq a\bigl(f(x)\bigr),\end{aligned}$$

provided A_m in the definition of a is chosen sufficiently large. Hence (2.6) again holds. □

2.6.4 Proof of Proposition 2.6

Proof of Proposition 2.6. To deal with the integrals over $\mathcal{X}_n^c$, we first observe that by (2.13) there exists a constant $C_{d,f} > 0$, depending only on d and f, such that

$$\begin{aligned}&\int_{\mathcal{X}_n^c} f(x) \int_0^1 \mathrm{B}_{k,n-k}(s) \log u_{x,s}\, ds\, dx \\ &\leq C_{d,f} \int_{\mathcal{X}_n^c} f(x)\biggl\{\log n + \log\biggl(1 + \frac{\|x\|}{\mu_\alpha^{1/\alpha}(f)}\biggr)\biggr\}\, dx = O\bigl(\max\{q_n \log n, q_n^{1-\epsilon}\}\bigr), \qquad (2.32)\end{aligned}$$

for every $\epsilon > 0$. Moreover,

$$\biggl|\int_{\mathcal{X}_n^c} f(x) \log f(x)\, dx\biggr| = O(q_n^{1-\epsilon}), \qquad (2.33)$$

for every $\epsilon > 0$. Now, a slightly simpler argument than that used in the proof of Lemma 2.10(ii) gives that for $r \in (0, r_x]$, we have

$$|h_x(r) - V_d f(x) r^d| \leq \frac{dV_d}{d+\tilde{\beta}} C_{n,\tilde{\beta}}(x) r^{d+\tilde{\beta}}.$$

We deduce, again using a slightly simplified version of the argument in Lemma 2.10(ii), that there exists $n_0 \in \mathbb{N}$ such that for $n \geq n_0$, $s \in [0, \frac{a_n}{n-1}]$ and $x \in \mathcal{X}_n$, we have

$$\bigl|V_d f(x) h_x^{-1}(s)^d - s\bigr| \leq \frac{2dV_d^{-\tilde{\beta}/d}}{d+\tilde{\beta}} s^{1+\tilde{\beta}/d} \frac{C_{n,\tilde{\beta}}(x)}{f(x)^{1+\tilde{\beta}/d}} \leq \frac{s}{2}. \qquad (2.34)$$

It follows from (2.32), (2.33), (2.34) and an almost identical argument to that leading to (2.15) that for every $n \geq n_0$ and $\epsilon > 0$,

$$\begin{aligned}|\mathbb{E}\hat{H}_n - H| &\leq \biggl|\int_{\mathcal{X}_n} f(x) \int_0^{\frac{a_n}{n-1}} \mathrm{B}_{k,n-k}(s) \log\biggl(\frac{V_d f(x) h_x^{-1}(s)^d}{s}\biggr)\, ds\, dx\biggr| + O\bigl(\max\{q_n^{1-\epsilon}, q_n \log n, n^{-1}\}\bigr) \\ &\leq 2\int_{\mathcal{X}_n} f(x) \int_0^{\frac{a_n}{n-1}} \mathrm{B}_{k,n-k}(s) \biggl|\frac{V_d f(x) h_x^{-1}(s)^d - s}{s}\biggr|\, ds\, dx + O\bigl(\max\{q_n^{1-\epsilon}, q_n \log n, n^{-1}\}\bigr) \\ &\leq \frac{4dV_d^{-\tilde{\beta}/d}}{d+\tilde{\beta}} \frac{\mathrm{B}_{k+\tilde{\beta}/d,n-k}}{\mathrm{B}_{k,n-k}} \int_{\mathcal{X}_n} \frac{C_{n,\tilde{\beta}}(x)}{f(x)^{\tilde{\beta}/d}}\, dx + O\bigl(\max\{q_n^{1-\epsilon}, q_n \log n, n^{-1}\}\bigr),\end{aligned}$$

as required. □

2.6.5 Completion of the proof of Lemma 2.7

To prove Lemma 2.7, it remains to bound several error terms arising from arguments that approximate the variance of the unweighted Kozachenko–Leonenko estimator $\hat{H}_n$, and then to show how these arguments may be adapted to yield the desired asyptotic expansion for $\mathrm{Var}(\hat{H}_n^w)$.

Bounds on $S_1, \ldots, S_5$

To bound S_1: By similar methods to those used to bound R_1 in the proof of Lemma 2.3 it is straightforward to show that for every $\epsilon > 0$, we have

$$S_1 = \int_{\mathcal{X}_n^c} f(x) \int_0^1 \mathrm{B}_{k,n-k}(s) \log^2 u_{x,s} \, ds \, dx = O\biggl(\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}\biggr).$$

To bound S_2: For every $\epsilon > 0$, we have that

$$S_2 = \int_{\mathcal{X}_n} f(x) \int_{\frac{a_n}{n-1}}^1 \mathrm{B}_{k,n-k}(s) \log^2 u_{x,s} \, ds \, dx = o(n^{-(3-\epsilon)}),$$

by very similar arguments to those used to bound R_2 in the proof of Lemma 2.3.

To bound S_3: We have

$$\begin{aligned}\log^2 u_{x,s} - \log^2\biggl(\frac{(n-1)s}{e^{\Psi(k)}f(x)}\biggr) \\ &= \biggl\{2\log\biggl(\frac{(n-1)s}{e^{\Psi(k)}f(x)}\biggr) + \log\biggl(\frac{V_d f(x) h_x^{-1}(s)^d}{s}\biggr)\biggr\} \log\biggl(\frac{V_d f(x) h_x^{-1}(s)^d}{s}\biggr).\end{aligned}$$

It therefore follows from Lemma 2.10(ii) that for every $\epsilon > 0$,

$$\begin{aligned} S_3 &= \int_{\mathcal{X}_n} f(x) \int_0^{\frac{a_n}{n-1}} \mathrm{B}_{k,n-k}(s) \biggl\{\log^2 u_{x,s} - \log^2\biggl(\frac{(n-1)s}{e^{\Psi(k)}f(x)}\biggr)\biggr\} \, ds \, dx \\ &= O\biggl\{\max\biggl(\frac{k^{\beta/d}}{n^{\beta/d}} \log n \, , \, \frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}\biggr)\biggr\}.\end{aligned}$$

To bound S_4: A simplified version of the argument used to bound R_4 in Lemma 2.3 of the main text shows that for every $\epsilon > 0$,

$$S_4 = \int_{\mathcal{X}_n} f(x) \int_{\frac{a_n}{n-1}}^1 \mathrm{B}_{k,n-k}(s) \log^2\biggl(\frac{(n-1)s}{e^{\Psi(k)}f(x)}\biggr) \, ds \, dx = o(n^{-(3-\epsilon)}).$$

To bound S_5: Very similar arguments to those used to bound R_1 in Lemma 2.3 show that for every $\epsilon > 0$,

$$S_5 = \int_{\mathcal{X}_n^c} f(x) \log^2 f(x) \, dx = O\biggl(\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}\biggr).$$

Bounds on T_1, T_2 and T_3

To bound T_1: Let $\mathrm{B} \sim \mathrm{Beta}(k-1, n-k-1)$. By (2.13), for every $\epsilon > 0$,

$$\begin{aligned}
T_{11} &:= \biggl| \int_{\mathcal{X}_n^c \times \mathcal{X}_n^c} f(x) f(y) \log f(y) \int_{\tilde{u}_{n,x,y}}^{\infty} \log(uf(x))\, d(\tilde{F}_{n,x} - F_{n,x}^{-})(u)\, dy\, dx \biggr| \\
&\leq \frac{n-2}{n-k-1} \int_{\mathcal{X}_n^c \times \mathcal{X}_n^c} f(x) f(y) |\log f(y)| \int_0^1 \bigl|\log(u_{x,s} f(x))\bigr| \mathrm{B}_{k-1,n-k-1}(s) \biggl| 1 - \frac{(n-2)s}{k-1} \biggr|\, ds\, dy\, dx \\
&\lesssim \int_{\mathcal{X}_n^c \times \mathcal{X}_n^c} f(x) f(y) |\log f(y)| \biggl[\mathbb{E}\biggl\{ \biggl(\log \frac{1}{\mathrm{B}} + \log \frac{1}{1-B} \biggr) \biggl| 1 - \frac{(n-2)\mathrm{B}}{k-1} \biggr| \biggr\} \\
&\qquad + \biggl\{ \log n + |\log f(x)| + \log\biggl(1 + \frac{\|x\|}{\mu_\alpha^{1/\alpha}(f)} \biggr) \biggr\} \mathbb{E}\biggl| 1 - \frac{(n-2)\mathrm{B}}{k-1} \biggr| \biggr]\, dy\, dx \\
&= o\biggl(\frac{k^{-\frac{1}{2} + \frac{2\alpha}{\alpha+d} - \epsilon}}{n^{\frac{2\alpha}{\alpha+d} - \epsilon}} \biggr),
\end{aligned}$$

where we used the Cauchy–Schwarz inequality and elementary properties of beta random variables to obtain the final bound.

Now let

$$u_n^*(x) := u_{x, a_n/(n-1)} = \frac{V_d (n-1) h_x^{-1}(\frac{a_n}{n-1})^d}{e^{\Psi(k)}},$$

and consider

$$T_{12} := \biggl| \int_{\mathcal{X}_n^c} \int_{\mathcal{X}_n} f(x) f(y) \log f(y) \int_{\tilde{u}_{n,x,y}}^{\infty} \log(uf(x))\, d(\tilde{F}_{n,x} - F_{n,x}^{-})(u)\, dy\, dx \biggr|.$$

If $\tilde{u}_{n,x,y} \geq u_n^*(x)$, then by very similar arguments to those used to bound R_1 and R_2 (cf. (2.13) and (2.14)), together with Cauchy–Schwarz,

$$\begin{aligned}
\int_{\tilde{u}_{n,x,y}}^{\infty} &\bigl|\log(uf(x))\bigr|\, d(\tilde{F}_{n,x} - F_{n,x}^{-})(u) \\
&\leq \int_{\frac{a_n}{n-1}}^{1} |\log(u_{x,s} f(x))| \{\mathrm{B}_{k-1,n-k}(s) + \mathrm{B}_{k,n-k-1}(s)\}\, ds \\
&\lesssim \frac{\log n + |\log f(x)| + \log\Bigl(1 + \frac{\|x\|}{\mu_\alpha^{1/\alpha}(f)}\Bigr)}{n^{3-\epsilon}}, \qquad (2.35)
\end{aligned}$$

for every $\epsilon > 0$. On the other hand, if $\tilde{u}_{n,x,y} < u_n^*(x)$, then $\|x - y\| < r_{n,u_n^*(x)} + r_{n,u_n^*(y)}$, where we have added the $r_{n,u_n^*(y)}$ term to aid a calculation later in the proof. Define the sequence

$$\rho_n := \bigl[c_n \log^{1/d}(n-1) \bigr]^{-1}.$$

From Lemma 2.10(ii),

$$\sup_{y \in \mathcal{X}_n} r_{n,u_n^*(y)} = \sup_{y \in \mathcal{X}_n} h_y^{-1}\Bigl(\frac{a_n}{n-1} \Bigr) \lesssim \sup_{y \in \mathcal{X}_n} \biggl\{ \frac{k \log n}{n f(y)} \biggr\}^{1/d} \leq \biggl(\frac{k \log n}{n \delta_n} \biggr)^{1/d} = o(\rho_n).$$

Now suppose that $x \in \mathcal{X}_n^c$ and $y \in \mathcal{X}_n$ satisfy $\|y - x\| \leq \rho_n$. Choose $n_0 \in \mathbb{N}$ large enough that $r_{n,u_n^*(y)} \leq \rho_n/2$ for all $y \in \mathcal{X}_n$, and that $\log(n-1) \geq \max\{(3/2)^d (8d^{1/2})^{d/\beta}, 12 V_d^{-1} 2^d\}$ for all $n \geq n_0$ and $k \in \{k_0^*, \ldots, k_1^*\}$. Then when $\beta \in (0,1]$ and $n \geq n_0$, using the fact that $B_x(\rho_n/2) \subseteq B_y(3\rho_n/2)$, we have

$$\int_{B_x(\rho_n/2)} f(w)\,dw \geq V_d f(y)(\rho_n/2)^d - V_d a(f(y))f(y)(\rho_n/2)^d(3\rho_n/2)^\beta$$
$$\geq V_d f(y)(\rho_n/2)^d\{1-(3c_n\rho_n/2)^\beta\} \geq \frac{1}{2}V_d(\rho_n/2)^d\delta_n \geq \frac{a_n}{n-1}. \tag{2.36}$$

Hence, for all $n \geq n_0$, $x \in \mathcal{X}_n^c$, $y \in \mathcal{X}_n$ with $\|y-x\| \leq \rho_n$ and $k \in \{k_0^*, \ldots, k_1^*\}$,

$$r_{n,u_n^*(x)} + r_{n,u_n^*(y)} \leq \rho_n. \tag{2.37}$$

On other hand, suppose instead that $x \in \mathcal{X}_n^c$ and $\rho_x^* := \inf_{y\in\mathcal{X}_n}\|y-x\| \geq \rho_n$. Since $\mathcal{X}_n$ is a closed subset of $\mathbb{R}^d$, we can find $y^* \in \mathcal{X}_n$ such that $\|y^*-x\| = \rho_x^*$, and set $\tilde{x} := \frac{\rho_n}{\rho_x^*}x + \bigl(1-\frac{\rho_n}{\rho_x^*}\bigr)y^*$. Then $\|\tilde{x}-y^*\| = \rho_n$, so from (2.36), we have $r_{n,u_n^*(\tilde{x})} \leq \rho_n/2$ for $n \geq n_0$ and $k \in \{k_0^*, \ldots, k_1^*\}$. Since $B_{\tilde{x}}(\rho_n/2) \subseteq B_x(\rho_x^*-\rho_n/2)$, we deduce that $r_{n,u_n^*(x)} \leq \rho_x^* - \rho_n/2$ and

$$\{y \in \mathcal{X}_n : \|x-y\| < r_{n,u_n^*(x)} + r_{n,u_n^*(y)}\} = \emptyset \tag{2.38}$$

for $n \geq n_0$ and $k \in \{k_0^*, \ldots, k_1^*\}$. But for $n \geq n_0$,

$$\sup_{x\in\mathcal{X}_n^c}\ \sup_{y\in\mathcal{X}_n:\|y-x\|\leq\rho_n} \frac{1}{f(y)}|f(x)-f(y)| \leq \frac{15d^{1/2}}{7}(c_n\rho_n)^\beta < \frac{1}{2}, \tag{2.39}$$

so that if $x \in \mathcal{X}_n^c$, $y \in \mathcal{X}_n$ and $\|x-y\| \leq \rho_n$, then $f(y) < 2\delta_n$ for $n \geq n_0$ and $k \in \{k_0^*, \ldots, k_1^*\}$.

It therefore follows from (2.35), (2.37), (2.38), (2.39) and the argument used to bound T_{11} that for each $\epsilon > 0$ and $n \geq n_0$,

$$\begin{aligned} T_{12} &\leq \int_{\mathcal{X}_n^c}\int_{\mathcal{X}_n} f(x)f(y)|\log f(y)|\mathbb{1}_{\{\|x-y\|<r_{n,u_n^*(x)}+r_{n,u_n^*(y)}\}} \\ &\qquad\qquad \int_0^\infty |\log(uf(x))|\,d(\tilde{F}_{n,x}-F_{n,x}^-)(u)\,dy\,dx + o(n^{-2}) \\ &\leq \int_{\mathcal{X}_n^c}\int_{y:f(y)<2\delta_n} f(x)f(y)|\log f(y)|\int_0^\infty |\log(uf(x))|\,d(\tilde{F}_{n,x}-F_{n,x}^-)(u)\,dy\,dx + o(n^{-2}) \\ &= o\biggl(\frac{k^{-\frac{1}{2}+\frac{2\alpha}{\alpha+d}-\epsilon}}{n^{\frac{2\alpha}{\alpha+d}-\epsilon}}\biggr). \end{aligned}$$

Finally for T_1, we define

$$T_{13} := \biggl|\int_{\mathcal{X}_n}\int_{B_x^c\left(\frac{r_{n,1}d_n}{f(x)^{1/d}}\right)} f(x)f(y)\log f(y)\int_{\tilde{u}_{n,x,y}}^\infty \log\bigl(uf(x)\bigr)\,d(\tilde{F}_{n,x}-F_{n,x}^-)(u)\,dy\,dx\biggr|.$$

By Lemma 2.10(ii) we can find $n_1 \in \mathbb{N}$ such that for $n \geq n_1$, $k \in \{k_0^*, \ldots, k_1^*\}$, $x \in \mathcal{X}_n$ and $s \leq a_n/(n-1)$, we have $V_d f(x)h_x^{-1}(s)^d \leq 2s$. Thus, for $n \geq n_1$, $k \in \{k_0^*, \ldots, k_1^*\}$, $x \in \mathcal{X}_n$ and $y \in B_x^c(\frac{r_{n,1}d_n}{f(x)^{1/d}})$,

$$\tilde{u}_{n,x,y} \geq \frac{24\log n}{f(x)} \geq \frac{2a_n}{f(x)e^{\Psi(k)}} \geq u_n^*(x).$$

Thus, from (2.35), $T_{13} = O(n^{-2}\log n)$. We conclude that for every $\epsilon > 0$,

$$|T_1| \leq T_{11} + T_{12} + T_{13} = o\biggl(\frac{k^{-\frac{1}{2}+\frac{2\alpha}{\alpha+d}-\epsilon}}{n^{\frac{2\alpha}{\alpha+d}-\epsilon}}\biggr).$$

To bound T_2: Fix $x \in \mathcal{X}_n$ and $z \in B_0(d_n)$. Choosing $n_2 \in \mathbb{N}$ large enough that $\frac{r_{n,1}d_n}{\delta_n^{1/d}} \leq (8d^{1/2})^{-1/\beta}c_n^{-1}$ for $n \geq n_2$, we have by Lemma 2.12 that

$$\sup_{y \in B_x\bigl(\frac{r_{n,1}d_n}{\delta_n^{1/d}}\bigr)} \biggl|\frac{f(y)}{f(x)} - 1\biggr| \leq \frac{1}{2}$$

for $n \geq n_2$, $k \in \{k_0^*, \ldots, k_1^*\}$. Also, for all $n \geq n_2, k \in \{k_0^*, \ldots, k_1^*\}$, we have

$$\begin{aligned}\bigl|f(y_{x,z})\log f(y_{x,z}) - f(x)\log f(x)\bigr| &\leq f(y_{x,z})|\log(f(y_{x,z})/f(x))| + |\log f(x)||f(y_{x,z}) - f(x)| \\ &\leq a(f(x))f(x)\|y_{x,z} - x\|^\beta\{|\log f(x)| + 4\}.\end{aligned}$$

Moreover, by arguments used to bound T_{11},

$$\begin{aligned}\biggl|\int_{\|z\|^d/f(x)}^\infty \log(uf(x))\, d(\tilde{F}_{n,x} - F_{n,x}^-)(u)\biggr| &\lesssim \mathbb{E}\biggl|\log(\mathrm{B})\biggl(1 - \frac{(n-2)\mathrm{B}}{k-1}\biggr)\biggr| \\ &+ \biggl\{\log n + |\log f(x)| + \log\biggl(1 + \frac{\|x\|}{\mu_\alpha^{1/\alpha}(f)}\biggr)\biggr\}\mathbb{E}\biggl|1 - \frac{(n-2)\mathrm{B}}{k-1}\biggr|,\end{aligned}$$

where $\mathrm{B} \sim \mathrm{Beta}(k-1, n-k-1)$. It follows that for every $\epsilon > 0$,

$$\begin{aligned}T_2 &= \frac{e^{\Psi(k)}}{V_d(n-1)}\int_{\mathcal{X}_n}\int_{B_0(d_n)}\{f(y_{x,z})\log f(y_{x,z}) - f(x)\log f(x)\} \\ &\qquad\qquad \int_{\|z\|^d/f(x)}^\infty \log(uf(x))\, d(\tilde{F}_{n,x} - F_{n,x}^-)(u)\, dz\, dx \\ &= O\biggl(\frac{k^{1/2}}{n}\max\biggl\{\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}, \frac{k^{\beta/d}}{n^{\beta/d}}\log^{2+\beta/d} n\biggr\}\biggr).\end{aligned}$$

To bound T_3: Note that by Fubini's theorem,

$$\begin{aligned}&\int_{\mathcal{X}_n} f(x)\log f(x)\int_{B_0(d_n)}\int_{\frac{\|z\|^d}{f(x)}}^\infty \log(uf(x))\, d(\tilde{F}_{n,x} - F_{n,x}^-)(u)\, dz\, dx \\ &\quad = V_d\int_{\mathcal{X}_n} f(x)\log f(x)\int_0^\infty uf(x)\log(uf(x))\, d(\tilde{F}_{n,x} - F_{n,x}^-)(u)\, dx \\ &\quad = V_d\int_{\mathcal{X}_n} f(x)\log f(x)\int_0^{u_n^*(x)} uf(x)\log(uf(x))\, d(\tilde{F}_{n,x} - F_{n,x}^-)(u)\, dx + O(n^{-(3-\epsilon)}),\end{aligned}$$

for every $\epsilon > 0$, where the order of the error term follows from the same argument used to obtain (2.35) and Lemma 2.10(i). Thus, for every $\epsilon > 0$,

$$\begin{aligned} T_3 &= \frac{k-1}{n-k-1}\int_{\mathcal{X}_n} f(x)\log f(x)\int_0^{\frac{a_n}{n-1}}\biggl\{\frac{V_d f(x) h_x^{-1}(s)^d}{s}\log(u_{x,s}f(x)) \\ &\qquad - \log\biggl(\frac{(n-1)s}{e^{\Psi(k)}}\biggr)\biggr\}\mathrm{B}_{k,n-k-1}(s)\biggl\{1-\frac{(n-2)s}{k-1}\biggr\}\,ds\,dx + O(n^{-(3-\epsilon)}) \\ &= O\biggl(\frac{k^{1/2}}{n}\max\biggl\{\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}, \frac{k^{\beta/d}}{n^{\beta/d}}\log n\biggr\}\biggr). \end{aligned}$$

Bounds on U_1 and U_2

To bound U_1: Using Lemma 2.10(i) and (2.13) as in our bounds on T_{11} we have that for every $\epsilon > 0$,

$$\begin{aligned} U_{11} &:= \biggl|\int_{\mathcal{X}_n^c} f(x)\int_0^{u_n^*(x)} \log\bigl(uf(x)\bigr)\,d(F_{n,x}^- - F_{n,x})(u)\,dx\biggr| \\ &\leq \int_{\mathcal{X}_n^c} f(x)\int_0^{\frac{a_n}{n-1}} |\log(u_{x,s}f(x))|\mathrm{B}_{k,n-k-1}(s)\biggl|\frac{(n-1)s-k}{n-k-1}\biggr|\,ds\,dx = o\biggl(\frac{k^{\frac{1}{2}+\frac{\alpha}{\alpha+d}-\epsilon}}{n^{1+\frac{\alpha}{\alpha+d}-\epsilon}}\biggr). \end{aligned} \tag{2.40}$$

Moreover, using arguments similar to those used to bound R_2 in the proof of Lemma 2.3, for every $\epsilon > 0$,

$$U_{12} := \biggl|\int_{\mathcal{X}} f(x)\int_{u_n^*(x)}^{\infty} \log\bigl(uf(x)\bigr)\,d(F_{n,x}^- - F_{n,x})(u)\,dx\biggr| = o(n^{-(3-\epsilon)}). \tag{2.41}$$

From (2.40), and (2.41), we have for every $\epsilon > 0$ that

$$|U_1| \leq U_{11} + U_{12} = o\biggl(\frac{k^{\frac{1}{2}+\frac{\alpha}{\alpha+d}-\epsilon}}{n^{1+\frac{\alpha}{\alpha+d}-\epsilon}}\biggr).$$

To bound U_2: By Lemma 2.10(ii) and letting $\mathrm{B} \sim \mathrm{Beta}(k+\beta/d, n-k-1)$, we have that for every $\epsilon > 0$,

$$\begin{aligned} U_{21} &:= \biggl|\int_{\mathcal{X}_n} f(x)\int_0^{\frac{a_n}{n-1}}\log\Bigl(\frac{V_d f(x)h_x^{-1}(s)^d}{s}\Bigr)\mathrm{B}_{k,n-k-1}(s)\Bigl\{\frac{(n-1)s-k}{n-k-1}\Bigr\}ds\,dx\biggr| \\ &\lesssim \frac{k^{\beta/d}}{n^{\beta/d}}\mathbb{E}\biggl(\biggl|\frac{(n-1)\mathrm{B}-k}{n-k-1}\biggr|\biggr)\int_{\mathcal{X}_n} a(f(x))f(x)^{1-\beta/d}\,dx = O\biggl(\frac{k^{1/2}}{n}\max\biggl\{\frac{k^{\beta/d}}{n^{\beta/d}}, \frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}\biggr\}\biggr). \end{aligned}$$

Moreover, we can use similar arguments to those used to bound R_4 in the proof of Lemma 2.3 to show that for every $\epsilon > 0$,

$$U_{22} := \biggl|\int_{\mathcal{X}_n} f(x)\int_{\frac{a_n}{n-1}}^1 \log\biggl(\frac{(n-1)s}{e^{\Psi(k)}}\biggr)\mathrm{B}_{k,n-k-1}(s)\biggl\{\frac{(n-1)s-k}{n-k-1}\biggr\}\,ds\,dx\biggr| = o(n^{-(3-\epsilon)}).$$

We deduce that for every $\epsilon > 0$,

$$|U_2| \leq U_{21} + U_{22} = O\biggl(\frac{k^{1/2}}{n}\max\biggl\{\frac{k^{\beta/d}}{n^{\beta/d}}, \frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}\biggr\}\biggr).$$

Bounds on $W_1, \ldots, W_4$

To bound W_1: We partition the region $([l_x, v_x] \times [l_y, v_y])^c$ into eight rectangles as follows:

$$\begin{aligned}\bigl([l_x, v_x] \times [l_y, v_y]\bigr)^c = \bigl([0, l_x) \times [0, l_y)\bigr) \cup \bigl([0, l_x) \times [l_y, v_y]\bigr) \cup \bigl([0, l_x) \times (v_y, \infty)\bigr) \cup \bigl([l_x, v_x] \times [0, l_y)\bigr) \\ \cup \bigl([l_x, v_x] \times (v_y, \infty)\bigr) \cup \bigl((v_x, \infty) \times [0, l_y)\bigr) \cup \bigl((v_x, \infty) \times [l_y, v_y]\bigr) \cup \bigl((v_x, \infty) \times (v_y, \infty)\bigr).\end{aligned}$$

Recall our shorthand $h(u, v) = \log(uf(x)) \log(vf(y))$. By Lemma 2.10(i) and the Cauchy–Schwarz inequality, as well as very similar arguments to those used to bound R_2 in the proof of Lemma 2.3, we can bound the contributions from each rectangle individually, to obtain that for every $\epsilon > 0$,

$$W_1 = \int_{\mathcal{X} \times \mathcal{X}} f(x)f(y) \int_{([l_x, v_x] \times [l_y, v_y])^c} h(u, v)\, d(F_{n,x,y} - F_{n,x}F_{n,y})(u, v)\, dx\, dy = o(n^{-(9/2-\epsilon)}).$$

To bound W_2: We have

$$W_2 = \int_{\mathcal{X} \times \mathcal{X}} f(x)f(y) \int_{l_x}^{v_x} \int_{l_y}^{v_y} h(u, v)\, d(G_{n,x,y} - F_{n,x}F_{n,y})(u, v)\, dx\, dy + \frac{1}{n}.$$

We write $\mathrm{B}_{a,b,c} := \Gamma(a)\Gamma(b)\Gamma(c)/\Gamma(a + b + c)$, and, for $s, t > 0$ with $s + t < 1$, let $\mathrm{B}_{a,b,c}(s, t) := s^{a-1}t^{b-1}(1 - s - t)^{c-1}/\mathrm{B}_{a,b,c}$ denote the density of a Dirichlet(a, b, c) random vector at (s, t). For $a, b > -1$, writing $I_n := [a_n^-/(n - 1), a_n^+/(n - 1)]$, let

$$\begin{aligned}\mathrm{B}^{(n)}_{k+a,n-k} &:= \int_{I_n} s^{k+a-1}(1 - s)^{n-k-1}\, ds, \\ \mathrm{B}^{(n)}_{k+a,n-k}(s) &:= s^{k+a-1}(1 - s)^{n-k-1}/\mathrm{B}^{(n)}_{k+a,n-k} \\ \mathrm{B}^{(n)}_{k+a,k+b,n-2k-1} &:= \int_{I_n \times I_n} s^{k+a-1}t^{k+b-1}(1 - s - t)^{n-2k-2}\, ds\, dt \\ \mathrm{B}^{(n)}_{k+a,k+b,n-2k-1}(s, t) &:= s^{k+a-1}t^{k+b-1}(1 - s - t)^{n-2k-2}/\mathrm{B}^{(n)}_{k+a,k+b,n-2k-1}.\end{aligned}$$

Then by the triangle and Pinsker's inequalities, and Beta tail bounds similar to those used previously, we have that

$$\begin{aligned}&\int_{I_n \times I_n} \bigl|\mathrm{B}_{k+a,k+b,n-2k-1}(s, t) - \mathrm{B}_{k+a,n-k}(s)\mathrm{B}_{k+b,n-k}(t)\bigr|\, ds\, dt \\ &\leq \left|\frac{\mathrm{B}^{(n)}_{k+a,k+b,n-2k-1}}{\mathrm{B}_{k+a,k+b,n-2k-1}} - 1\right| + \left|\frac{\mathrm{B}^{(n)}_{k+a,n-k}\mathrm{B}^{(n)}_{k+b,n-k}}{\mathrm{B}_{k+a,n-k}\mathrm{B}_{k+b,n-k}} - 1\right| \\ &\qquad + \left\{2\int_{I_n \times I_n} \mathrm{B}^{(n)}_{k+a,k+b,n-2k-1}(s, t) \log\Bigl(\frac{\mathrm{B}^{(n)}_{k+a,k+b,n-2k-1}(s, t)}{\mathrm{B}^{(n)}_{k+a,n-k}(s)\mathrm{B}^{(n)}_{k+b,n-k}(t)}\Bigr)\, ds\, dt\right\}^{1/2} \\ &= \left\{2\int_0^1 \int_0^{1-t} \mathrm{B}_{k+a,k+b,n-2k-1}(s, t) \log\Bigl(\frac{\mathrm{B}_{k+a,k+b,n-2k-1}(s, t)}{\mathrm{B}_{k+a,n-k}(s)\mathrm{B}_{k+b,n-k}(t)}\Bigr)\, ds\, dt\right\}^{1/2} + o(n^{-2}) \\ &= 2^{1/2}\biggl[\log\Bigl(\frac{\Gamma(n + a + b - 1)\Gamma(n - k)^2}{\Gamma(n - 2k - 1)\Gamma(n + a)\Gamma(n + b)}\Bigr) + (n - 2k - 2)\psi(n - 2k - 1) \\ &\qquad - (n - k - 1)\{\psi(n + b - k - 1) + \psi(n + a - k - 1)\} + n\psi(n + a + b - 1)\biggr]^{1/2} + o(n^{-2}) \\ &= \frac{k}{n}\{1 + o(1)\}. \qquad (2.42)\end{aligned}$$

As a first step towards bounding W_2 note that

$$\begin{aligned}
W_{21} &:= \int_{\mathcal{X}_n \times \mathcal{X}_n} f(x)f(y) \int_{l_x}^{v_x} \int_{l_y}^{v_y} h(u,v)\, d(G_{n,x,y} - F_{n,x}F_{n,y})(u,v)\, dx\, dy \\
&= \int_{\mathcal{X}_n \times \mathcal{X}_n} f(x)f(y) \int_{I_n \times I_n} \log(u_{x,s}f(x)) \log(u_{y,t}f(y)) \\
&\qquad \big\{\mathrm{B}_{k,k,n-2k-1}(s,t) - \mathrm{B}_{k,n-k}(s)\mathrm{B}_{k,n-k}(t)\big\}\, ds\, dt\, dx\, dy \\
&= \int_{\mathcal{X}_n \times \mathcal{X}_n} f(x)f(y) \int_{I_n \times I_n} \log\biggl(\frac{(n-1)s}{e^{\Psi(k)}}\biggr) \log\biggl(\frac{(n-1)t}{e^{\Psi(k)}}\biggr) \\
&\qquad \big\{\mathrm{B}_{k,k,n-2k-1}(s,t) - \mathrm{B}_{k,n-k}(s)\mathrm{B}_{k,n-k}(t)\big\}\, ds\, dt\, dx\, dy + W_{211} \\
&= -\frac{1}{n} + O\biggl(\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{1+\frac{\alpha}{\alpha+d}-\epsilon}}\biggr) + O(n^{-2}) + W_{211}, \qquad (2.43)
\end{aligned}$$

for every $\epsilon > 0$. But, by Lemma 2.10(ii) and (2.42), for every $\epsilon > 0$,

$$\begin{aligned}
|W_{211}| &= \biggl|\int_{\mathcal{X}_n \times \mathcal{X}_n} f(x)f(y) \int_{I_n \times I_n} \biggl\{2\log\biggl(\frac{V_d h_x^{-1}(s)^d f(x)}{s}\biggr) \log\biggl(\frac{(n-1)t}{e^{\Psi(k)}}\biggr) \\
&\qquad + \log\biggl(\frac{V_d h_x^{-1}(s)^d f(x)}{s}\biggr) \log\biggl(\frac{V_d h_y^{-1}(t)^d f(y)}{t}\biggr)\biggr\} \\
&\qquad \big\{\mathrm{B}_{k,k,n-2k-1}(s,t) - \mathrm{B}_{k,n-k}(s)\mathrm{B}_{k,n-k}(t)\big\}\, ds\, dt\, dx\, dy\biggr| \\
&\le 2\biggl|\int_{\mathcal{X}_n \times \mathcal{X}_n} f(x)f(y) \int_{I_n} \log\biggl(\frac{V_d h_x^{-1}(s)^d f(x)}{s}\biggr) \\
&\qquad \Bigl[\big\{\log(n-1) - \Psi(n-k-1) + \log(1-s)\big\}\mathrm{B}_{k,n-k-1}(s) \\
&\qquad - \big\{\log(n-1) - \Psi(n)\big\}\mathrm{B}_{k,n-k}(s)\Bigr]\, ds\, dx\, dy\biggr| + O\biggl(\max\biggl\{\frac{k^{1+\frac{2\beta}{d}}}{n^{1+\frac{2\beta}{d}}}, \frac{k^{1+\frac{2\alpha}{\alpha+d}-\epsilon}}{n^{1+\frac{2\alpha}{\alpha+d}-\epsilon}}\biggr\}\biggr) \\
&= O\biggl(\frac{k^{1/2}}{n} \max\biggl\{\frac{k^{\beta/d}}{n^{\beta/d}}, \frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}\biggr\}\biggr). \qquad (2.44)
\end{aligned}$$

Moreover, by Lemma 2.10(i) and (ii) and very similar arguments, for every $\epsilon > 0$,

$$\begin{aligned}
W_{22} &:= \int_{\mathcal{X}_n \times \mathcal{X}_n^c} f(x)f(y) \int_{l_x}^{v_x} \int_{l_y}^{v_y} h(u,v)\, d(G_{n,x,y} - F_{n,x}F_{n,y})(u,v)\, dx\, dy \\
&= O\biggl(\frac{k^{1+\frac{\alpha}{\alpha+d}-\epsilon}}{n^{1+\frac{\alpha}{\alpha+d}-\epsilon}} \max\biggl\{\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}, \frac{k^{\beta/d}}{n^{\beta/d}}, \frac{1}{k^{1/2}}\biggr\}\biggr) \\
W_{23} &:= \int_{\mathcal{X}_n^c \times \mathcal{X}_n^c} f(x)f(y) \int_{l_x}^{v_x} \int_{l_y}^{v_y} h(u,v)\, d(G_{n,x,y} - F_{n,x}F_{n,y})(u,v)\, dx\, dy \\
&= O\biggl(\frac{k^{1+\frac{2\alpha}{\alpha+d}-\epsilon}}{n^{1+\frac{2\alpha}{\alpha+d}-\epsilon}}\biggr). \qquad (2.45)
\end{aligned}$$

Incorporating our restrictions on k, we conclude from (2.43), (2.44) and (2.45) that for every $\epsilon > 0$,

$$|W_2| \le \biggl|W_{21} + \frac{1}{n}\biggr| + 2|W_{22}| + |W_{23}| = O\biggl(\frac{k^{1/2}}{n} \max\biggl\{\frac{k^{\beta/d}}{n^{\beta/d}}, \frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}\biggr\}\biggr).$$

To bound W_3: We write h_u, h_v and h_{uv} for the partial derivatives of $h(u,v)$ and write, for example,

$(h_uF)(u,v) = h_u(u,v)F(u,v)$. We find on integrating by parts that, writing $F = F_{n,x,y} - G_{n,x,y}$,

$$\begin{aligned}
&\int_{[l_x,v_x]\times[l_y,v_y]} (h\,dF)(u,v) - \int_{l_x}^{v_x}\int_{l_y}^{v_y} (h_{uv}F(u,v))\,du\,dv \\
&= \int_{l_x}^{v_x} \big[(h_uF)(u,l_y) - (h_uF)(u,v_y)\big]du + \int_{l_y}^{v_y} \big[(h_vF)(l_x,v) - (h_vF)(v_x,v)\big]dv \\
&\qquad + (hF)(v_x,v_y) + (hF)(l_x,l_y) - (hF)(v_x,l_y) - (hF)(l_x,v_y). \qquad (2.46)
\end{aligned}$$

Using standard binomial tail bounds as used to bound W_1 together with (2.13) we therefore see that for every $\epsilon > 0$,

$$\begin{aligned}
W_{31} &:= \int_{\mathcal{X}\times\mathcal{X}} f(x)f(y)\Big\{\int_{l_x}^{v_x}\int_{l_y}^{v_y}(h\,dF)(u,v) - \int_{l_x}^{v_x}\int_{l_y}^{v_y}(h_{uv}F)(u,v)\,du\,dv\Big\}dx\,dy \\
&= -\int_{\mathcal{X}\times\mathcal{X}} f(x)f(y)\bigg\{\int_{l_x}^{v_x}(h_uF)(u,v_y)\,du + \int_{l_y}^{v_y}(h_vF)(v_x,v)\,dv\bigg\}\,dx\,dy + o(n^{-(9/2-\epsilon)}). \quad (2.47)
\end{aligned}$$

Now, uniformly for $u \in [l_x, v_x]$ and $(x,y) \in \mathcal{X}\times\mathcal{X}$ and for every $\epsilon > 0$,

$$\begin{aligned}
F(u,v_y) &= \mathbb{1}_{\{\|x-y\|\le r_{n,u}\}}\binom{n-2}{k-1}p_{n,x,u}^{k-1}(1-p_{n,x,u})^{n-k-1} + o(n^{-(9/2-\epsilon)}) \\
&= \mathbb{1}_{\{\|x-y\|\le r_{n,u}\}}\frac{\mathrm{B}_{k,n-k}(p_{n,x,u})}{n-1} + o(n^{-(9/2-\epsilon)}) \\
&\le \mathbb{1}_{\{\|x-y\|\le r_{n,v_x}\}}\frac{1}{(2\pi k)^{1/2}}\{1+o(1)\} + o(n^{-(9/2-\epsilon)}). \qquad (2.48)
\end{aligned}$$

By (2.39) and the arguments leading up to it, we have

$$\sup_{x\in\mathcal{X}_n^c}\ \sup_{y\in\mathcal{X}_n\cap B_x(r_{n,v_x}+r_{n,v_y})}\Big|\frac{f(x)}{f(y)} - 1\Big| \to 0. \qquad (2.49)$$

We therefore have by (2.13) that, for every $\epsilon > 0$,

$$\int_{\mathcal{X}_n^c\times\mathcal{X}} f(x)f(y)\int_{l_x}^{v_x}(h_uF)(u,v_y)\,du\,dy\,dx = O\bigg(\frac{k^{-\frac{1}{2}+\frac{2\alpha}{\alpha+d}-\epsilon}}{n^{\frac{2\alpha}{\alpha+d}-\epsilon}}\bigg). \qquad (2.50)$$

Now, using Lemma 2.10(ii), for $x \in \mathcal{X}_n$,

$$\max\{|l_xf(x)-1|, |v_xf(x)-1|\} \lesssim a(f(x))\Big(\frac{k}{nf(x)}\Big)^{\beta/d} + \frac{\log^{1/2} n}{k^{1/2}}. \qquad (2.51)$$

We also need some control over $vf(y)$. By (2.39) and the work leading up to it, for $n \ge \max(n_0, 5), x \in \mathcal{X}_n$ and $\|y-x\| \le r_{n,v_x} + r_{n,v_y}$,

$$f(y) \ge \Big\{1 - \frac{15d^{1/2}}{7}(c_n\rho_n)^{\beta}\Big\}\delta_n \ge \delta_n/2 \ge k/(n-1).$$

Thus $a(f(y)) \le c_n^{\beta}$ and using (2.49) we may apply Lemma 2.10(ii) to the set

$$\mathcal{X}_n' = \mathcal{X}_n \cup \{y : \|y-x\| \le r_{n,v_x} + r_{n,v_y} \text{ for some } x \in \mathcal{X}_n\}.$$

From this and (2.49), for any $x \in \mathcal{X}_n$ and $y \in B_x(r_{n,v_x} + r_{n,v_y})$,

$$\max(|l_y f(y) - 1|, |v_y f(y) - 1|) \lesssim a(f(y))\Big(\frac{k}{nf(x)}\Big)^{\beta/d} + \frac{\log^{1/2} n}{k^{1/2}}. \tag{2.52}$$

Using (2.49) again, we have that $a(f(y_{x,z})) \lesssim f(x)^{-\epsilon}$ for each $\epsilon > 0$, uniformly for $x \in \mathcal{X}_n$ and $\|z\| \leq \{v_x f(x)\}^{1/d} + \{v_y f(x)\}^{1/d}$. From (2.48), (2.51) and (2.52) we therefore have that

$$\begin{aligned}&\Big|\int_{\mathcal{X}_n \times \mathcal{X}} f(x)f(y) \int_{l_x}^{v_x} (h_u F)(u, v_y)\,du\,dy\,dx\Big| \\ &\lesssim k^{-1/2} \int_{\mathcal{X}_n \times \mathcal{X}} f(x)f(y)\mathbb{1}_{\{\|x-y\| < r_{n,v_x}\}} |\log(v_y f(y))| \log(v_x/l_x)\,dy\,dx \\ &= O\Big(\max\Big\{\frac{k^{1/2+2\beta/d}}{n^{1+2\beta/d}}, \frac{\log n}{nk^{1/2}}, \frac{k^{\frac{1}{2}+\frac{\alpha}{\alpha+d}-\epsilon}}{n^{1+\frac{\alpha}{\alpha+d}-\epsilon}}\Big\}\Big)\end{aligned} \tag{2.53}$$

for every $\epsilon > 0$. By (2.47), (2.50) and (2.53) we therefore have that

$$W_{31} = O\Big(\max\Big\{\frac{k^{1/2+2\beta/d}}{n^{1+2\beta/d}}, \frac{\log n}{nk^{1/2}}, \frac{k^{-1/2+\frac{2\alpha}{\alpha+d}-\epsilon}}{n^{\frac{2\alpha}{\alpha+d}-\epsilon}}\Big\}\Big). \tag{2.54}$$

Finally, by (2.13) and (2.49), we have since $F = 0$ when $\|x - y\| > r_{n,u} + r_{n,v}$ that

$$W_{32} := \int_{\mathcal{X}_n^c \times \mathcal{X}} f(x)f(y) \int_{l_x}^{v_x} \int_{l_y}^{v_y} (h_{uv}F)(u, v)\,du\,dv\,dx\,dy = O\Big(\frac{k^{\frac{2\alpha}{\alpha+d}-\epsilon}}{n^{\frac{2\alpha}{\alpha+d}-\epsilon}}\Big). \tag{2.55}$$

Combining (2.54) and (2.55) we have that

$$W_3 = W_{31} + W_{32} = O\Big(\max\Big\{\frac{k^{1/2+2\beta/d}}{n^{1+2\beta/d}}, \frac{\log n}{nk^{1/2}}, \frac{k^{\frac{2\alpha}{\alpha+d}-\epsilon}}{n^{\frac{2\alpha}{\alpha+d}-\epsilon}}\Big\}\Big).$$

To bound W_4: Let $p_\cap := \int_{B_x(r_{n,u}) \cap B_y(r_{n,v})} f(y)\,dy$ and let $(N_1, N_2, N_3, N_4) \sim \mathrm{Multi}(n-2, p_{n,x,u} - p_\cap, p_{n,y,v} - p_\cap, p_\cap, 1 - p_{n,x,u} - p_{n,y,v} + p_\cap)$. Further, let

$$F^{(1)}_{n,x,y}(u, v) := \mathbb{P}(N_1 + N_3 \geq k, N_2 + N_3 \geq k),$$

so that

$$\begin{aligned}(F_{n,x,y} - F^{(1)}_{n,x,y})(u, v) &= \mathbb{P}(N_1 + N_3 = k-1, N_2 + N_3 \geq k)\mathbb{1}_{\{\|x-y\| \leq r_{n,u}\}} \\ &+ \mathbb{P}(N_2 + N_3 = k-1, N_1 + N_3 \geq k)\mathbb{1}_{\{\|x-y\| \leq r_{n,v}\}} \\ &+ \mathbb{P}(N_1 + N_3 = k-1, N_2 + N_3 = k-1)\mathbb{1}_{\{\|x-y\| \leq r_{n,u} \wedge r_{n,v}\}}.\end{aligned}$$

Now $\mathbb{P}(N_1 + N_3 = k-1) = \binom{n-2}{k-1} p_{n,x,u}^{k-1}(1 - p_{n,x,u})^{n-k-1} \leq (2\pi k)^{-1/2}\{1 + o(1)\}$ and $F_{n,x,y}(u, v) =$

$G_{n,x,y}(u,v)$ if $\|x-y\| > r_{n,u}+r_{n,v}$, and so, by (2.51) and (2.52), we have that

$$\begin{aligned}\int_{\mathcal{X}_n\times\mathcal{X}} & f(x)f(y)\int_{l_x}^{v_x}\int_{l_y}^{v_y}\frac{(F_{n,x,y}-G_{n,x,y})(u,v)}{uv}\,du\,dv\,dx\,dy\\ &= \int_{\mathcal{X}_n\times\mathcal{X}} f(x)f(y)\int_{l_x}^{v_x}\int_{l_y}^{v_y}\frac{(F_{n,x,y}^{(1)}-G_{n,x,y})(u,v)}{uv}\,du\,dv\,dx\,dy\\ &\qquad + O\biggl(\max\biggl\{\frac{\log n}{nk^{1/2}}\,,\,\frac{k^{\frac{1}{2}+\frac{2\beta}{d}}}{n^{1+\frac{2\beta}{d}}}\,,\,\frac{k^{\frac{1}{2}+\frac{\alpha}{\alpha+d}-\epsilon}}{n^{1+\frac{\alpha}{\alpha+d}-\epsilon}}\biggr\}\biggr). \qquad (2.56)\end{aligned}$$

We can now approximate $F_{n,x,y}^{(1)}(u,v)$ by $\Phi_\Sigma(k^{1/2}\{uf(x)-1\}, k^{1/2}\{vf(x)-1\})$ and $G_{n,x,y}(u,v)$ by $\Phi(k^{1/2}\{uf(x)-1\})\Phi(k^{1/2}\{vf(x)-1\})$. To avoid repetition, we focus on the former of these terms. To this end, for $i=3,\ldots,n$, let

$$Y_i := \begin{pmatrix}\mathbb{1}_{\{X_i\in B_x(r_{n,u})\}}\\ \mathbb{1}_{\{X_i\in B_y(r_{n,v})\}}\end{pmatrix},$$

so that $\sum_{i=3}^n Y_i = \begin{pmatrix}N_1+N_3\\ N_2+N_3\end{pmatrix}$. We also define

$$\begin{aligned}\mu &:= \mathbb{E}(Y_i) = \begin{pmatrix}p_{n,x,u}\\ p_{n,y,v}\end{pmatrix}\\ V &:= \mathrm{Cov}(Y_i) = \begin{pmatrix}p_{n,x,u}(1-p_{n,x,u}) & p_\cap - p_{n,x,u}p_{n,y,v}\\ p_\cap - p_{n,x,u}p_{n,y,v} & p_{n,y,v}(1-p_{n,y,v})\end{pmatrix},\end{aligned}$$

When $x\in\mathcal{X}_n$ and $y\in B_x^\circ(r_{n,v_x}+r_{n,v_y})$ we have that, writing Δ for the symmetric difference and using (2.49), $\mathbb{P}(X_1\in B_x(r_{n,u})\Delta B_y(r_{n,v}))>0$ and so V is invertible. We may therefore set $Z_i := V^{-1/2}(Y_i-\mu)$. Then by the Berry–Esseen bound of Götze (1991), writing $\mathcal{C}$ for the set of closed, convex subsets of $\mathbb{R}^2$ and letting $Z\sim N_2(0,I)$, there exists a universal constant $C_2>0$ such that

$$\sup_{C\in\mathcal{C}}\biggl|\mathbb{P}\biggl(\frac{1}{(n-2)^{1/2}}\sum_{i=3}^n Z_i\in C\biggr)-\mathbb{P}(Z\in C)\biggr| \le \frac{C_2\mathbb{E}(\|Z_3\|^3)}{(n-2)^{1/2}}. \qquad (2.57)$$

The distribution of Z_3 depends on x,y,u and v, but, recalling the substitution $y=y_{x,z}$ as defined in (2.22), we claim that for $x\in\mathcal{X}_n$, $y=y_{x,z}\in B_x(r_{n,u}+r_{n,v})$, $u\in[l_x,v_x]$ and $v\in[l_y,v_y]$,

$$\mathbb{E}(\|Z_3\|^3) \lesssim \Bigl(\frac{n}{k\|z\|}\Bigr)^{1/2}. \qquad (2.58)$$

To establish this, note that for $x\in\mathcal{X}_n$ and $\|y-x\|\le r_{n,v_x}+r_{n,v_y}$, we have by (2.49), (2.51) and (2.52) that $\|y-x\|\lesssim(\frac{k}{nf(x)})^{1/d}$. Thus, for $v\in[l_y,v_y]$, and using Lemma 2.12, we also have that

$$\begin{aligned}|vf(x)-1| &\le \max(|v_yf(y)-1|,|l_yf(y)-1|)+v_y|f(y)-f(x)|\\ &\lesssim a(f(x)\wedge f(y))\biggl(\frac{k}{nf(x)}\biggr)^{\beta/d}+\frac{\log^{1/2}n}{k^{1/2}}. \qquad (2.59)\end{aligned}$$

Now, by the definition of l_x and v_x,

$$\max\bigl\{|p_{n,x,u}-k/(n-1)|\,,\,|p_{n,y,v}-k/(n-1)|\bigr\} \le \frac{3k^{1/2}\log^{1/2}n}{n-1} \qquad (2.60)$$

for all $x, y \in \mathcal{X}$ and $u \in [l_x, v_x], v \in [l_y, v_y]$. Next, we bound $|\frac{n-2}{k} p_\cap - \alpha_z|$ for $x \in \mathcal{X}_n$ and $y = y_{x,z}$ with $\|z\| \leq \{v_x f(x)\}^{1/d} + \{v_y f(x)\}^{1/d}$. First suppose that $u \geq v$. We may write

$$B_x(r_{n,u}) \cap B_y(r_{n,v}) = \{B_x(r_{n,v}) \cap B_y(r_{n,v})\} \cup [\{B_x(r_{n,u}) \setminus B_x(r_{n,v})\} \cap B_y(r_{n,v})],$$

where this is a disjoint union. Writing $I_{a,b}(x) := \int_0^x \mathrm{B}_{a,b}(s)\, ds$ for the regularised incomplete beta function and recalling that μ_d denotes Lebesgue measure on $\mathbb{R}^d$, we have

$$\mu_d\big(B_x(r_{n,v}) \cap B_y(r_{n,v})\big) = V_d r_{n,v}^d I_{\frac{d+1}{2},\frac{1}{2}}\biggl(1 - \frac{\|x-y\|^2}{4r_{n,v}^2}\biggr) = \frac{ve^{\Psi(k)}}{n-1} I_{\frac{d+1}{2},\frac{1}{2}}\biggl(1 - \frac{\|z\|^2}{4\{vf(x)\}^{2/d}}\biggr)$$

and

$$\alpha_z = I_{\frac{d+1}{2},\frac{1}{2}}\biggl(1 - \frac{\|z\|^2}{4}\biggr).$$

Now,

$$\biggl|\frac{d}{dr} I_{\frac{d+1}{2},\frac{1}{2}}\biggl(1 - \frac{r^2}{4}\biggr)\biggr| = \frac{(1-r^2/4)^{\frac{d-1}{2}}}{\mathrm{B}_{(d+1)/2,1/2}} \leq \frac{1}{\mathrm{B}_{(d+1)/2,1/2}}.$$

Hence by the mean value inequality,

$$\biggl|\mu_d\big(B_x(r_{n,v}) \cap B_y(r_{n,v})\big) - \frac{e^{\Psi(k)}\alpha_z}{(n-1)f(x)}\biggr| \leq \frac{e^{\Psi(k)}}{n-1}\biggl[\frac{v\|z\||1 - \{vf(x)\}^{-1/d}|}{\mathrm{B}_{(d+1)/2,1/2}} + \frac{\alpha_z}{f(x)}|1 - vf(x)|\biggr].$$

It follows that for all $x \in \mathcal{X}_n$, $y \in B_x(r_{n,v_x} + r_{n,v_y})$ and $v \in [l_y, v_y]$,

$$\biggl|\int_{B_x(r_{n,v}) \cap B_y(r_{n,v})} f(w)\, dw - \frac{e^{\Psi(k)}\alpha_z}{n-1}\biggr| \lesssim \frac{k}{n} a(f(x) \wedge f(y))\biggl(\frac{k}{nf(x)}\biggr)^{\beta/d} + \frac{k^{1/2}\log^{1/2} n}{n}$$

using (2.59) and Lemma 2.12. We also have by (2.60) that

$$\begin{aligned}\int_{\{B_x(r_{n,u}) \setminus B_x(r_{n,v})\} \cap B_y(r_{n,v})} f(w)\, dw &\leq p_{n,x,u} - p_{n,x,v} \\ &\lesssim \frac{k}{n} a(f(x) \wedge f(y))\biggl(\frac{k}{nf(x)}\biggr)^{\beta/d} + \frac{k^{1/2}\log^{1/2} n}{n}.\end{aligned}$$

Thus, when $x \in \mathcal{X}_n$, $y = y_{x,z} \in B_x(r_{n,v_x} + r_{n,v_y})$, $u \in [l_x, v_x]$, $v \in [l_y, v_y]$ and $u \geq v$,

$$\biggl|\frac{n-2}{k} p_\cap - \alpha_z\biggr| \lesssim a(f(x) \wedge f(y))\biggl(\frac{k}{nf(x)}\biggr)^{\beta/d} + \frac{\log^{1/2} n}{k^{1/2}}. \tag{2.61}$$

We can prove the same bound when $v > u$ similarly, using (2.51), (2.59) and Lemma 2.12. We will also require a lower bound on $p_{n,x,u} + p_{n,y,v} - 2p_\cap$ in the region where $B_x(r_{n,u}) \cap B_y(r_{n,v}) \neq \emptyset$, i.e., $\|z\| \leq \{uf(x)\}^{1/d} + \{vf(x)\}^{1/d}$. By the mean value theorem,

$$1 - I_{\frac{d+1}{2},\frac{1}{2}}(1-\delta^2) \geq 2^{1/2}\delta \max\biggl\{\frac{2^{-d/2}}{\mathrm{B}_{(d+1)/2,1/2}}\, ,\, 1 - I_{\frac{d+1}{2},\frac{1}{2}}(1/2)\biggr\}$$

for all $\delta \in [0,1]$. Thus, for $u \geq v$, with $v \in [l_y, v_y]$, $x \in \mathcal{X}_n$, and $y = y_{x,z}$ with $\|z\| \leq 2\{vf(x)\}^{1/d}$, by (2.59) we have,

$$\begin{aligned}\mu_d\big(B_x(r_{n,u})\cap B_y(r_{n,v})^c\big) &\geq \mu_d\big(B_x(r_{n,v})\cap B_y(r_{n,v})^c\big)\\ &= V_d r_{n,v}^d\biggl\{1-I_{\frac{d+1}{2},\frac{1}{2}}\biggl(1-\frac{\|x-y\|^2}{4r_{n,v}^2}\biggr)\biggr\} \gtrsim \frac{k\|z\|}{nf(x)}.\end{aligned}$$

When $\|z\| > 2\{vf(x)\}^{1/d}$ we simply have $\mu_d\big(B_x(r_{n,v})\cap B_y(r_{n,v})^c\big) = V_d r_{n,v}^d$ and the same overall bound applies. Moreover, the same lower bound for $\mu_d\big(B_y(r_{n,v})\cap B_x(r_{n,u})^c\big)$ holds when $u<v$, $u\in[l_x,v_x]$, $x\in\mathcal{X}_n$, and $y=y_{x,z}\in B_x(r_{n,v_x}+r_{n,v_y})$. We deduce that for all $x\in\mathcal{X}_n$, $y=y_{x,z}\in B_x(r_{n,v_x}+r_{n,v_y})$, $u\in[l_x,v_x]$ and $v\in[l_y,v_y]$,

$$p_{n,x,u}+p_{n,y,v}-2p_\cap \geq \max\{p_{n,x,u}-p_\cap\,,\,p_{n,y,v}-p_\cap\} \gtrsim \frac{k}{n}\|z\|. \tag{2.62}$$

We are now in a position to bound $\mathbb{E}(\|Z_3\|^3)$ above for $x\in\mathcal{X}_n$, $y=y_{x,z}\in B_x(r_{n,v_x}+r_{n,v_y})$, $u\in[l_x,v_x]$, $v\in[l_y,v_y]$. We write

$$\begin{aligned}\mathbb{E}(\|Z_3\|^3) = p_\cap\biggl\|V^{-1/2}\begin{pmatrix}1-p_{n,x,u}\\1-p_{n,y,v}\end{pmatrix}\biggr\|^3 + (p_{n,x,u}-p_\cap)\biggl\|V^{-1/2}\begin{pmatrix}1-p_{n,x,u}\\-p_{n,y,v}\end{pmatrix}\biggr\|^3\\ + (p_{n,y,v}-p_\cap)\biggl\|V^{-1/2}\begin{pmatrix}-p_{n,x,u}\\1-p_{n,y,v}\end{pmatrix}\biggr\|^3 + (1-p_{n,x,u}-p_{n,y,v}+p_\cap)\biggl\|V^{-1/2}\begin{pmatrix}p_{n,x,u}\\p_{n,y,v}\end{pmatrix}\biggr\|^3,\end{aligned} \tag{2.63}$$

and bound each of these terms in turn. First,

$$\begin{aligned}p_\cap\biggl\|V^{-1/2}\begin{pmatrix}1-p_{n,x,u}\\1-p_{n,y,v}\end{pmatrix}\biggr\|^3 &= p_\cap|V|^{-3/2}\{(1-p_{n,x,u})(1-p_{n,y,v})(p_{n,x,u}+p_{n,y,v}-2p_\cap)\}^{3/2}\\ &= p_\cap\biggl\{\frac{(1-p_{n,x,u})(1-p_{n,y,v})}{p_\cap-p_{n,x,u}p_{n,y,v}+\frac{(p_{n,x,u}-p_\cap)(p_{n,y,v}-p_\cap)}{p_{n,x,u}+p_{n,y,v}-2p_\cap}}\biggr\}^{3/2}\\ &\leq p_\cap\min\biggl\{\frac{p_{n,x,u}+p_{n,y,v}}{|V|},\frac{1}{p_\cap-p_{n,x,u}p_{n,y,v}}\biggr\}^{3/2} \lesssim n^{1/2}/k^{1/2},\end{aligned} \tag{2.64}$$

using (2.60) and (2.61), and where we derive the final bound from the left hand side of the minimum if $\|z\|\geq 1$ and the right hand side if $\|z\|<1$. Similarly,

$$(p_{n,x,u}-p_\cap)\biggl\|V^{-1/2}\begin{pmatrix}1-p_{n,x,u}\\-p_{n,y,v}\end{pmatrix}\biggr\|^3 \leq (p_{n,x,u}-p_\cap)p_{n,y,v}^{3/2}|V|^{-3/2} \lesssim \Bigl(\frac{n}{k\|z\|}\Bigr)^{1/2}, \tag{2.65}$$

where we have used (2.62) for the final bound. By symmetry, the same bound holds for the third term on the right-hand side of (2.63). Finally, very similar arguments yield

$$(1-p_{n,x,u}-p_{n,y,v}+p_\cap)\biggl\|V^{-1/2}\begin{pmatrix}p_{n,x,u}\\p_{n,y,v}\end{pmatrix}\biggr\|^3 \lesssim (k/n)^{3/2}. \tag{2.66}$$

Combining (2.64), (2.65) and (2.66) gives (2.58).

Writing $\mathbf{\Phi}_A(\cdot)$ for the measure associated with the $N_2(0,A)$ distribution for invertible A, and ϕ_A for the corresponding density, we have by Pinsker's inequality and a Taylor expansion of the log-determinant function that

$$
\begin{aligned}
2\sup_{C\in\mathcal{C}}|\mathbf{\Phi}_A(C)-\mathbf{\Phi}_B(C)|^2 &\leq \int_{\mathbb{R}^2}\phi_A\log\frac{\phi_A}{\phi_B} \\
&= \frac{1}{2}\{\log|B|-\log|A|+\operatorname{tr}(B^{-1}(A-B))\} \leq \|B^{-1/2}(A-B)B^{-1/2}\|^2,
\end{aligned}
$$

provided $\|B^{-1/2}(A-B)B^{-1/2}\|\leq 1/2$. Hence

$$
\sup_{C\in\mathcal{C}}|\mathbf{\Phi}_A(C)-\mathbf{\Phi}_B(C)| \leq \min\{1, 2\|B^{-1/2}(A-B)B^{-1/2}\|\}.
$$

We now take $A=(n-2)V/k$, $B=\Sigma$ and use the submultiplicativity of the Frobenius norm along with (2.60) and (2.61) and the fact that $\|\Sigma^{-1/2}\| = \{(1+\alpha_z)^{-1}+(1-\alpha_z)^{-1}\}^{1/2}$ to deduce that

$$
\sup_{C\in\mathcal{C}}|\mathbf{\Phi}_A(C)-\mathbf{\Phi}_B(C)| \lesssim \frac{1}{\|z\|}\biggl\{a(f(x)\wedge f(y))\biggl(\frac{k}{nf(x)}\biggr)^{\beta/d}+\frac{\log^{1/2}n}{k^{1/2}}\biggr\} \tag{2.67}
$$

for $x\in\mathcal{X}_n$, $y\in B_x^\circ(r_{n,v_x}+r_{n,v_y})$, $u\in[l_x,v_x]$, $v\in[l_y,v_y]$. Now let $u=f(x)^{-1}(1+k^{-1/2}s)$ and $v=f(x)^{-1}(1+k^{-1/2}t)$. By the mean value theorem, (2.51) and (2.59),

$$
\begin{aligned}
&\biggl|\Phi_\Sigma\biggl(k^{-1/2}\biggl\{(n-2)\mu-\begin{pmatrix}k\\k\end{pmatrix}\biggr\}\biggr)-\Phi_\Sigma(s,t)\biggr| \\
&\qquad\leq \frac{1}{(2\pi)^{1/2}}\biggl\{\biggl|\frac{(n-2)p_{n,x,u}-k}{k^{1/2}}-s\biggr|+\biggl|\frac{(n-2)p_{n,y,v}-k}{k^{1/2}}-t\biggr|\biggr\} \\
&\qquad\lesssim k^{1/2}a(f(x)\wedge f(y))\biggl(\frac{k}{nf(x)}\biggr)^{\beta/d}+k^{-1/2}.
\end{aligned} \tag{2.68}
$$

It follows by (2.57), (2.58), (2.67) and (2.68) that for $x\in\mathcal{X}_n$ and $y\in B_x^\circ(r_{n,v_x}+r_{n,v_y})$,

$$
\begin{aligned}
&\sup_{u\in[l_x,v_x],v\in[l_y,v_y]}|F_{n,x,y}^{(1)}(u,v)-\Phi_\Sigma(s,t)| \\
&\lesssim \min\biggl\{1, \frac{\log^{1/2}n}{k^{1/2}\|z\|}+a(f(x)\wedge(f(y))\biggl(\frac{k}{nf(x)}\biggr)^{\beta/d}\Bigl(k^{1/2}+\frac{1}{\|z\|}\Bigr)\biggr\}.
\end{aligned}
$$

Therefore, by (2.51) and (2.52), and since $f(y)\geq f(x)/2$ for $x\in\mathcal{X}_n$, $y\in B_x(r_{n,v_x}+r_{n,v_y})$ and $n\geq n_0$, we conclude that for each $\epsilon>0$ and $n\geq n_0$

$$
\begin{aligned}
&\biggl|\int_{\mathcal{X}_n\times\mathcal{X}}f(x)f(y)\int_{l_x}^{v_x}\int_{l_y}^{v_y}\frac{F_{n,x}^{(1)}(u,v)-\Phi_\Sigma(s,t)}{uv}\mathbb{1}_{\{\|x-y\|\leq r_{n,u}+r_{n,v}\}}\,du\,dv\,dy\,dx\biggr| \\
&\quad\lesssim \frac{k}{n}\int_{\mathcal{X}_n}f(x)\biggl\{\frac{\log^{1/2}n}{k^{1/2}}+a(f(x)/2)\biggl(\frac{k}{nf(x)}\biggr)^{\beta/d}\biggr\}^2 \\
&\qquad\qquad\int_{B_0(3)}\sup_{u\in[l_x,v_x],v\in[l_{y_{x,z}},v_{y_{x,z}}]}|F_{n,x,y_{x,z}}^{(1)}(u,v)-\Phi_\Sigma(s,t)|\,dz\,dx \\
&\quad= O\biggl(\frac{k}{n}\max\biggl\{\frac{\log^{5/2}n}{k^{3/2}}, \frac{k^{\frac{1}{2}+\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}, \frac{k^{-1/2+\beta/d}\log n}{n^{\beta/d}}, \frac{k^{1/2+2\beta/d}}{n^{2\beta/d}}\biggr\}\biggr).
\end{aligned} \tag{2.69}
$$

By similar (in fact, rather simpler) means we can establish the same bound for the approximation of $G_{n,x,y}$ by $\Phi(k^{1/2}\{uf(x)-1\})\Phi(k^{1/2}\{vf(x)-1\})$.

To conclude the proof for the unweighted case, we write $\mathcal{X}_n = \mathcal{X}_n^{(1)} \cup \mathcal{X}_n^{(2)}$, where

$$\mathcal{X}_n^{(1)} := \{x : f(x) \geq k^{\frac{d}{2\beta}}\delta_n\}\,, \quad \mathcal{X}_n^{(2)} := \{x : \delta_n \leq f(x) < k^{\frac{d}{2\beta}}\delta_n\},$$

and deal with these two regions separately. We have by Slepian's inequality that $\Phi_\Sigma(s,t) \geq \Phi(s)\Phi(t)$ for all s and t. Hence, recalling that $s = s_{x,u} = k^{1/2}\{uf(x)-1\}$ and $t = t_{x,v} = k^{1/2}\{vf(x)-1\}$, by (2.49), (2.51) and (2.59), for every $\epsilon > 0$,

$$\begin{aligned}
&\int_{\mathcal{X}_n^{(2)}\times\mathcal{X}} f(x)f(y)\int_{l_x}^{v_x}\int_{l_y}^{v_y}\frac{\Phi_\Sigma(s,t)-\Phi(s)\Phi(t)}{uv}\mathbb{1}_{\{\|x-y\|\leq r_{n,u}+r_{n,v}\}}du\,dv\,dy\,dx \\
&\quad\leq \frac{e^{\Psi(k)}}{V_d(n-1)k}\int_{\mathcal{X}_n^{(2)}}\int_{\mathbb{R}^d} f(y_{x,z})\frac{\mathbb{1}_{\{\|x-y_{x,z}\|\leq r_{n,v_x}+r_{n,v_{y_{x,z}}}\}}}{f(x)^2 l_x l_{y_{x,z}}} \\
&\qquad\qquad\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\{\Phi_\Sigma(s,t)-\Phi(s)\Phi(t)\}\,ds\,dt\,dz\,dx \\
&\quad\lesssim \frac{1}{n}\int_{\mathcal{X}_n^{(2)}} f(x)\int_{B_0(2)}\alpha_z\,dz\,dx = o\Bigl(\frac{k^{(1+\frac{d}{2\beta})\frac{\alpha}{\alpha+d}-\epsilon}}{n^{1+\frac{\alpha}{\alpha+d}-\epsilon}}\Bigr),
\end{aligned} \tag{2.70}$$

where to obtain the final error term, we have used the fact that $\int_{B_0(2)}\alpha_z\,dz = V_d$. By (2.51) and (2.52) we have, for each $\epsilon > 0$,

$$\begin{aligned}
&\int_{\mathcal{X}_n^{(1)}\times\mathcal{X}} f(x)f(y)\int_{l_x}^{v_x}\int_{l_y}^{v_y}\frac{\Phi_\Sigma(s,t)-\Phi(s)\Phi(t)}{uv}\mathbb{1}_{\{\|x-y\|\leq r_{n,u}+r_{n,v}\}}\,du\,dv\,dy\,dx \\
&\leq \frac{e^{\Psi(k)}}{V_d(n-1)k}\int_{\mathcal{X}_n^{(1)}}\int_{\mathbb{R}^d} f(y_{x,z})\frac{\mathbb{1}_{\{\|x-y_{x,z}\|\leq r_{n,v_x}+r_{n,v_{y_{x,z}}}\}}}{f(x)^2 l_x l_{y_{x,z}}}\alpha_z\,dz\,dx \\
&= \frac{e^{\Psi(k)}}{(n-1)k}\int_{\mathcal{X}_n^{(1)}} f(x)\,dx + O\biggl(\max\biggl\{\frac{\log^{1/2} n}{nk^{1/2}}\,,\,\frac{k^{\beta/d}}{n^{1+\beta/d}}\,,\,\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{1+\frac{\alpha}{\alpha+d}-\epsilon}}\biggr\}\biggr) \\
&= \frac{e^{\Psi(k)}}{(n-1)k} + O\biggl(\max\biggl\{\frac{\log^{1/2} n}{nk^{1/2}}\,,\,\frac{k^{\beta/d}}{n^{1+\beta/d}}\,,\,\frac{k^{(1+\frac{d}{2\beta})\frac{\alpha}{\alpha+d}-\epsilon}}{n^{1+\frac{\alpha}{\alpha+d}-\epsilon}}\biggr\}\biggr).
\end{aligned} \tag{2.71}$$

By Lemma 2.10(ii) as for (2.59) we have, for $x \in \mathcal{X}_n^{(1)}, y \in B_x(r_{n,v_x}+r_{n,v_y})$,

$$\max_{v\in\{v_x,v_y\}}|vf(x)-1-3k^{-1/2}\log^{1/2} n| \lesssim a(f(x)\wedge f(y))\Bigl(\frac{k}{nf(x)}\Bigr)^{\beta/d} = o(k^{-1/2}), \tag{2.72}$$

with similar bounds holding for l_x and l_y. A corresponding lower bound of the same order for the left-hand side of (2.71) follows from (2.72) and the fact that

$$\int_{-2\sqrt{\log n}}^{2\sqrt{\log n}}\int_{-2\sqrt{\log n}}^{2\sqrt{\log n}}\{\Phi_\Sigma(s,t)-\Phi(s)\Phi(t)\}\,ds\,dt = \alpha_z + O(n^{-2})$$

uniformly for $z \in \mathbb{R}^d$. It now follows from (2.56), (2.69), (2.70) and (2.71) that for each $\epsilon > 0$,

$$W_4 = O\biggl(\max\biggl\{\frac{\log^{5/2} n}{nk^{1/2}}\,,\,\frac{k^{\frac{3}{2}+\frac{\alpha-\epsilon}{\alpha+d}}}{n^{1+\frac{\alpha-\epsilon}{\alpha+d}}}\,,\,\frac{k^{3/2+2\beta/d}}{n^{1+2\beta/d}}\,,\,\frac{k^{(1+\frac{d}{2\beta})\frac{\alpha-\epsilon}{\alpha+d}}}{n^{1+\frac{\alpha-\epsilon}{\alpha+d}}}\,,\,\frac{k^{\frac{1}{2}+\frac{\beta}{d}}\log n}{n^{1+\frac{\beta}{d}}}\biggr\}\biggr),$$

as required.

We now turn our attention to the variance of the weighted Kozachenko–Leonenko estimator

$\hat{H}_n^w$. We first claim that

$$\mathrm{Var}\biggl(\sum_{j=1}^k w_j \log \xi_{(j),1}\biggr) = \sum_{j,l=1}^k w_j w_l \,\mathrm{Cov}(\log \xi_{(j),1}, \log \xi_{(l),1}) = V(f) + o(1). \tag{2.73}$$

By (2.18), (2.19) and Lemma 2.3, for j such that $w_j \neq 0$,

$$\mathrm{Var} \log \xi_{(j),1} = V(f) + o(1)$$

as $n \to \infty$. For $l > j$, using similar arguments to those used in the proof of Lemma 2.3, and writing $u_{x,s}^{(k)} := u_{x,s} = V_d(n-1)h_x^{-1}(s)^d e^{-\Psi(k)}$ for clarity, we have

$$\begin{aligned}\mathbb{E}(\log \xi_{(j),1} \log \xi_{(l),1}) &= \int_{\mathcal{X}} f(x) \int_0^1 \int_0^{1-s} \log(u_{x,s}^{(j)}) \log(u_{x,s+t}^{(l)}) \mathrm{B}_{j,l-j,n-l}(s,t)\, dt\, ds\, dx \\ &= \int_{\mathcal{X}} f(x) \int_0^1 \int_0^{1-s} \log\Bigl(\frac{(n-1)s}{f(x)e^{\Psi(j)}}\Bigr) \log\Bigl(\frac{(n-1)(s+t)}{f(x)e^{\Psi(l)}}\Bigr) \mathrm{B}_{j,l-j,n-l}(s,t) dt\, ds\, dx + o(1) \\ &= \int_{\mathcal{X}} f(x) \log^2 f(x)\, dx + o(1)\end{aligned}$$

as $n \to \infty$, uniformly for $1 \leq j < l \leq k_1^*$. Now (2.73) follows on noting that $\sup_{k \geq k_d} \|w\| < \infty$.

Next we claim that

$$\mathrm{Cov}\biggl(\sum_{j=1}^k w_j \log \xi_{(j),1}, \sum_{l=1}^k w_l \log \xi_{(l),2}\biggr) = o(n^{-1}) \tag{2.74}$$

as $n \to \infty$. In view of (2.20) and the fact that $\sup_{k \geq k_d} \|w\| < \infty$, it is sufficient to show that

$$\mathrm{Cov}\bigl(\log(f(X_1)\xi_{(j),1}), \log(f(X_2)\xi_{(l),2})\bigr) = o(n^{-1})$$

as $n \to \infty$, whenever $w_j, w_l \neq 0$. We suppose without loss of generality here that $j < l$, since the $j = l$ case is dealt with in (2.27). We broadly follow the same approach used to bound $W_1, \ldots, W_4$, though we require some new (similar) notation. Let $F'_{n,x,y}$ denote the conditional distribution function of $(\xi_{(j),1}, \xi_{(l),2})$ given $X_1 = x, X_2 = y$ and let $F_{n,x}^{(j)}$ denote the conditional distribution function of $\xi_{(j),1}$ given $X_1 = x$. Let

$$r_{n,u}^{(j)} := \biggl\{\frac{ue^{\Psi(j)}}{V_d(n-1)}\biggr\}^{1/d}, \quad p_{n,x,u}^{(j)} := h_x(r_{n,u}^{(j)}).$$

Recall the definitions of $a_{n,j}^{\pm}$ given in the proof of Lemma 2.13, and let $v_{x,j} := \inf\{u \geq 0 : (n-1)p_{n,x,u}^{(j)} = a_{n,j}^+\}$ and $l_{x,j} := \inf\{u \geq 0 : (n-1)p_{n,x,u}^{(j)} = a_{n,j}^-\}$. For pairs (u,v) with $u \leq v_{x,j}$ and $v \leq v_{y,l}$, let $(M_1, M_2, M_3) \sim \mathrm{Multi}(n-2; p_{n,x,u}^{(j)}, p_{n,y,v}^{(l)}, 1 - p_{n,x,u}^{(j)} - p_{n,y,v}^{(l)})$ and write

$$G'_{n,x,y}(u,v) := \mathbb{P}(M_1 \geq j, M_2 \geq l).$$

Also write

$$\Sigma' := \begin{pmatrix} 1 & (j/l)^{1/2}\alpha'_z \\ (j/l)^{1/2}\alpha'_z & 1 \end{pmatrix},$$

where $\alpha'_z := V_d^{-1}\mu_d\bigl(B_0(1) \cap B_z(\exp(\Psi(l) - \Psi(j))^{1/d})\bigr)$. Writing W'_i for remainder terms to be bounded later, we have

$$\begin{aligned}
&\operatorname{Cov}\bigl(\log(f(X_1)\xi_{(j),1}), \log(f(X_2)\xi_{(l),2})\bigr) \\
&= \int_{\mathcal{X}\times\mathcal{X}} f(x)f(y) \int_{[l_{y,l},v_{y,l}]\times[l_{x,j},v_{x,j}]} h(u,v)\, d(F'_{n,x,y} - F^{(j)}_{n,x}F^{(l)}_{n,y})(u,v)\, dx\, dy + W'_1 \\
&= \int_{\mathcal{X}\times\mathcal{X}} f(x)f(y) \int_{[l_{y,l},v_{y,l}]\times[l_{x,j},v_{x,j}]} h(u,v)\, d(F'_{n,x,y} - G'_{n,x,y})(u,v)\, dx\, dy - \frac{1}{n} + \sum_{i=1}^{2} W'_i \\
&= \int_{\mathcal{X}_n\times\mathcal{X}} f(x)f(y) \int_{l_{y,l}}^{v_{y,l}} \int_{l_{x,j}}^{v_{x,j}} \frac{(F'_{n,x,y} - G'_{n,x,y})(u,v)}{uv}\, du\, dv\, dx\, dy - \frac{1}{n} + \sum_{i=1}^{3} W'_i \\
&= \frac{V_d^{-1}e^{\Psi(j)}}{(n-1)(jl)^{1/2}} \int_{\mathbb{R}^d} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{\Phi_{\Sigma'}(s,t) - \Phi(s)\Phi(t)\}\, ds\, dt\, dz - \frac{1}{n} + \sum_{i=1}^{4} W'_i \\
&= \frac{V_d^{-1}e^{\Psi(j)}}{(n-1)l} \int_{\mathbb{R}^d} \alpha'_z\, dz - \frac{1}{n} + \sum_{i=1}^{4} W_i = O\Bigl(\frac{1}{nk}\Bigr) + \sum_{i=1}^{4} W'_i
\end{aligned} \tag{2.75}$$

as $n \to \infty$. The final equality here follows from the fact that, for Borel measurable sets $K, L \subseteq \mathbb{R}^d$,

$$\int_{\mathbb{R}^d} \mu_d\bigl((K+z)\cap L\bigr)\, dz = \mu_d(K)\mu_d(L), \tag{2.76}$$

so that $\int_{\mathbb{R}^d} \alpha'_z\, dz = V_d e^{\Psi(l)-\Psi(j)}$.

To bound W'_1: Very similar arguments to those used to bound W_1 show that $W'_1 = o(n^{-(9/2-\epsilon)})$ as $n \to \infty$, for every $\epsilon > 0$.

To bound W'_2: Similar to our work used to bound W_2, we may show that

$$\int_{\frac{a^{-}_{n,j}}{n-1}}^{\frac{a^{+}_{n,j}}{n-1}} \int_{\frac{a^{-}_{n,l}}{n-1}}^{\frac{a^{+}_{n,l}}{n-1}} |\mathrm{B}_{j+a,l+b,n-j-l-1}(s,t) - \mathrm{B}_{j+a,n-j}(s)\mathrm{B}_{l+b,n-l}(t)|\, dt\, ds \leq \frac{(jl)^{1/2}}{n}\{1+o(1)\}$$

as $n \to \infty$, for fixed $a, b > -1$. Also,

$$\int_0^1\int_0^{1-s} \log\Bigl(\frac{(n-1)s}{e^{\Psi(j)}}\Bigr)\log\Bigl(\frac{(n-1)t}{e^{\Psi(l)}}\Bigr)\{\mathrm{B}_{j,l,n-j-l-1}(s,t) - \mathrm{B}_{j,n-j}(s)\mathrm{B}_{l,n-l}(t)\}dtds - \frac{1}{n} + O(n^{-2})$$

as $n \to \infty$. Using these facts and very similar arguments to those used to bound W_2 we have for every $\epsilon > 0$ that

$$W'_2 = O\biggl(\frac{k^{1/2}}{n}\max\biggl\{\frac{k^{\beta/d}}{n^{\beta/d}}, \frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}\biggr\}\biggr).$$

To bound W'_3: Similarly to (2.46) and the surrounding work, we can show that for every $\epsilon > 0$,

$$W'_3 = O\biggl(\max\biggl\{\frac{\log n}{nk^{1/2}}, \frac{k^{\frac{1}{2}+\frac{2\beta}{d}}}{n^{1+\frac{2\beta}{d}}}, \frac{k^{\frac{2\alpha}{\alpha+d}-\epsilon}}{n^{\frac{2\alpha}{\alpha+d}-\epsilon}}\biggr\}\biggr).$$

To bound W'_4: Let $(N_1, N_2, N_3, N_4) \sim \mathrm{Multi}(n-2; p^{(j)}_{n,x,u} - p_\cap, p^{(l)}_{n,y,v} - p_\cap, p_\cap, 1 - p^{(j)}_{n,x,u} - p^{(l)}_{n,y,v} + p_\cap)$, where $p_\cap := \int_{B_x(r^{(j)}_{n,u})\cap B_y(r^{(l)}_{n,v})} f(w)\, dw$. Further, let

$$F'^{,(1)}_{n,x,y} := \mathbb{P}(N_1 + N_3 \geq j, N_2 + N_3 \geq l).$$

Then, as in (2.56), we have

$$\begin{aligned}\int_{\mathcal{X}_n\times\mathcal{X}} f(x)f(y)\int_{l_{x,j}}^{v_{x,j}}\int_{l_{y,l}}^{v_{y,l}} \frac{(F'_{n,x,y}-G'_{n,x,y})(u,v)}{uv}\,du\,dv\,dx\,dy \\ = \int_{\mathcal{X}_n\times\mathcal{X}} f(x)f(y)\int_{l_{x,j}}^{v_{x,j}}\int_{l_{y,l}}^{v_{y,l}} \frac{(F^{',(1)}_{n,x,y}-G'_{n,x,y})(u,v)}{uv}\,du\,dv\,dx\,dy \\ + O\biggl(\max\biggl\{\frac{\log n}{nk^{1/2}}, \frac{k^{\frac{1}{2}+\frac{2\beta}{d}}}{n^{1+\frac{2\beta}{d}}}, \frac{k^{\frac{1}{2}+\frac{\alpha}{\alpha+d}-\epsilon}}{n^{1+\frac{\alpha}{\alpha+d}-\epsilon}}\biggr\}\biggr).\end{aligned}$$

We can now approximate $F^{',(1)}_{n,x,y}(u,v)$ by $\Phi_{\Sigma'}(j^{1/2}\{uf(x)-1\}, l^{1/2}\{vf(x)-1\})$ and $G'_{n,x,y}(u,v)$ by $\Phi(j^{1/2}\{uf(x)-1\})\Phi(l^{1/2}\{vf(x)-1\})$. This is rather similar to the corresponding approximation in the bounds on W_4, so we only present the main differences. First, let

$$Y'_i := \begin{pmatrix} \mathbb{1}_{\{X_i\in B_x(r^{(j)}_{n,u})\}} \\ \mathbb{1}_{\{X_i\in B_y(r^{(l)}_{n,v})\}} \end{pmatrix}.$$

We also define

$$\mu' := \mathbb{E}(Y'_i) = \begin{pmatrix} p^{(j)}_{n,x,u} \\ p^{(l)}_{n,y,v} \end{pmatrix}$$

and

$$V' := \mathrm{Cov}(Y'_i) = \begin{pmatrix} p^{(j)}_{n,x,u}(1-p^{(j)}_{n,x,u}) & p_\cap - p^{(j)}_{n,x,u}p^{(l)}_{n,y,v} \\ p_\cap - p^{(j)}_{n,x,u}p^{(l)}_{n,y,v} & p^{(l)}_{n,y,v}(1-p^{(l)}_{n,y,v}) \end{pmatrix},$$

and set $Z'_i := V'^{-1/2}(Y'_i-\mu)$. Our aim is to provide a bound on $p_\cap$. Since the function

$$(r,s)\mapsto \mu_d\bigl(B_0(r^{1/d})\cap B_z(s^{1/d})\bigr),$$

is Lipschitz we have for $x\in\mathcal{X}_n, y = x+f(x)^{-1/d}r^{(j)}_{n,1}z \in B_x(r^{(j)}_{n,v_{x,j}}+r^{(l)}_{n,v_{y,l}}), u\in[l_{x,j},v_{x,j}]$ and $v\in[l_{y,l},v_{y,l}]$ that

$$\biggl|\frac{n-2}{e^{\Psi(j)}}p_\cap - \alpha'_z\biggr| \lesssim a(f(x)\wedge(f(y))\biggl(\frac{k}{nf(x)}\biggr)^{\beta/d} + \frac{\log^{1/2} n}{k^{1/2}}, \tag{2.77}$$

using similar equations to (2.51), (2.52) and (2.59). From this and similar bounds to (2.60), we find that $|V'|\gtrsim k^2/n^2$ and $\|(V')^{-1/2}\|\lesssim(n/k)^{1/2}$. We therefore have

$$\mathbb{E}\|Z'_3\|^3 \le \|(V')^{-1/2}\|^3\mathbb{E}\|Y'_3-\mu'\|^3 \lesssim n^{1/2}/k^{1/2},$$

which is as in the $l=j$ case except with the factor of $\|z\|^{-1/2}$ missing. Note now that

$$\limsup_{n\to\infty}\ \sup_{\substack{(j,l):j<l\\ w_j,w_l\neq 0}}\ \sup_{z\in B_0(1+e^{(\Psi(l)-\Psi(j))/d})} \|(\Sigma')^{-1/2}\| < \infty.$$

Hence, using (2.77), similar bounds to (2.60) and the same arguments as leading up to (2.67),

$$\sup_{C\in\mathcal{C}} |\mathbf{\Phi}_A(C) - \mathbf{\Phi}_B(C)| \lesssim a(f(x)\wedge f(y))\biggl(\frac{k}{nf(x)}\biggr)^{\beta/d} + \frac{\log^{1/2} n}{k^{1/2}}, \tag{2.78}$$

where $B := \Sigma'$ and

$$A := (n-2)\begin{pmatrix} j^{-1}p_{n,x,u}^{(j)}(1-p_{n,x,u}^{(j)}) & j^{-1/2}l^{-1/2}(p_\cap - p_{n,x,u}^{(j)}p_{n,y,v}^{(l)}) \\ j^{-1/2}l^{-1/2}(p_\cap - p_{n,x,u}^{(j)}p_{n,y,v}^{(l)}) & l^{-1}p_{n,y,v}^{(l)}(1-p_{n,y,v}^{(l)}) \end{pmatrix}.$$

Now let $u := f(x)^{-1}(1+j^{-1/2}s)$ and $v := f(x)^{-1}(1+l^{-1/2}t)$. Similarly to (2.68), we have

$$\begin{aligned}&\biggl|\Phi_{\Sigma'}\Bigl(\frac{(n-2)p_{n,x,u}^{(j)}-j}{j^{1/2}},\frac{(n-2)p_{n,y,v}^{(l)}-l}{l^{1/2}}\Bigr) - \Phi_{\Sigma'}(s,t)\biggr| \\ &\qquad\qquad\lesssim k^{1/2}a(f(x)\wedge f(y))\biggl(\frac{k}{nf(x)}\biggr)^{\beta/d} + k^{-1/2}.\end{aligned}$$

Similarly to the arguments leading up to (2.69), it follows that

$$\begin{aligned}&\biggl|\int_{\mathcal{X}_n\times\mathcal{X}} f(x)f(y)\int_{l_{x,j}}^{v_{x,j}}\int_{l_{y,l}}^{v_{y,l}} \frac{F_{n,x}^{',(1)}(u,v)-\Phi_{\Sigma'}(s,t)}{uv}\mathbb{1}_{\{\|x-y\|\leq r_{n,u}^{(j)}+r_{n,v}^{(l)}\}}\,du\,dv\,dy\,dx\biggr| \\ &= O\biggl(\frac{k}{n}\max\biggl\{\frac{\log^{3/2} n}{k^{3/2}}, \frac{k^{\frac{1}{2}+\frac{\alpha}{\alpha+d}-\epsilon}}{n^{\frac{\alpha}{\alpha+d}-\epsilon}}, \frac{k^{-1/2+\beta/d}\log n}{n^{\beta/d}}, \frac{k^{1/2+2\beta/d}}{n^{2\beta/d}}\biggr\}\biggr),\end{aligned}$$

where the power on the first logarithmic factor is smaller because of the absence of the factor of the $\|z\|^{-1}$ term in (2.78). The remainder of the work required to bound W_4' is very similar to the work done from (2.70) to (2.71), using also (2.76), so is omitted. We conclude that

$$W_4' = O\biggl(\max\biggl\{\frac{\log^{\frac{3}{2}} n}{nk^{\frac{1}{2}}}, \frac{k^{\frac{3}{2}+\frac{\alpha-\epsilon}{\alpha+d}}}{n^{1+\frac{\alpha-\epsilon}{\alpha+d}}}, \frac{k^{\frac{3}{2}+\frac{2\beta}{d}}}{n^{1+\frac{2\beta}{d}}}, \frac{k^{(1+\frac{d}{2\beta})\frac{\alpha-\epsilon}{\alpha+d}}}{n^{1+\frac{\alpha-\epsilon}{\alpha+d}}}, \frac{k^{\frac{1}{2}+\frac{\beta}{d}}\log n}{n^{1+\frac{\beta}{d}}}\biggr\}\biggr).$$

The equation (2.75), together the bounds on $W_1',\ldots,W_4'$ just proved, establish the claim (2.74). We finally conclude from (2.73) and (2.74) that

$$\begin{aligned}\mathrm{Var}(\hat{H}_n^w) &= \frac{1}{n}\mathrm{Var}\biggl(\sum_{j=1}^k w_j\log\xi_{(j),1}\biggr) + \biggl(1-\frac{1}{n}\biggr)\mathrm{Cov}\biggl(\sum_{j=1}^k w_j\log\xi_{(j),1}, \sum_{l=1}^k w_l\log\xi_{(l),2}\biggr) \\ &= V(f) + o(n^{-1}),\end{aligned}$$

as required.

2.6.6 Proof of Theorem 2.8

Proof of Theorem 2.8. For the first part of the theorem we aim to apply Theorem 25.21 of van der Vaart (1998), and follow the notation used there. With $\dot{\mathcal{P}} := \{\lambda(\log f + H(f)) : \lambda\in\mathbb{R}\}$ we will first show that the entropy functional H is differentiable at f relative to the tangent set $\dot{\mathcal{P}}$, with efficient influence function $\tilde{\psi}_f = -\log f - H(f)$. Following Example 25.16 in van der Vaart (1998), for $g\in\dot{\mathcal{P}}$, the paths $f_{t,g}$ defined in (2.10) of the main text are differentiable in quadratic mean at $t=0$ with score function g. Note that $\int_{\mathcal{X}} gf = 0$ and $\int_{\mathcal{X}} g^2 f < \infty$ for all $g\in\dot{\mathcal{P}}$. It is convenient to define, for $t \geq 0$, the set $A_t := \{x \in \mathcal{X} : 8t|g(x)| \leq 1\}$, on which we may expand e^{-2tg} easily as a Taylor series. By Hölder's inequality, for $\epsilon \in (0, 1/2)$,

$$\int_{A_t^c} f|\log f| \leq (8t)^{2(1-\epsilon)} \int_{\mathcal{X}} f|g|^{2(1-\epsilon)}|\log f| \leq (8t)^{2(1-\epsilon)} \Big\{\int_{\mathcal{X}} g^2 f\Big\}^{1-\epsilon} \Big\{\int_{\mathcal{X}} f|\log f|^{1/\epsilon}\Big\}^{\epsilon} = o(t)$$

as $t \searrow 0$. Moreover,

$$\int_{A_t^c} f \log(1 + e^{-2tg}) \leq \int_{A_t^c} (\log 2 + 2t|g|)f \leq 16t^2(4\log 2 + 1) \int_{\mathcal{X}} g^2 f.$$

We also have that

$$\begin{aligned} |c(t)^{-1} - 1| &= \Big|\int_{\mathcal{X}} \Big(\frac{2}{1 + e^{-2tg}} - 1 - tg\Big) f\Big| \\ &\leq \int_{A_t} \Big|\frac{e^{-2tg} - 1 + 2tg + tg(e^{-2tg} - 1)}{1 + e^{-2tg}}\Big| f + \int_{A_t^c} (1 + t|g|)f \\ &\leq \frac{16}{3}t^2 \int_{A_t} g^2 f + 72t^2 \int_{A_t^c} g^2 f \leq 72t^2 \int_{\mathcal{X}} g^2 f. \end{aligned} \tag{2.79}$$

It follows that

$$\begin{aligned} &\Big|t^{-1}\{H(f_{t,g}) - H(f)\} + \int_{\mathcal{X}} \{\log f + H(f)\} fg\Big| \\ &= \Big|\frac{1}{t} \int_{\mathcal{X}} \Big\{\Big(1 - \frac{2c(t)}{1+e^{-2tg}}\Big)\log f - \frac{2c(t)}{1+e^{-2tg}} \log\Big(\frac{2c(t)}{1+e^{-2tg}}\Big) + tg(1 + \log f)\Big\} f\Big| \\ &\leq \frac{1}{t} \int_{A_t} f \Big|\{e^{-2tg} - 1 + 2tg + tg(e^{-2tg} - 1)\} \log f - 2\log\Big(\frac{2}{1 + e^{-2tg}}\Big) + tg(1 + e^{-2tg})\Big| + o(1) \\ &\leq \frac{16}{3} t \int_{\mathcal{X}} g^2 f|\log f| + 22t \int_{\mathcal{X}} g^2 f + o(1) \to 0. \end{aligned}$$

The conclusion (2.11) therefore follows from van der Vaart (1998, Theorem 25.21).

We now establish the second part of the theorem. First, by our previous bound on $c(t)$ in (2.79), for $12t < \{\int_{\mathcal{X}} g^2 f\}^{-1/2}$ we have that

$$\|f_{t,g}\|_\infty \leq 2c(t)\|f\|_\infty \leq \frac{2\|f\|_\infty}{1 - 72t^2 \int_{\mathcal{X}} g^2 f} \leq 4\|f\|_\infty,$$

and $\mu_\alpha(f_{t,g}) \leq 4\mu_\alpha(f)$.

We now study the smoothness properties of $f_{t,g}$. This requires some involved calculations, because we first need to understand corresponding properties of g. To this end, for an m times differentiable function $g : \mathbb{R}^d \to \mathbb{R}$, define

$$M_g^*(x) := \max\Big\{\max_{t=1,\ldots,m} \|g^{(t)}(x)\|\,,\ \sup_{y \in B_x^\circ(r_a(x))} \frac{\|g^{(m)}(y) - g^{(m)}(x)\|}{\|y - x\|^{\beta - m}}\Big\}$$

and

$$D_g := \max\Big\{1,\ \sup_{\delta \in (0, \|f\|_\infty)} \frac{\sup_{x : f(x) \geq \delta} M_g^*(x)}{a(\delta)^{m+1}}\Big\}.$$

Let $\mathcal{J}_m$ denote the set of multisets of elements $\{1, \ldots, d\}$ of cardinality at most m, and for $J = \{j_1, \ldots, j_s\} \in \mathcal{J}_m$, define $g_J(x) := \frac{\partial^s g}{\prod_{\ell=1}^s \partial x_\ell}(x)$. Moreover, for $i \in \{1, \ldots, s\}$, let $\mathcal{P}_i(J)$ denote the set of partitions of J into i non-empty multisets. As an illustration, if $d = 2$, then

$$\begin{aligned}\mathcal{J}_3 = \bigl\{&\emptyset, \{1\}, \{2\}, \{1,1\}, \{1,2\}, \{2,1\}, \{2,2\},\\ &\{1,1,1\},\{1,1,2\},\{1,2,1\},\{1,2,2\},\{2,1,1\},\{2,1,2\},\{2,2,1\},\{2,2,2\}\bigr\}.\end{aligned}$$

Moreover, if $J = \{1,1,2\} \in \mathcal{J}_3$, then

$$\mathcal{P}_2(J) = \Bigl\{\bigl\{\{1,1\},\{2\}\bigr\}, \bigl\{\{1,2\},\{1\}\bigr\}, \bigl\{\{1,2\},\{1\}\bigr\}\Bigr\}.$$

Then, by induction, and writing $g^* := g_1 = \log f + H(f)$, it may be shown that

$$g_J^*(x) = \sum_{i=1}^{\mathrm{card}(J)} \frac{(-1)^{i-1}(i-1)!}{f^i} \sum_{\{P_1,\ldots,P_i\}\in\mathcal{P}_i(J)} f_{P_1}\ldots f_{P_i}.$$

Now, the cardinality of $\mathcal{P}_i(J)$ is given by a Stirling's number of the second kind:

$$\mathrm{card}\bigl(\mathcal{P}_i(J)\bigr) = \frac{1}{i!}\sum_{\ell=0}^{i}(-1)^{i-\ell}\binom{i}{\ell}\ell^{\mathrm{card}(J)} =: S\bigl(\mathrm{card}(J), i\bigr),$$

say. Thus, if $\mathrm{card}(J) \leq m$, then

$$|g_J^*(x)| \leq \sum_{i=1}^{\mathrm{card}(J)} (i-1)!S\bigl(\mathrm{card}(J), i\bigr)a(f(x))^i \leq \frac{1}{2}m^{m+1}m!a(f(x))^m. \tag{2.80}$$

Moreover, if $\|y-x\| \leq r_a(x)$ and $m \geq 1$, then

$$\begin{aligned}|g_J^*(y) - g_J^*(x)| \leq \sum_{i=1}^{\mathrm{card}(J)} (i-1)! \sum_{\{P_1,\ldots,P_i\}\in\mathcal{P}_i(J)} &\biggl\{\frac{|f_{P_1}\ldots f_{P_i}(y) - f_{P_1}\ldots f_{P_i}(x)|}{f^i(y)}\\ &+ \frac{|f_{P_1}\ldots f_{P_i}(x)|}{f^i(y)}\biggl|\frac{f^i(y)}{f^i(x)} - 1\biggr|\biggr\}.\end{aligned}$$

Now, by Lemma 2.12,

$$\biggl|\frac{f^i(y)}{f^i(x)} - 1\biggr| \leq i\biggl|\frac{f(y)}{f(x)} - 1\biggr|\biggl(1 + \biggl|\frac{f(y)}{f(x)} - 1\biggr|\biggr)^{i-1} \leq \Bigl(\frac{71}{56}\Bigr)^{i-1} i\biggl|\frac{f(y)}{f(x)} - 1\biggr|.$$

Moreover, by induction and Lemma 2.12 again,

$$|f_{P_1}\ldots f_{P_i}(y) - f_{P_1}\ldots f_{P_i}(x)| \leq 8d^{1/2}\Bigl\{\Bigl(\frac{71}{56}\Bigr)^i - 1\Bigr\}a(f(x))^i f^i(x)\|y-x\|^{\beta-m}.$$

We deduce that (even when $m = 0$),

$$|g_J^*(y) - g_J^*(x)| \leq 8d^{1/2}\Bigl(\frac{71}{41}\Bigr)^m m!(m+1)^{m+2}a(f(x))^{m+1}\|y-x\|^{\beta-m}. \tag{2.81}$$

Comparing (2.80) and (2.81), we see that

$$D_{g^*} \leq 8d^{1/2}\Bigl(\frac{71}{41}\Bigr)^m m!(m+1)^{m+2} =: D. \tag{2.82}$$

Now let $q(y) := (1+e^{-2ty})^{-1}$, so that $f_{t,g}(x) = 2c(t)q\bigl(g(x)\bigr)f(x)$. Similar inductive arguments to those used above yield that when $J \in \mathcal{J}_m$ with $m \geq 1$ and g is m times differentiable,

$$(q \circ g)_J(x) = \sum_{i=1}^{\mathrm{card}(J)} q^{(i)}\bigl(g(x)\bigr) \sum_{\{P_1,\ldots,P_i\}\in\mathcal{P}_i(J)} g_{P_1}\ldots g_{P_i}(x),$$

and we now bound the derivatives of q. By induction,

$$q^{(i)}(y) = (2t)^i \sum_{\ell=1}^{i} (-1)^{i-\ell} \frac{a_\ell^{(i)} e^{-2t\ell y}}{(1+e^{-2ty})^{\ell+1}},$$

where for each $i \in \mathbb{N}$, we have $a_1^{(i)} = 1$, $a_i^{(i)} = i!$ and $a_\ell^{(i)} = \ell(a_\ell^{(i-1)} + a_{\ell-1}^{(i-1)})$ for $\ell \in \{2,\ldots,i-1\}$. Since $\max_{1\leq\ell\leq i} a_\ell^{(i)} \leq (2i)^{i-1}$ (again by induction), we deduce that

$$(1+e^{-2ty})|q^{(i)}(y)| \leq 2^{2i-1} i^i t^i. \tag{2.83}$$

Writing $s := \mathrm{card}(J)$, it follows that

$$\begin{aligned} |(q\circ g)_J(x)| &\leq q\bigl(g(x)\bigr) \sum_{i=1}^{s} 2^{2i-1} i^i t^i S(s,i) a(f(x))^{i(m+1)} D_g^i \\ &\leq q\bigl(g(x)\bigr) s^{s+1} 2^{2s-1} \max(1,t)^s B_s a(f(x))^{s(m+1)} D_g^s, \end{aligned} \tag{2.84}$$

where $B_s := \sum_{i=1}^{s} S(s,i)$ denotes the sth Bell number. We can now apply the multivariate Leibniz rule, so that for a multi-index $\omega = (\omega_1,\ldots,\omega_d)$ with $|\omega| \leq m$, and for $t \leq 1$ and $m \geq 1$,

$$\begin{aligned} \biggl|\frac{\partial^\omega f_{t,g^*}(x)}{\partial x^\omega}\biggr| &= \biggl|2c(t) \sum_{\nu:\nu\leq\omega} \binom{\omega}{\nu} \frac{\partial^\nu q\bigl(g^*(x)\bigr)}{\partial x^\nu} \frac{\partial^{\omega-\nu} f(x)}{\partial x^{\omega-\nu}}\biggr| \\ &\leq 2^{3m-1} m^{m+1} B_m D_{g^*}^m a(f(x))^{m^2+m} f_{t,g^*}(x). \end{aligned} \tag{2.85}$$

Now, in order to control $\bigl|\frac{\partial^\omega f_{t,g^*}(y)}{\partial x^\omega} - \frac{\partial^\omega f_{t,g^*}(x)}{\partial x^\omega}\bigr|$, we first note that by (2.81) and (2.82), we have for $\|y-x\| \leq r_a(x)$, $i \in \mathbb{N}$, $J \in \mathcal{J}_m$ with $\mathrm{card}(J) = s$ and $\{P_1,\ldots,P_i\} \in \mathcal{P}_i(J)$,

$$|g^*_{P_1}\ldots g^*_{P_i}(y) - g^*_{P_1}\ldots g^*_{P_i}(x)| \leq (2D)^i a(f(x))^{i(m+1)} \|y-x\|^{\beta-m}. \tag{2.86}$$

Thus, by (2.83), (2.86), the mean value theorem and Lemma 2.12, for $t \leq 1$, $\|y-x\| \leq r_a(x)$ and $m \geq 1$,

$$\begin{aligned} &|(q\circ g^*)_J(y) - (q\circ g^*)_J(x)| \\ &\leq \biggl|\sum_{i=1}^{s} q^{(i)}(g^*(x)) \sum_{\{P_1,\ldots,P_i\}\in\mathcal{P}_i(J)} \{g^*_{P_1}\ldots g^*_{P_i}(y) - g^*_{P_1}\ldots g^*_{P_i}(x)\}\biggr| \\ &\quad + \biggl|\sum_{i=1}^{s} \{q^{(i)}(g^*(y)) - q^{(i)}(g^*(x))\} \sum_{\{P_1,\ldots,P_i\}\in\mathcal{P}_i(J)} g^*_{P_1}\ldots g^*_{P_i}(y)\biggr| \\ &\leq D^m q(g^*(x)) a(f(x))^{m^2+m+1} \|y-x\|^{\beta-m} \frac{B_m 2^{3m+5} d^{1/2} (m+1)^{m+1} (1+e^{2tg^*(x)})}{e^{2tg^*(x)} + e^{-2t|g^*(y)-g^*(x)|}} \\ &\leq D^m q(g^*(x)) a(f(x))^{m^2+m+1} \|y-x\|^{\beta-m} B_m 2^{3m+5} d^{1/2} (m+1)^{m+1} \Bigl(\frac{56}{41}\Bigr)^{2t}. \end{aligned} \tag{2.87}$$

Using the multivariate Leibnitz rule again, together with (2.84), (2.87) and Lemma 2.12, for $t \leq 1$, $\|y-x\| \leq r_a(x)$ and $|\omega| = m \geq 1$,

$$\begin{aligned}
&\biggl|\frac{\partial^\omega f_{t,g^*}(y)}{\partial x^\omega} - \frac{\partial^\omega f_{t,g^*}(x)}{\partial x^\omega}\biggr| \\
&\leq 2c(t) \sum_{\nu:\nu \leq \omega} \binom{\omega}{\nu} \biggl\{ \biggl|\frac{\partial^{\omega-\nu} f(y)}{\partial y^{\omega-\nu}}\biggr| \biggl|\frac{\partial^\nu q(g^*(y))}{\partial x^\nu} - \frac{\partial^\nu q(g^*(x))}{\partial x^\nu}\biggr| + \biggl|\frac{\partial^\nu q(g^*(x))}{\partial x^\nu}\biggr| \biggl|\frac{\partial^\nu f(y)}{\partial x^\nu} - \frac{\partial^\nu f(x)}{\partial x^\nu}\biggr| \biggr\} \\
&\leq 2^{4m+9} d^{1/2} B_m (m+1)^{m+1} D^m a(f(x))^{m^2+m+1} f_{t,g^*}(x) \|y-x\|^{\beta-m} \\
&=: C'_m D^m a(f(x))^{m^2+m+1} f_{t,g^*}(x) \|y-x\|^{\beta-m}. \qquad (2.88)
\end{aligned}$$

This also holds in the case $m = 0$. Now note that if $12t < \{\int_{\mathcal{X}} (g^*)^2 f\}^{-1/2}$ we have

$$f(x) = \frac{1 + e^{-2tg^*(x)}}{2c(t)} f_{t,g^*}(x) \geq \frac{f_{t,g^*}(x)}{4}.$$

Finally, define the function

$$\tilde{a}(\delta) := d^{m/2} C'_m D^m a(\delta/4)^{m^2+m+1}. \qquad (2.89)$$

Then $\tilde{a} \in \mathcal{A}$ and from (2.85) and (2.88), we have $M_{f_{t,g^*},\tilde{a},\beta}(x) \leq \tilde{a}(f_{t,g^*}(x))$. We conclude that for $t < \min\bigl(1, \{144 \int g^2 f\}^{-1/2}\bigr)$, we have that $f_{t,g^*} \in \mathcal{F}_{d,\theta'}$, where $\theta' = (\alpha, \beta, 4\gamma, 4\nu, \tilde{a}) \in \Theta$. The result follows on noting that $f_{t,g_\lambda} = f_{t\lambda,g^*}$. □

Chapter 3

Tests of independence based on mutual information

3.1 Introduction

Independence is a fundamental concept in statistics and many related fields. The assumption of independence is made in countless statistical models; as a simple example, the linear model $Y = X\beta + \epsilon$ under random design often assumes that X and ϵ are independent. Often we would like to confirm that the assumption of independence is reasonable, as if this assumption is violated then standard theory may not apply. Testing independence and measuring dependence are very well established areas of statistics with the idea of the correlation between two random variables dating back to the end of the 19th century when it was introduced by Francis Galton (Stigler, 1989), and subsequently expanded upon by Pearson. Since then many new measures of dependence have been developed and studied, each with its own advantages and disadvantages, and there is no universally accepted measure. For surveys of well-established measures see, for example, Schweizer (1981), Joe (1989), Mari and Kotz (2001) and the references therein. We give an overview of more recently introduced quantities below; see also Josse and Holmes (2014).

One area in which measuring dependence plays a central role is independent component analysis (ICA), a special case of blind source separation, in which a linear transformation of the data is sought so that the transformed data is maximally independent; see e.g. Comon (1994), Bach and Jordan (2002) and Samworth and Yuan (2012). Independence tests may then be carried out to check the convergence of the ICA algorithm. In many applications the aim is simply to establish whether or not there is dependence between two variables, and tests of independence are required; see Nguyen and Eisenstein (2017) for a recent example in computational linguistics or Steuer et al. (2002) and Albert et al. (2015) and the references therein for biological examples. In addition, the problem of measuring dependence has applications in feature selection (Torkkola, 2003; Song et al., 2012), in which one seeks a set of features which contains the maximum possible information about a response, and in evaluating the quality of a clustering in cluster analysis (Vinh, Epps and Bailey, 2010).

In the contingency table setting where observations are categorical, the testing problem reduces to testing the equality of two discrete distributions and the chi-squared test is commonly used. Here we will focus on the case of continuous random variables. Classical nonparametric approaches to measuring dependence and independence testing in such cases include Pearson's correlation coefficient, Kendall's tau and Spearman's rank correlation coefficient. Though these approaches are widely used they suffer from a lack of power against many alternatives; indeed Pearson correlation measures linear relationships between variables and Kendall's tau and Spearman's rank measure monotonic relationships. Hoeffding's test of independence (Hoeffding, 1948) is able to detect a wider class of departures from independence but, together with these other classical methods, is only applicable in the case of univariate variables. Tests such as Kendall's tau, Spearman's rank and Hoeffding's test in which the test statistic depends on the data only through their rankings have the advantage of being distribution-free, that is the null distribution of the test statistic does not depend on the marginal distributions of the data and critical values can be tabulated in advance. The concept of ranks in the multidimensional setting is less clear, and distribution-free tests are more difficult to construct.

Recent research has focused on constructing tests that can be used for more complex data and that are consistent against wider classes of alternatives. The concept of distance covariance was introduced in Székely, Rizzo and Bakirov (2007) and can be expressed as a weighted L_2 norm between the characteristic function of the joint distribution and the product of the marginal characteristic functions. This concept has also been studied in high-dimensions in Székely and Rizzo (2013). In Sejdinovic et al. (2013) tests based on distance covariance were shown to be equivalent to a reproducing kernel Hilbert space (RKHS) test for a specific choice of kernel. RKHS tests have been widely studied in the machine learning community with early understanding of the subject given by Bach and Jordan (2002) and Gretton et al. (2005), in which the Hilbert–Schmidt independence criterion was proposed. These tests are based on embedding the joint distribution and product of the marginal distributions into a Hilbert space and considering the norm of their difference in this space. One drawback of the kernel paradigm here is the computational complexity, though the recent works Jitkrittum, Szabó and Gretton (2016) and Zhang et al. (2017) attempt to address this issue. The choice of kernel also affects the results in RKHS methods. Other methods include those based on partitioning the sample space; see, for example, Gretton and Györfi (2010) and Heller et al. (2016). These have the advantage of being distribution-free, though partitions of the sample space must be chosen.

We now formalise the independence testing problem considered in this chapter. Let $Z = (X, Y)$ and suppose we observe independent and identically distributed copies $Z_1, \ldots, Z_n$ of Z. The property $X \perp\!\!\!\perp Y$ of independence is often characterised as either the joint distribution function, density function or characteristic function factorising as the product of the corresponding marginal functions.We wish to test the hypotheses

$$H_0 : X \perp\!\!\!\perp Y \quad \text{vs.} \quad H_1 : X \not\!\perp\!\!\!\perp Y.$$

Many related problems have also been studied, such as testing mutual independence between a group of random variables (see e.g. Bai et al. (2009) for the Gaussian case) and testing conditional independence. The concept of conditional independence is particularly useful in graphical modelling (Lauritzen, 1996) and causal inference and there is a large literature on the corresponding conditional independence problem (e.g. Su and White, 2008; Zhang et al., 2011). In Fan, Feng and Xia (2017) the problem of conditional independence testing in graphical models is reduced to independence testing through a linearity assumption and then a distance covariance-based test is used. We will not explicitly consider conditional independence in this chapter except to say that our approach is rather flexible and it is likely some of our work will extend to this setting.

A very natural measure of dependence is given by mutual information, defined between random variables X and Y with joint density f and marginal densities f_X and f_Y by

$$I(X;Y) = I(f) := \int f(x,y) \log \frac{f(x,y)}{f_X(x) f_Y(y)} \, dx \, dy = H(X) + H(Y) - H(X,Y), \tag{3.1}$$

where H denotes differential entropy defined in Chapter 2. This is the Kullback–Leibler divergence between the joint distribution of (X,Y) and the product of the marginal distributions. It is non-negative and equal to zero if and only if X and Y are independent. As noted in Comon (1994) mutual information is very useful in ICA and indeed many methods for fitting ICA models are based on mutual information or approximations thereof. Another attractive feature of mutual information as a measure of dependence is that it is invariant to invertible transformations of X and Y. Indeed, if X takes values in $\mathcal{X} \subseteq \mathbb{R}^p$, and g is a differentiable invertible function on $\mathcal{X}$ then

$$H(g(X)) = H(X) + \mathbb{E} \log |J(X)|,$$

where J is the Jacobian of the transformation $x \mapsto g(x)$. Therefore,

$$\begin{aligned} I(g(X);Y) &= H(X) + \mathbb{E} \log |J(X)| + H(Y) - H(X,Y) - \mathbb{E} \log |J(X)| \\ &= H(X) + H(Y) - H(X,Y) = I(X;Y), \end{aligned}$$

where we used in the above the fact that J is also the Jacobian of the transformation $(x,y) \mapsto (g(x), y)$. Moreover, the concept of mutual information is easily generalised to more complex situations though objects such as the conditional mutual information

$$\begin{aligned} I(X;Y|Z) &:= H(X|Z) + H(Y|Z) - H(X,Y|Z) \\ &= H(X,Z) + H(Y,Z) - H(Z) - H(X,Y,Z) \end{aligned}$$

and the mutual information between p random variables

$$I(X_1; \ldots; X_p) := \sum_{j=1}^{p} H(X_j) - H(X_1, \ldots, X_p). \tag{3.2}$$

These quantities are non-negative and equal to zero if and only if we have conditional independence or mutual independence respectively. They are also expressible purely in terms of differential entropy.

The estimation of mutual information of course plays a crucial role in tests based on this quantity. Many estimators are based on the expansion (3.1) in terms of differential entropy, which allows one to estimate mutual information through entropy estimation. In Miller and Fisher (2003) the authors perform ICA based on (3.2) using entropy estimators based on sample spacings. Recall that the Kozachenko–Leonenko entropy estimator based on a d-dimensional sample $Z_1, \ldots, Z_n$ is given by

$$\hat{H}_n = \hat{H}_n(Z_1, \ldots, Z_n) = \frac{1}{n} \sum_{i=1}^{n} \log \biggl(\frac{V_d (n-1) \rho_{(k),i}^d}{e^{\Psi(k)}} \biggr),$$

where $\rho_{(k),i}$ is the kth nearest neighbour distance for the ith observation. The KSG estimator is a popular estimator of mutual information which is described in Kraskov, Stögbauer and Grassberger (2004). This is based on the Kozachenko–Leonenko estimator but uses a data-driven, local choice of k for the marginal entropy estimation. For simplicity we will instead consider the estimator

$$\hat{I}_n = \hat{I}_n(Z_1, \ldots, Z_n) = \hat{H}_n^X(X_1, \ldots, X_n) + \hat{H}_n^Y(Y_1, \ldots, Y_n) - \hat{H}_n^{XY}(Z_1, \ldots, Z_n), \tag{3.3}$$

where on the right hand side we have (weighted) Kozachenko–Leonenko estimators of $H(X), H(Y)$ and $H(X, Y)$ defined in Chapter 2 with appropriate choices of k (and w). This is similar to the idea for an independence test considered in Goria et al. (2005), though the null distribution for their test statistic is not studied and no test is formally defined.

A common approach to testing when the null distribution of the test statistics is unknown is to use a permutation test. These are a general type of resampling method in which the null distribution is simulated by randomly permuting the data a large number of times and calculating the test statistic for each of these new data sets. As a simple example, when testing for equality of means between two sets of data one would randomly define new sets of data of the same size by sampling without replacement from the pooled data. In this way the resampled data sets will have the same means on average, and the null distribution of the test statistic can be approximated (Romano, 1990). In the context of independence testing with paired data $(X_1, Y_1), \ldots, (X_n, Y_n)$, for a random permutation π one would consider the new data set $(X_1, Y_{\pi(1)}), \ldots, (X_n, Y_{\pi(n)})$ in which $X_i \perp\!\!\!\perp Y_{\pi(i)}$ whenever $\pi(i) \neq i$. In Albert et al. (2015) a permutation test of independence is proposed which is shown to be consistent.

The aim of this chapter is to propose tests of independence based on entropy estimation and to provide a theoretical understanding of these tests. In Section 3.2 we make the assumption that the marginal distributions of X and Y are known and propose a simple test of independence. We show that, under our regularity conditions, the power of our test converges to 1 provided the mutual information is above some threshold that may be $o(n^{-1/2})$ as $n \to \infty$; to the best of our knowledge this is the first time that such a local power analysis has been carried out for an independence test. In Section 3.3 we no longer assume that the marginal distributions are known and formally consider a permutation test. We show that this test is consistent whenever our regularity conditions are satisfied and X and Y are not independent. Again to the best of our knowledge, this is the first study of nearest neighbour methods when some of the components have been permuted. Proofs of our results are presented in Section 3.4.

We now introduce some notation used throughout this chapter. We will denote by f, f_X and f_Y the joint density of (X, Y), the marginal density of X and the marginal density of Y with respect to the appropriate Lebesgue measure, and for given $d_X, d_Y \in \mathbb{N}$ and density f on $\mathbb{R}^{d_X + d_Y}$ we use the convention that

$$f_X(x) = \int_{\mathbb{R}^{d_Y}} f(x, y)\, dy, \quad \text{and} \quad f_Y(y) = \int_{\mathbb{R}^{d_X}} f(x, y)\, dx.$$

For given marginal densities f_X on $\mathbb{R}^{d_X}$ and f_Y and $\mathbb{R}^{d_Y}$ we also define the product density $f_X f_Y$ on $\mathbb{R}^{d_X + d_Y}$ by $f_X f_Y(x, y) = f_X(x) f_Y(y)$. For a density function g we denote by $\mathbb{P}_g(\cdot)$ and $\mathbb{E}_g(\cdot)$ probabilities and expectations respectively when the true underlying joint density of (X, Y) is g.

3.2 A test in the case of known marginal distributions

To define our test formally recall the mutual information estimator $\hat{I}_n$ introduced in (3.3), and write $(k_X, w_X), (k_Y, w_Y)$ and (k_{XY}, w_{XY}) for the tuning parameters selected for $\hat{H}_n^X, \hat{H}_n^Y$ and $\hat{H}_n^{XY}$ respectively. Since $I(X;Y) \geq 0$ with equality if and only if X and Y are independent we will reject the null hypothesis of independence when $\hat{I}_n$ is significantly large. Defining the critical values

$$C_q^{(n)} := \inf\{r \in \mathbb{R} : \mathbb{P}_{f_X f_Y}(\hat{I}_n > r) \leq q\},$$

the test that rejects the null hypothesis if and only if $\hat{I}_n > C_q^{(n)}$ has size at most q. We suppose in this section that the marginal distributions of X and Y are known, and have densities f_X and f_Y respectively. Making this assumption allows us to simulate $\hat{I}_n$ under H_0 and find the critical values of the test to arbitrary precision, so we assume for simplicity that the critical values are known. Observe that, under regularity conditions, the estimators $\hat{H}_n^X, \hat{H}_n^Y$ and $\hat{H}_n^{XY}$ are efficient and under H_0 we then have that

$$\hat{I}_n = -\frac{1}{n}\sum_{i=1}^n [\log f_X(X_i) + \log f_Y(Y_i) - \log\{f_X(X_i) f_Y(Y_i)\}] + o_p(n^{-1/2}) = o_p(n^{-1/2})$$

as $n \to \infty$. The critical values therefore satisfy

$$C_q^{(n)} = n^{-1/2} \inf\{r \in \mathbb{R} : \mathbb{P}_{f_X f_Y}(n^{1/2}\hat{I}_n > r) \leq q\} = o(n^{-1/2})$$

as $n \to \infty$. Now, under regularity conditions and a fixed alternative, writing $V(X;Y) = V(f) := \operatorname{Var}\log\bigl(\frac{f(X,Y)}{f_X(X) f_Y(Y)}\bigr)$, we have, again by the efficiency of the entropy estimators, that

$$\begin{aligned} n^{1/2}(\hat{I}_n - I) &= n^{1/2}\biggl\{-\frac{1}{n}\sum_{i=1}^n \log\biggl(\frac{f_X(X_i) f_Y(Y_i)}{f(X_i, Y_i)}\biggr) - H(X) - H(Y) + H(X,Y)\biggr\} + o_p(1) \\ &\xrightarrow{d} N\bigl(0, V(X;Y)\bigr). \end{aligned}$$

Hence, for a fixed alternative f we have that

$$\mathbb{P}_f(\hat{I}_n \geq C_q^{(n)}) - \bar{\Phi}\biggl(\frac{n^{1/2}C_q^{(n)} - n^{1/2} I(X;Y)}{V^{1/2}(X;Y)}\biggr) \to 0 \tag{3.4}$$

as $n \to \infty$. We will not use this approximation explicitly in our following analysis though it provides some heuristic justification that our test is consistent against alternatives with $I(X;Y)$ greater than $n^{-1/2}$.

The remainder of this section is devoted to a rigorous study of the power of our test that is compatible with a local alternative f_n having mutual information $I_n \to 0$. Recalling the definitions of $\mathcal{F}_{d,\theta}$ and Θ in Section 2.2, for $d_X, d_Y \in \mathbb{N}$ and $\vartheta = (\theta, \theta_X, \theta_Y) \in \Theta^3$, define

$$\mathcal{F}_{d_X, d_Y, \vartheta} := \Bigl\{f \in \mathcal{F}_{d_X + d_Y, \theta} : f_X \in \mathcal{F}_{d_X, \theta_X}, f_Y \in \mathcal{F}_{d_Y, \theta_Y}, f_X f_Y \in \mathcal{F}_{d_X + d_Y, \theta}\Bigr\}$$

and, for $b \geq 0$, let

$$\mathcal{F}_{d_X, d_Y, \vartheta}(b) := \Bigl\{f \in \mathcal{F}_{d_X, d_Y, \vartheta} : I(f) > b\Bigr\}.$$

Given $d \in \mathbb{N}$ and $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$ additionally define

$$\tau_1(d, \theta) = \min\left\{ \frac{2\alpha}{5\alpha + 3d}, \frac{\alpha - d}{2\alpha}, \frac{4(\beta \wedge 1)}{4(\beta \wedge 1) + 3d} \right\}$$

and

$$\tau_2(d, \theta) = \min\left\{ 1 - \frac{d}{2\beta}, 1 - \frac{d/4}{\lfloor d/4 \rfloor + 1} \right\}$$

(cf. Theorem 2.1 in Section 2.2). Note that $\min_{i=1,2} \tau_i(d, \theta) > 0$ exactly when $\alpha > d$ and $\beta > d/2$. The following theorem constitutes our main result on this test and shows that, under regularity conditions, it is consistent against any alternative with sufficiently large mutual information. Recall the definition of $\mathcal{W}^{(k)}$ in Section 2.2.

Theorem 3.1. *Fix $d_X, d_Y \in \mathbb{N}$, set $d_{XY} = d_X + d_Y$ and fix $\vartheta = (\theta_{XY}, \theta_X, \theta_Y) \in \Theta^3$ with*

$$\min_{\mathrm{L} \in \{XY, X, Y\}} \min_{i=1,2} \tau_i(d_{\mathrm{L}}, \theta_{\mathrm{L}}) > 0.$$

Let $k_0^* = k_{0,n}^*, k_X^* = k_{X,n}^*, k_Y^* = k_{Y,n}^*$ and $k_{XY}^* = k_{XY,n}^*$ denote any deterministic sequences of positive integers with $k_0^* \leq \min\{k_X^*, k_Y^*, k_{XY}^*\}$, with $k_0^* / \log^5 n \to \infty$ and with

$$\max_{\mathrm{L} \in \{XY, X, Y\}} \max\left\{ \frac{k_{\mathrm{L}}^*}{n^{\tau_1(d_{\mathrm{L}}, \theta_{\mathrm{L}}) - \epsilon}}, \frac{k_{\mathrm{L}}^*}{n^{\tau_2(d_{\mathrm{L}}, \theta_{\mathrm{L}})}} \right\} \to 0$$

for some $\epsilon > 0$. Also suppose that $w_X = w_X^{(k_X)} \in \mathcal{W}^{(k_X)}, w_Y = w_Y^{(k_Y)} \in \mathcal{W}^{(k_Y)}$ and $w_{XY} = w_{XY}^{(k_{XY})} \in \mathcal{W}^{(k_{XY})}$, and that $\limsup_n \max_{\mathrm{L} \in \{XY, X, Y\}} \|w_{\mathrm{L}}\| < \infty$. Then there exists a sequence (b_n) such that $b_n = o(n^{-1/2})$ and for each $q \in (0, 1)$

$$\inf_{f \in \mathcal{F}_{d_X, d_Y, \vartheta}(b_n)} \mathbb{P}_f(\hat{I}_n > C_q^{(n)}) \to 1$$

uniformly for $k_X \in \{k_0^*, \ldots, k_X^*\}, k_Y \in \{k_0^*, \ldots, k_Y^*\}$ and $k_{XY} \in \{k_0^*, \ldots, k_{XY}^*\}$.

An outline of the proof of Theorem 3.1 is as follows. For $I > C_q^{(n)}$ we have

$$\begin{aligned} \mathbb{P}_f(\hat{I}_n > C_q^{(n)}) &= \mathbb{P}_f(\hat{I}_n - I > C_q^{(n)} - I) \\ &\geq 1 - \mathbb{P}_f(|\hat{I}_n - I| \geq I - C_q^{(n)}) \geq 1 - \frac{\mathbb{E}_f\{(\hat{I}_n - I)^2\}}{(I - C_q^{(n)})^2}. \end{aligned} \tag{3.5}$$

by Markov's inequality. Applying Theorem 2.1 in Section 2.2 we see that

$$\sup_{f \in \mathcal{F}_{d_X, d_Y, \vartheta}} |n\mathbb{E}_f[\{\hat{I}_n - I(f)\}^2] - V(f)| \to 0$$

uniformly for $k_X \in \{k_0^*, \ldots, k_X^*\}, k_Y \in \{k_0^*, \ldots, k_Y^*\}$ and $k_{XY} \in \{k_0^*, \ldots, k_{XY}^*\}$. The next step of the proof is to bound the critical values $C_q^{(n)}$ above, which can be done again using Theorem 2.1 in Section 2.2. We must finally understand the behaviour of $V(X;Y)$, particularly when $I(X;Y)$ is small. It is clear that

$$I(X;Y) = 0 \implies X \perp\!\!\!\perp Y \implies V(X;Y) = 0,$$

but we will require an upper bound on $V(X;Y)$ that vanishes as $I(X;Y) \to 0$. To gain intuition consider the following example. When X and Y are standard univariate normal random variables with $\mathrm{Cov}(X,Y)=\rho$ we have $I(X;Y)=-(1/2)\log(1-\rho^2)$ and $V(X;Y)=\rho^2$, which are asymptotically equivalent, up to a factor of 2, for small ρ. Next consider the following rough calculation based on a Taylor expansion of the exponential function around the origin:

$$\begin{aligned}0&=2\int f_X(x)f_Y(y)\,dx\,dy-2=2\int f(x,y)e^{\log f_X(x)f_Y(y)-\log f(x,y)}\,dx\,dy-2\\&\approx 2\int f(x,y)\biggl\{\log\frac{f_X(x)f_Y(y)}{f(x,y)}+\frac{1}{2}\log^2\frac{f_X(x)f_Y(y)}{f(x,y)}\biggr\}\,dz=V(X;Y)+I(X;Y)^2-2I(X;Y).\end{aligned}$$

This seems to suggest that the above relationship between $I(X;Y)$ and $V(X;Y)$ for bivariate Gaussians is fairly general. The following is a bound, possibly rather conservative, that is strong enough for our purposes.

Lemma 3.2. *Fix $d_X, d_Y \in \mathbb{N}$ and $\vartheta \in \Theta^3$. Then*

$$\sup_{f\in\mathcal{F}_{d_X,d_Y,\vartheta}(0)} I(f)^{-1/4}V(f)<\infty.$$

This has the consequence that the asymptotic distribution of our mutual information estimator is more concentrated about its mean when $I(f)$ is smaller; thus we may detect smaller departures from independence than we might expect from a first glance at (3.4). Formal proofs of Lemma 3.2 and Theorem 3.1 are given in Section 3.4.

3.3 A test in the case of unknown marginal distributions

In this section we consider the, perhaps more realistic, setting in which the marginal distributions of X and Y and the critical values of the previous test $C_q^{(n)}$ are not known. We propose a test similar to the test used in the previous section in which we estimate the critical values by permuting our sample to attempt to mimic the behaviour of the test statistic under H_0. For some large positive integer B simulate $\pi_1,\ldots,\pi_B$ uniformly from S_n, the permutation group of $\{1,\ldots,n\}$, and for $b=1,\ldots,B$ set $Z_i^{(b)}:=(X_i,Y_{\pi_b(i)})$ and also set $\hat{I}_n^{(b)}:=\hat{I}_n(Z_1^{(b)},\ldots,Z_n^{(b)})$. We can now estimate $C_q^{(n)}$ by

$$\hat{C}_q^{(n),B}:=\inf\biggl\{r\in\mathbb{R}:(B+1)^{-1}\sum_{b=0}^{B}\mathbb{1}_{\{\hat{I}_n^{(b)}\geq r\}}\leq q\biggr\},$$

the $(1-q)^{th}$ quantile of $\{\hat{I}_n^{(0)},\ldots,\hat{I}_n^{(B)}\}$, adopting the convention $\hat{I}_n^{(0)}:=\hat{I}_n$. We reject H_0 if and only if $\hat{I}_n>\hat{C}_q^{(n),B}$. The following result controls the size of the test, and follows from the fact that, under H_0, the sequence $(\hat{I}_n^{(0)},\ldots,\hat{I}_n^{(B)})$ is exchangeable.

Theorem 3.3. *We have $\mathbb{P}_{f_Xf_Y}(\hat{I}_n>\hat{C}_q^{(n),B})\leq q$ for any marginal densities f_X and f_Y and $q\in(0,1)$.*

Note that we have $\hat{I}_n>\hat{C}_q^{(n),B}$ if and only if

$$(B+1)^{-1}\sum_{b=0}^{B}\mathbb{1}_{\{\hat{I}_n^{(b)}\geq\hat{I}_n\}}\leq q. \tag{3.6}$$

Thus, by Markov's inequality,

$$\mathbb{P}(\hat{I}_n \leq \hat{C}_q^{(n),B}) \leq \frac{1}{q(B+1)} \sum_{b=0}^{B} \mathbb{P}(\hat{I}_n^{(b)} \geq \hat{I}_n) = \frac{1}{q(B+1)} \{1 + B\mathbb{P}(\hat{I}_n^{(1)} \geq \hat{I}_n)\}$$
$$= \frac{1}{q(B+1)} \{1 + B\mathbb{P}(\hat{H}_n^{XY} \geq \hat{H}_n^{(1)})\},$$

where $\hat{H}_n^{(1)}$ is the (weighted) Kozachenko–Leonkenko estimator applied to $Z_1^{(1)}, \ldots, Z_n^{(1)}$. Taking $B = B_n \to \infty$, we see that it is enough to show that $\mathbb{P}(\hat{H}_n^{XY} \geq \hat{H}_n^{(1)}) \to 0$ under H_1 to prove the consistency of the test. In fact (3.6) shows that estimating the marginal entropies is unnecessary to carry out the test, since $\hat{I}_n^{(b)} - \hat{I}_n = \hat{H}_n^{XY} - \hat{H}_n^{(1)}$.

In the remainder of this section we work under H_1 with a fixed alternative, where X and Y are not independent and we therefore have $I(X;Y) > 0$, and discuss the power of our test. For simplicity we will restrict our attention from this point to test statistics $\hat{I}_n$ based on unweighted entropy estimators, as weighting will be seen to be unnecessary in achieving consistency. Corresponding results for the test based on weighted estimators will hold straightforwardly. Writing

$$\begin{aligned}\mathbb{P}(\hat{H}_n^{XY} \geq \hat{H}_n^{(1)}) &= \mathbb{P}\big(\hat{H}_n^{XY} - H(X,Y) \geq \hat{H}_n^{(1)} - H(X) - H(Y) + I(X;Y)\big) \\ &\leq \mathbb{P}\Big(|\hat{H}_n^{XY} - H(X,Y)| \geq \frac{1}{2} I(X;Y)\Big) + \mathbb{P}\Big(|\hat{H}_n^{(1)} - H(X) - H(Y)| \geq \frac{1}{2} I(X;Y)\Big), \quad (3.7)\end{aligned}$$

we see that it is sufficient to show that $\hat{H}_n^{XY}$ is a consistent estimator of $H(X,Y)$ and $\hat{H}_n^{(1)}$ is a consistent estimator of $H(X) + H(Y)$ under suitable regularity conditions. To ease notation we write $k = k_{XY}$ where this will not cause confusion. Write $\rho_{(k),i,(1)}$ for the distance from $Z_i^{(1)}$ to its k^{th} nearest neighbour in the sample $Z_1^{(1)}, \ldots, Z_n^{(1)}$ and $\xi_i^{(1)} = e^{-\Psi(k)} V_d (n-1) \rho_{(k),i,(1)}^d$ so that

$$\hat{H}_n^{(1)} = \frac{1}{n} \sum_{i=1}^{n} \log \xi_i^{(1)}.$$

We will work with the following conditions.

(A1)(α) We have $\|f_X\|_\infty, \|f_Y\|_\infty < \infty$ and $\int_{\mathcal{X}} \|z\|^\alpha f(z)\, dz < \infty$.

(A2) There exists $\mathcal{X}_n \subset \{z : f_X f_Y(z) > 0\}$ such that

$$\sup_{\delta \in [0,2]} \sup_{z \in \mathcal{X}_n} \left| \frac{1}{V_d r_{z,\delta}^d f_X f_Y(z)} \int_{B_z(r_{z,\delta})} f_X f_Y(w)\, dw - 1 \right| \to 0$$

as $n \to \infty$, where $r_{z,\delta}^d := \frac{\delta e^{\Psi(k)}}{V_d (n-1) f_X f_Y(z)}$.

Our condition **(A2)** ensures that the density $f_X f_Y$ is smooth enough for us to use the approximation $\xi_i^{(1)} f_X f_Y(Z_i^{(1)}) \approx k e^{-\Psi(k)} \approx 1$. This approximation is the basis of such k-nearest neighbour estimators, and in this case, together with **(A1)**(α), will allow us to show that

$$\hat{H}_n^{(1)} = \frac{1}{n} \sum_{i=1}^{n} \log \xi_i^{(1)} \approx -\frac{1}{n} \sum_{i=1}^{n} \log f_X f_Y(Z_i^{(1)}) \approx H(f_X f_Y) = H(f_X) + H(f_Y).$$

The following lemmas formalise this approximation to establish the consistency of $\hat{H}_n^{(1)}$ as an estimator of $H(X) + H(Y)$, and are proved in Section 3.4.

Lemma 3.4. *Suppose that **(A1)**(α) holds for some $\alpha > 0$ and also that **(A2)** holds with $p_n := \int_{\mathcal{X}_n^c} f_X f_Y = o(\log^{-1} n)$. Suppose also that $k/\log^2 n \to \infty$ and $k \log^2 n/n \to 0$ as $n \to \infty$. Then*

$$\mathbb{E}(\hat{H}_n^{(1)}) \to H(X) + H(Y)$$

as $n \to \infty$.

Lemma 3.5. *Suppose that **(A1)**(α) holds for some $\alpha > 0$ and also that **(A2)** holds with $p_n = \int_{\mathcal{X}_n^c} f_X f_Y = o(\log^{-2} n)$. Suppose also that $k/\log^4 n \to \infty$ and $k \log^3 n/n \to 0$ as $n \to \infty$. Then*

$$\operatorname{Var} \hat{H}_n^{(1)} \to 0$$

as $n \to \infty$.

The following analogous conditions on the joint density f are sufficient for the estimator $\hat{H}_n^{XY}$ of $H(X,Y)$ to be consistent. This is much simpler to prove than the consistency of $\hat{H}_n^{(1)}$ for $H(X) + H(Y)$ and we omit a formal lemma in the interests of brevity.

(A3)(α) We have $\|f\|_\infty < \infty$ and $\int_{\mathcal{X}} \|z\|^\alpha f(z)\,dz < \infty$.

(A4) There exists $\mathcal{Y}_n \subset \mathcal{X}$ such that

$$\sup_{\delta \in (0,2]} \sup_{z \in \mathcal{Y}_n} \left| \frac{1}{V_d s_{z,\delta}^d f(z)} \int_{B_z(s_{z,\delta})} f(w)\,dw - 1 \right| \to 0$$

as $n \to \infty$, where $s_{z,\delta}^d := \frac{\delta e^{\Psi(k)}}{V_d(n-1)f(z)}$.

The following result summarises our work on the power of the permutation test against a fixed alternative.

Theorem 3.6. *Let $d_X, d_Y \in \mathbb{N}$ be given and let f be a density function on $\mathbb{R}^{d_X+d_Y}$ satisfying **(A1)**(α) and **(A3)**(α) for some $\alpha > 0$, **(A2)** with $\int_{\mathcal{X}_n^c} f_X f_Y = o(\log^{-2} n)$, **(A4)** with $\int_{\mathcal{Y}_n^c} f = o(\log^{-2} n)$ and $I(f) > 0$. Let $k = k^{XY}$ satisfy $k/\log^4 n \to \infty$ and $k \log^3 n/n \to 0$ and let $B = B_n$ define a sequence of positive integers such that $B \to \infty$ as $n \to \infty$. Then*

$$\mathbb{P}_f(\hat{I}_n > \hat{C}_q^{(n),B}) \to 1,$$

as $n \to \infty$.

This follows from a straightforward combination of (3.7), Lemmas 3.4 and 3.5 and the consistency of $\hat{H}_n^{XY}$ as an estimator of $H(X,Y)$.

3.4 Proofs of main results

Proof of Lemma 3.2. For $x \in \mathbb{R}$ we write $x_- := \max(0, -x)$. First, by Pinkser's inequality,

$$\mathbb{E}\biggl\{\log \frac{f(Z)}{f_X f_Y(Z)}\biggr\}_- = \int_{\{z: f(z) \le f_X f_Y(z)\}} f(z) \log \frac{f_X f_Y(z)}{f(z)}\,dz \le \int_{f \le f_X f_Y} f(z) \biggl\{ \frac{f_X f_Y(z)}{f(z)} - 1 \biggr\}\,dz$$
$$= \sup_A \biggl\{ \int_A f_X f_Y - \int_A f \biggr\} \le \{I(X;Y)/2\}^{1/2}.$$

Thus,

$$\mathbb{E}\Big|\log \frac{f(Z)}{f_X f_Y(Z)}\Big| = I(X;Y) + 2\mathbb{E}\Big\{\log \frac{f(Z)}{f_X f_Y(Z)}\Big\}_- \leq I(X;Y) + \{2I(X;Y)\}^{1/2}. \tag{3.8}$$

We therefore have that

$$\begin{aligned} V(X;Y) &\leq \mathbb{E}\log^2 \frac{f(Z)}{f_X f_Y(Z)} = \mathbb{E}\Big[\Big|\log \frac{f(Z)}{f_X f_Y(Z)}\Big|^{1/2}\Big|\log \frac{f(Z)}{f_X f_Y(Z)}\Big|^{3/2}\Big] \\ &\leq 4\Big\{\mathbb{E}\Big|\log \frac{f(Z)}{f_X f_Y(Z)}\Big|\Big\}^{1/2}\{\mathbb{E}|\log f_X(X)|^3 + \mathbb{E}|\log f_Y(Y)|^3 + \mathbb{E}|\log f(Z)|^3\}^{1/2}]. \end{aligned} \tag{3.9}$$

By Lemma 2.11 in Chapter 2 we may combine (3.8) and (3.9) to conclude that $V(X;Y) = O(I^{1/4})$ as $I \to 0$, uniformly for $f \in \mathcal{F}_{d_X,d_Y,\vartheta}$. The result follows on using Lemma 2.11 in Chapter 2 again to see that

$$\sup_{f \in \mathcal{F}_{d_X,d_Y,\vartheta}} V(f) < \infty, \tag{3.10}$$

so that $I(X;Y)^{-1/4}V(X;Y)$ is also bounded above when $I(X;Y)$ is bounded below. □

Proof of Theorem 3.1. It will be convenient to define the set $K_n := \{k_0^*, \dots, k_X^*\} \times \{k_0^*, \dots, k_Y^*\} \times \{k_0^*, \dots, k_{XY}^*\}$ and to write $\kappa = (k_X, k_Y, k_{XY})$. Then, writing $I_n^* := \frac{1}{n}\sum_{i=1}^n \log \frac{f(Z_i)}{f_X f_Y(Z_i)}$, we have by Theorem 2.1 in Chapter 2 that

$$\sup_{\kappa \in K_n} \sup_{f \in \mathcal{F}_{d_X,d_Y,\vartheta}} n\mathbb{E}_f\{(\hat{I}_n - I_n^*)^2\} \to 0.$$

Thus,

$$\begin{aligned} \epsilon_n^3 &:= \sup_{\kappa \in K_n} \sup_{f \in \mathcal{F}_{d_X,d_Y,\vartheta}} |n\mathbb{E}_f[\{\hat{I}_n - I(f)\}^2] - V(f)| \\ &\leq \sup_{\kappa \in K_n} \sup_{f \in \mathcal{F}_{d_X,d_Y,\vartheta}} \Big\{n\mathbb{E}_f\{(\hat{I}_n - I_n^*)^2\} + 2\big[n\mathbb{E}_f\{(\hat{I}_n - I_n^*)^2\}V(f)\big]^{1/2}\Big\} \to 0, \end{aligned} \tag{3.11}$$

where we use (3.10) to bound $V(f)$ above. We now have, since $f_X f_Y \in \mathcal{F}_{d_X,d_Y,\vartheta}$, that

$$\mathbb{P}_{f_X f_Y}(n^{1/2}\hat{I}_n \geq \epsilon_n) \leq \frac{n\mathbb{E}_{f_X f_Y}\hat{I}_n^2}{\epsilon_n^2} \leq \frac{\sup_{f \in \mathcal{F}_{d_X,d_Y,\vartheta}} |n\mathbb{E}_f[\{\hat{I}_n - I(f)\}^2] - V(f)|}{\epsilon_n^2} \leq \epsilon_n.$$

Choosing $n_0 \in \mathbb{N}$ such that we have $\epsilon_n \leq q$ for $n \geq n_0$ we have that

$$\sup_{\kappa \in K_n} \sup_{f \in \mathcal{F}_{d_X,d_Y,\vartheta}} C_q^{(n)} \leq n^{-1/2}\epsilon_n.$$

for all $n \geq n_0$. Now consider $b_n := \max(2\epsilon_n n^{-1/2}, n^{-4/7}\log n)$. By (3.5) and Lemma 3.2 we have for $n \geq n_0$ that

$$\begin{aligned}\inf_{\kappa \in K_n} \inf_{f \in \mathcal{F}_{d_X,d_Y,\vartheta}(b_n)} \mathbb{P}_f(\hat{I}_n > C_q^{(n)}) &\geq 1 - \sup_{\kappa \in K_n} \sup_{f \in \mathcal{F}_{d_X,d_Y,\vartheta}(b_n)} \frac{\mathbb{E}_f[\{\hat{I}_n - I(f)\}^2]}{\{I(f) - C_q^{(n)}\}^2} \\ &\geq 1 - \sup_{f \in \mathcal{F}_{d_X,d_Y,\vartheta}(b_n)} \frac{4\{V(f) + \epsilon_n^3\}}{nI(f)^2} \\ &\geq 1 - \frac{4}{\log^{7/4} n} \sup_{f \in \mathcal{F}_{d_X,d_Y,\vartheta}(0)} \frac{V(f)}{I(f)^{1/4}} - \epsilon_n \to 1,\end{aligned}$$

as required. □

Proof of Theorem 3.3. We first claim that $(\hat{I}_n^{(0)}, \hat{I}_n^{(1)}, \ldots, \hat{I}_n^{(B)})$ is an exchangeable sequence under H_0. Indeed, let $\sigma \in S_{B+1}$ be arbitrary. Then, since $(X_i, Y_i)_{i=1}^n \overset{d}{=} (X_i, Y_{\tau(i)})_{i=1}^n$ for any $\tau \in S_n$ under H_0, for any Borel set $A \subseteq \mathbb{R}^{B+1}$ we have

$$\begin{aligned}&\mathbb{P}\big((\hat{I}_n^{(\sigma(0))}, \ldots, \hat{I}_n^{(\sigma(B))}) \in A\big) \\ &= \frac{1}{(n!)^B} \sum_{\tau_1,\ldots,\tau_B \in S_n} \mathbb{P}\big((\hat{I}_n^{(\sigma(0))}, \ldots, \hat{I}_n^{(\sigma(B))}) \in A | \pi_1 = \tau_1, \ldots, \pi_B = \tau_B\big) \\ &= \frac{1}{(n!)^B} \sum_{\tau_1,\ldots,\tau_B \in S_n} \mathbb{P}\big((\hat{I}_n^{(0)}, \ldots, \hat{I}_n^{(B)}) \in A | \pi_1 = \tau_{\sigma(1)}\tau_{\sigma(0)}^{-1}, \ldots, \pi_B = \tau_{\sigma(B)}\tau_{\sigma(0)}^{-1}\big) \\ &= \mathbb{P}\big((\hat{I}_n^{(0)}, \ldots, \hat{I}_n^{(B)}) \in A\big).\end{aligned}$$

We now have

$$\mathbb{P}(\hat{I}_n > \hat{C}_q^{(n),B}) \leq \frac{\lfloor q(B+1) \rfloor}{B+1} \leq q,$$

where the first inequality would be an equality if we knew that ties among $\hat{I}_n^{(0)}, \ldots, \hat{I}_n^{(B)}$ had probability zero. □

Proof of Lemma 3.4. Throughout the proof, we write $a \lesssim b$ to mean that there exists $C > 0$, depending only on $d \in \mathbb{N}$ and f, such that $a \leq Cb$. The first step of the proof is to show that we may restrict attention to the event on which the random permutation π_1 does not have too many fixed points. To do this we will need bounds on $\hat{H}_n^{(1)}$ that do not depend on π_1. Writing $\rho_{(k),i,X}$ for the distance from X_i to its k^{th} nearest neighbour in the sample $X_1, \ldots, X_n$ and defining $\rho_{(k),i,Y}$ similarly, we have

$$\max\{\rho_{(k),i,X}^2, \rho_{(k),\pi_1(i),Y}^2\} \leq \rho_{(k),i,(1)}^2 \leq \rho_{(n-1),i,X}^2 + \rho_{(n-1),\pi_1(i),Y}^2.$$

Using the fact that $0 \leq \log(a+b) \leq \log 2 + |\log a| + |\log b|$ for $a, b > 0$ such that $a + b \geq 1$, we have that

$$\begin{aligned}|\log \xi_i^{(1)}| &\leq \left|\log\left(\frac{V_d(n-1)}{e^{\Psi(k)}}\right)\right| + d|\log \rho_{(k),i,X}| + \frac{d}{2}\log\left(\frac{\rho_{(k),i,(1)}^2}{\rho_{(k),i,X}^2}\right) \\ &\leq \left|\log\left(\frac{V_d(n-1)}{e^{\Psi(k)}}\right)\right| + 3d|\log \rho_{(k),i,X}| + \frac{d}{2}\log 2 + d|\log \rho_{(n-1),i,X}| + d|\log \rho_{(n-1),\pi_1(i),Y}| \\ &\leq \left|\log\left(\frac{V_d(n-1)}{e^{\Psi(k)}}\right)\right| + \frac{d}{2}\log 2 + 5d \max_{j=1,\ldots,n} \max\{-\log \rho_{(1),j,X}\,,\, \log \rho_{(n-1),j,X}\,,\, \log \rho_{(n-1),j,Y}\}.\end{aligned} \tag{3.12}$$

By the triangle inequality and Markov's inequality we have that

$$\begin{aligned}
\mathbb{E}\log\Bigl\{\max_{i=1,\ldots,n}\rho_{(n-1),i,X}\Bigr\}-\log 2 &\le \mathbb{E}\log\Bigl(\max_{i=1,\ldots,n}\|X_i\|\Bigr)\le \mathbb{E}\Bigl\{\log\Bigl(\max_{i=1,\ldots,n}\|X_i\|\Bigr)\Bigr\}_+\\
&= \int_0^\infty \mathbb{P}\Bigl(\max_{i=1,\ldots,n}\|X_i\|\ge e^M\Bigr)\,dM\\
&\le \alpha^{-1}\{\log n+\log\mathbb{E}(\|X_1\|^\alpha)\}+\alpha^{-1}n\mathbb{E}(\|X_1\|^\alpha)\exp(-\log n-\log\mathbb{E}(\|X_1\|^\alpha))\\
&\le \alpha^{-1}\{\log n+\log\mathbb{E}(\|Z_1\|^\alpha)+1\}, \qquad (3.13)
\end{aligned}$$

and the same bound holds for $\mathbb{E}\log\rho_{(n-1),i,Y}$. Similarly, for $n\ge(V_d\|f_X\|_\infty)^{1/3}$,

$$\begin{aligned}
\mathbb{E}\Bigl\{\log\min_{i=1,\ldots,n}\rho_{(1),i,X}\Bigr\}_- &\le \int_0^\infty \mathbb{P}\Bigl(\min_{i=1,\ldots,n}\rho_{(1),i,X}\le e^{-M}\Bigr)\,dM\\
&\le 3d^{-1}\log n+n\int_{3d^{-1}\log n}^\infty\{1-(1-V_d\|f_X\|_\infty e^{-Md})^{n-1}\}\,dM\\
&\le 3d^{-1}\log n+n(n-1)V_d\|f_X\|_\infty\int_{3d^{-1}\log n}^\infty e^{-Md}\,dM\\
&= 3d^{-1}\log n+d^{-1}n^{-2}(n-1)V_d\|f_X\|_\infty. \qquad (3.14)
\end{aligned}$$

Define $S_n^l\subset S_n$ to be the set of permutations with exactly l fixed points. Then

$$\begin{aligned}
\mathbb{P}(\pi_1\in\cup_{l=l_n+1}^n S_n^l) &\le \frac{1}{n!}\binom{n}{l_n+1}|\{\pi\in S_n:\pi(1)=1,\ldots,\pi(l_n+1)=l_n+1\}|\\
&= \frac{1}{n!}\binom{n}{l_n+1}(n-l_n-1)!=\frac{1}{(l_n+1)!}\sim\frac{1}{\sqrt{2\pi(l_n+1)}}\Bigl(\frac{e}{l_n+1}\Bigr)^{l_n+1}
\end{aligned}$$

by Stirling's approximation. Thus, using (3.12), (3.13), (3.14) and choosing $l_n=\lfloor\log\log n\rfloor$ so that $l_n\log l_n/\log\log n\to\infty$, we have

$$|\mathbb{E}(\hat{H}_n^{(1)}\mathbb{1}_{\{\pi_1\in\cup_{l=l_n+1}^n S_n^l\}})|\lesssim\frac{\log n}{(l_n+1)!}\to 0$$

as $n\to\infty$.

The next step is to show that, when π_1 has fewer than l_n fixed points, the dominant contribution to $\hat{H}_n^{(1)}$ comes from those i such that $\pi_1(i)\ne i$ and the k nearest neighbours of $Z_i^{(1)}$ are among the $Z_j^{(1)}$ such that $\pi_1(j)\ne j$. Suppose that $\pi_1=\pi\in S_n^l$ and, without loss of generality, suppose that $1,\ldots,l$ are the fixed points. We use an argument that involves covering $\mathbb{R}^d$ by cones; cf. Section 20.7 in Biau and Devroye (2015). Define the cone

$$\mathcal{C}(z,\theta)=\bigl\{w\in\mathbb{R}^d\setminus\{0\}:\cos^{-1}(z^Tw/(\|z\|\|w\|))\le\theta\bigr\}\cup\{0\}.$$

There exists a constant $C_{\pi/6}<\infty$ depending only on d such that we may cover $\mathbb{R}^d$ by $C_{\pi/6}$ cones of angle $\pi/6$ centred at $Z_1^{(1)}$. In each cone, mark the k nearest points to $Z_1^{(1)}$ among $Z_2^{(1)},\ldots,Z_n^{(1)}$. Now consider a point $Z_i^{(1)}$ that is not marked, and let $Z_{i_1}^{(1)},\ldots,Z_{i_k}^{(1)}$ be the k marked points in a cone containing $Z_i^{(1)}$. By Lemma 20.5 of Biau and Devroye (2015) we have, for each $j=1,\ldots,k$, that

$$\|Z_i^{(1)}-Z_{i_j}^{(1)}\|<\|Z_i^{(1)}-Z_1^{(1)}\|.$$

Thus, the unmarked $Z_i^{(1)}$ is not one of the k nearest neighbours of $Z_1^{(1)}$, and only the marked points, of which there are at most $kC_{\pi/6}$, may have $Z_1^{(1)}$ as one of their k nearest neighbours. This immediately generalises to show that at most $klC_{\pi/6}$ of the points $Z_{l+1}^{(1)}, \ldots, Z_n^{(1)}$ may have one of $Z_1^{(1)}, \ldots, Z_l^{(1)}$ among their k nearest neighbours. Thus using (3.12), (3.13) and (3.14) and defining the event $A_i' := \{$The k nearest neighbours of $Z_i^{(1)}$ are among $Z_{l+1}^{(1)}, \ldots, Z_n^{(1)}\}$, we have that

$$\left|\mathbb{E}(\hat{H}_n^{(1)}|\pi_1 = \pi) - \mathbb{E}\biggl(\frac{1}{n}\sum_{i=l+1}^n \log \xi_i^{(1)} \mathbb{1}_{A_i'}\biggm|\pi_1 = \pi\biggr)\right| \lesssim \frac{kl_n}{n}\log n.$$

By (3.12), Lemma 2.10 as in (2.13) and Hölder's inequality we have for $i \geq l+1$ that

$$\mathbb{E}\bigl(|\log \xi_i^{(1)}|\mathbb{1}_{\{Z_i^{(1)} \in \mathcal{X}_n^c\}}|\pi_1 = \pi\bigr) \lesssim p_n \log n + p_n^{1-\epsilon}$$

for each $\epsilon > 0$. It therefore suffices to consider

$$\mathbb{E}\biggl(\frac{1}{n}\sum_{i=l+1}^n \log \xi_i^{(1)} \mathbb{1}_{A_i}\biggm|\pi_1 = \pi\biggr),$$

where $A_i := A_i' \cap \{Z_i^{(1)} \in \mathcal{X}_n\}$.

Now, as above,

$$\begin{aligned}
&\left|\mathbb{E}\biggl(\frac{1}{n}\sum_{i=l+1}^n \log \xi_i^{(1)} \mathbb{1}_{A_i}\biggm|\pi_1 = \pi\biggr) - H(X) - H(Y)\right| \\
&= \left|\mathbb{E}\biggl(\frac{1}{n}\sum_{i=l+1}^n \log \xi_i^{(1)} \mathbb{1}_{A_i}\biggm|\pi_1 = \pi\biggr) + \mathbb{E}\bigl(\log f_X(X_1)f_Y(Y_2)\bigr)\right| \\
&\lesssim \left|\mathbb{E}\biggl(\frac{1}{n}\sum_{i=l+1}^n \log\bigl(\xi_i^{(1)} f_X f_Y(Z_i^{(1)})\bigr) \mathbb{1}_{A_i}\biggm|\pi_1 = \pi\biggr)\right| + \frac{kl_n}{n} + p_n \log n + p_n^{1-\epsilon},
\end{aligned}$$

for each $\epsilon > 0$, and the remainder of the proof is devoted to studying, on the event A_i, the convergence of $\xi_i^{(1)} f_X f_Y(Z_i^{(1)})$ to 1. We again work on the event $\pi_1 = \pi \in S_n^l$ and assume that the fixed points are $1, \ldots, l$. Write $\mathbb{P}_\pi(\cdot) := \mathbb{P}(\cdot|\pi_1 = \pi)$. Recalling the definition of $r_{z,\delta}$ in **(A2)** we define the random variables

$$B_i^\delta := \sum_{\substack{l<j\leq n \\ j\neq i}} \mathbb{1}_{\{\|Z_j^{(1)} - Z_i^{(1)}\| \leq r_{Z_i^{(1)},\delta}\}}.$$

For $0 < \epsilon < 1$ we have

$$\begin{aligned}
\mathbb{P}_\pi(\{\xi_i^{(1)} f_X f_Y(Z_i^{(1)}) - 1\}\mathbb{1}_{A_i} \geq \epsilon) &= \mathbb{P}_\pi(A_i, \{\rho_{(k),i,(1)} \geq r_{Z_i^{(1)},1+\epsilon}\}) \leq \mathbb{P}_\pi(B_i^{1+\epsilon} \leq k, Z_i^{(1)} \in \mathcal{X}_n) \\
&= \int_{\mathcal{X}_n} f_X f_Y(z)\mathbb{P}_\pi(B_i^{1+\epsilon} \leq k|Z_i^{(1)} = z)\, dz.
\end{aligned}$$

To bound the above we study the bias and the variance of B_i^δ. We have by **(A2)** that

$$\begin{aligned}
\mathbb{E}_\pi(B_i^{1+\epsilon}|Z_i^{(1)} = z) &\geq (n - l_n - 3)\int_{B_z(r_{z,1+\epsilon})} f_X f_Y(w)\, dw \\
&= (n - l_n - 3)V_d r_{z,1+\epsilon}^d f_X f_Y(z)\{1 + o(1)\} = k(1+\epsilon)\{1 + o(1)\}
\end{aligned}$$

uniformly for $\epsilon \in (-1,1)$ and $z \in \mathcal{X}_n$. Similarly,

$$\mathbb{E}_\pi(B_i^{1+\epsilon}|Z_i^{(1)} = z) \leq 2 + (n - l_n - 3)\int_{B_z(r_{z,1+\epsilon})} f_X f_Y(w)\,dw = k(1+\epsilon)\{1+o(1)\}$$

uniformly for $\epsilon \in (-1,1)$ and $z \in \mathcal{X}_n$. Note that if $j_2 \notin \{j_1, \pi(j_1), \pi^{-1}(j_1)\}$ then

$$\mathrm{Cov}_\pi(\mathbb{1}_{\{\|Z_{j_1}^{(1)}-z\| \leq r_{z,1+\epsilon}\}}, \mathbb{1}_{\{\|Z_{j_2}^{(1)}-z\| \leq r_{z,1+\epsilon}\}}|Z_i^{(1)} = z) = 0.$$

Also, for $j \notin \{i, \pi(i), \pi^{-1}(i)\}$ we have

$$\mathrm{Var}_\pi(\mathbb{1}_{\{\|Z_j^{(1)}-z\| \leq r_{z,1+\epsilon}\}}|Z_i^{(1)} = z) \leq \int_{B_z(r_{z,1+\epsilon})} f_X f_Y(w)\,dw = \frac{(1+\epsilon)k}{n}\{1+o(1)\}$$

uniformly for $\epsilon \in (-1,1)$ and $z \in \mathcal{X}_n$. When $j =\in \{i, \pi(i), \pi^{-1}(i)\}$ we simply bound the variance above by 1 so that, using the Cauchy–Schwarz inequality,

$$\mathrm{Var}_\pi(B_i^{1+\epsilon}|Z_i^{(1)}) \leq (n - l_n - 3)\frac{3(1+\epsilon)k}{n}\{1+o(1)\} + 4 = 3(1+\epsilon)k\{1+o(1)\}.$$

We have now shown that, given $\epsilon \in (0,1)$, there exists n_0 such that for $n \geq n_0$ we have $\mathbb{E}_\pi(B_i^{1+\epsilon}|Z_i^{(1)} = z) > k$ for all $z \in \mathcal{X}_n$ and

$$\begin{aligned}\mathbb{P}_\pi(\{\xi_i^{(1)} f_X f_Y(Z_i^{(1)}) - 1\}\mathbb{1}_{A_i} \geq \epsilon) &\leq \int_{\mathcal{X}_n} f_X f_Y(z) \frac{\mathrm{Var}_\pi(B_i^{1+\epsilon}|Z_i^{(1)} = z)}{\{\mathbb{E}_\pi(B_i^{1+\epsilon}|Z_i^{(1)} = z) - k\}^2}\,dz \\ &\leq \frac{3(1+\epsilon)}{\epsilon^2}\frac{1}{k}\{1+o(1)\} \to 0 \qquad (3.15)\end{aligned}$$

as $n \to \infty$. Using very similar arguments and increasing n_0 if necessary we also have for $\epsilon \in (0,1)$ that

$$\begin{aligned}&\mathbb{P}_\pi(\{\xi_i^{(1)} f_X f_Y(Z_i^{(1)}) - 1\}\mathbb{1}_{A_i} \leq -\epsilon) \leq \mathbb{P}_\pi(B_i^{1-\epsilon} \geq k, Z_i^{(1)} \in \mathcal{X}_n) \\ &\leq \int_{\mathcal{X}_n} f_X f_Y(z) \frac{\mathrm{Var}_\pi(B_i^{1-\epsilon}|Z_i^{(1)} = z)}{\{k - \mathbb{E}_\pi(B_i^{1-\epsilon}|Z_i^{(1)} = z)\}^2}\,dz \leq \frac{3(1-\epsilon)}{\epsilon^2}\frac{1}{k}\{1+o(1)\} \to 0\end{aligned}$$

as $n \to \infty$.

We have now established that $\log(\xi_i^{(1)} f_X f_Y(Z_i^{(1)}))\mathbb{1}_{A_i}|\pi_1 = \pi \xrightarrow{p} 0$, and our aim now is to show that these random variables are bounded in $\mathcal{L}_2$ and so uniformly integrable, so we have convergence in $\mathcal{L}_1$. First, by Markov's inequality, for $k \geq 3$ and $\epsilon \in (0,2]$,

$$\mathbb{P}_\pi(\xi_i^{(1)} f_X f_Y(Z_i^{(1)}) \leq \epsilon, A_i) \leq \frac{n-l-3}{k-2}\int_{\mathcal{X}_n} f_X f_Y(z) \int_{B_z(r_{z,\epsilon})} f_X f_Y(w)\,dw\,dz = \epsilon\{1+o(1)\}.$$

Now, by (3.15),

$$\mathbb{P}_\pi(\xi_i^{(1)} f_X f_Y(Z_i^{(1)}) \geq 1 + k^{-1/2}\log n, A_i) = O(\log^{-2} n)$$

as $n \to \infty$. Also, by Markov's inequality applied twice,

$$\begin{aligned}
&\mathbb{P}_\pi(\xi_i^{(1)} f_X f_Y(Z_i^{(1)}) \geq M, A_i) \leq \mathbb{P}_\pi(B_i^M \leq k, Z_i^{(1)} \in \mathcal{X}_n) \\
&\leq \frac{n-l_n-1}{n-l_n-1-k} \int_{\mathcal{X}_n} f_X f_Y(z) \mathbb{P}_\pi(\|Z-z\| > r_{z,M})\, dz + \frac{2}{n-l_n-1-k} \\
&\leq \frac{n-l_n-1}{n-l_n-1-k} \max\{1, 2^{\alpha-1}\} \int_{\mathcal{X}_n} f_X f_Y(z) \frac{\mathbb{E}\|Z\|^\alpha + \|z\|^\alpha}{r_{z,M}^\alpha}\, dz + \frac{2}{n-l_n-1-k} \lesssim \Big(\frac{n}{kM}\Big)^{\alpha/d}.
\end{aligned}$$

We now integrate by parts to see that, writing g and G for the density and distribution function respectively of $\xi_i^{(1)} f_X f_Y(Z_i^{(1)})|\pi_1 = \pi$ on the event A_i,

$$\begin{aligned}
&\mathbb{E}_\pi(\log^2(\xi_i^{(1)} f_X f_Y(Z_i^{(1)}))\mathbb{1}_{A_i}) = \bigg\{\int_0^1 + \int_1^{1+k^{-1/2}\log n} + \int_{1+k^{-1/2}\log n}^{\frac{n\log^{2d/\alpha} n}{k}} + \int_{\frac{n\log^{2d/\alpha} n}{k}}^\infty \bigg\} \log^2 x\, g(x)\, dx \\
&\leq \int_0^1 \frac{-2\log x}{x} G(x)\, dx + \log^2(1+k^{-1/2}\log n) + \log^2\bigg(\frac{n\log^{2d/\alpha} n}{k}\bigg)\bar{G}(1+k^{-1/2}\log n) \\
&\quad + 2\log^2\bigg(\frac{n\log^{2d/\alpha} n}{k}\bigg)\bar{G}\bigg(\frac{n\log^{2d/\alpha} n}{k}\bigg) + \int_{\frac{n\log^{2d/\alpha} n}{k}}^\infty \frac{2\log x\, \bar{G}(x)}{x}\, dx = O(1)
\end{aligned}$$

as $n \to \infty$. We have now established the required uniform integrability. Since all of our bounds are uniform in $\pi \in \cup_{l=1}^{l_n} S_n^l$ we have now shown that

$$\begin{aligned}
&\bigg|\mathbb{E}\bigg(\frac{1}{n}\sum_{i=l+1}^n \log\big(\xi_i^{(1)} f_X f_Y(Z_i^{(1)})\big)\mathbb{1}_{A_i}|\pi_1 = \pi\bigg)\bigg| \\
&\leq \frac{n-l_n}{n} \max_{1\leq l\leq l_n} \sup_{\pi \in S_n^l} \mathbb{E}(|\log(\xi_{l+1,(1)} f_X f_Y(Z_{l+1}^{(1)}))|\mathbb{1}_{A_{l+1}}|\pi_1 = \pi) \to 0
\end{aligned}$$

as $n \to \infty$. This concludes the proof. □

Proof of Lemma 3.5. We start by writing

$$\operatorname{Var} \hat{H}_n^{(1)} = \frac{1}{n}\operatorname{Var}\log\xi_{1,(1)} + (1-n^{-1})\operatorname{Cov}(\log\xi_{1,(1)}\, ,\, \log\xi_{2,(1)}). \tag{3.16}$$

We have, similarly to in the proof of Lemma 3.4, that

$$\mathbb{E}\log^2\xi_{1,(1)} \to \mathbb{E}\log^2(f_X(X_1)f_Y(Y_2)),$$

and so

$$\frac{1}{n}\operatorname{Var}\log\xi_{1,(1)} = \frac{1}{n}\operatorname{Var}\log(f_X(X_1)f_Y(Y_2))\{1+o(1)\} = O(n^{-1})$$

as $n \to \infty$. It is now sufficient to consider the covariance term in (3.16). Using Cauchy–Schwarz we write

$$\begin{aligned}
&|\operatorname{Cov}(\log\xi_{1,(1)}\, ,\, \log\xi_{2,(1)})| \\
&= |\operatorname{Cov}\big(\log(\xi_{1,(1)} f_X f_Y(Z_1^{(1)})) - \log f_X f_Y(Z_1^{(1)})\, ,\, \log(\xi_{2,(1)} f_X f_Y(Z_2^{(1)})) - \log f_X f_Y(Z_2^{(1)})\big)| \\
&\leq \operatorname{Var}\log(\xi_{1,(1)} f_X f_Y(Z_1^{(1)})) + 2\big\{\operatorname{Var}\log(\xi_{1,(1)} f_X f_Y(Z_1^{(1)}))\operatorname{Var}\log f_X f_Y(Z_1^{(1)})\big\}^{1/2} \\
&\qquad + \operatorname{Cov}(\log f_X f_Y(Z_1^{(1)})\, ,\, \log f_X f_Y(Z_2^{(1)}))
\end{aligned}$$

and deal with each of these three terms separately. Firstly, again by similar methods to those used in the proof of Lemma 3.4, we have that

$$\operatorname{Var}\log(\xi_{1,(1)}f_Xf_Y(Z_1^{(1)})) \leq \mathbb{E}\log^2(\xi_{1,(1)}f_Xf_Y(Z_1^{(1)})) \to 0$$

as $n \to \infty$, and so the first term vanishes. Now note that, by **(A1)**(α),

$$\mathbb{E}\log^2 f_Xf_Y(Z_1^{(1)}) \leq 2\{\mathbb{E}\log^2 f_X(X) + \mathbb{E}\log^2 f_Y(Y)\} < \infty, \tag{3.17}$$

and so we also have that the second term vanishes as $n \to \infty$. Now,

$$\begin{aligned}\mathbb{P}(\{\pi_1(1) = 2\} \cup \{\pi_1(2) = 1\}) &= \mathbb{P}(\pi_1(1) = 2) + \mathbb{P}(\pi_1(2) = 1) - \mathbb{P}(\pi_1(1) = 2, \pi_1(2) = 1) \\ &= 2n^{-1} - n^{-1}(n-1)^{-1} = O(n^{-1}),\end{aligned}$$

and on the complementary event $Z_1^{(1)}$ and $Z_2^{(1)}$ are independent. Hence, using (3.17),

$$\operatorname{Cov}(\log f_Xf_Y(Z_1^{(1)})\, , \log f_Xf_Y(Z_2^{(1)})) = O(1/n)$$

as $n \to \infty$, and the result follows. □

Chapter 4

Local nearest neighbour classification with applications to semi-supervised learning

4.1 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, is arguably the simplest and most intuitive nonparametric classifier. It was introduced in the seminal work of Fix and Hodges (1951), later republished as Fix and Hodges (1989), and early understanding of some of its theoretical properties was provided in Cover and Hart (1967), Duda and Hart (1973) and Stone (1977). 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 (1995), Audibert and Tsybakov (2007), Hall et al. (2008), Biau, Cérou and Guyader (2010), Samworth (2012), Chaudhuri and Dasgupta (2014) and Celisse and Mary-Huard (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 is computed only over a compact subset of $\mathbb{R}^d$. Unfortunately, such restrictions are typically imposed purely for mathematical convenience, and leave open the question of the effect of tail behaviour on the excess risk.

The first goal of this chapter, therefore, is to provide a new asymptotic expansion for the global excess risk of a k-nearest neighbour classifier (Theorem 4.1), where we allow the feature vectors to have unbounded support. Our expansion elucidates conditions under which the dominant contribution to the excess risk comes from the locus of points at which each class label is equally likely to occur, but we also show that if these conditions are not satisfied, the dominant contribution may arise from the tails of the marginal distribution of the features.

The proof of Theorem 4.1 also reveals a local bias-variance trade-off that motivates a modification of the standard k-nearest neighbour classifier in semi-supervised classification settings, where, in addition to the labelled training data, we have access to a further, 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 frequently 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 to depend on an estimate of $\bar{f}$ at the test point in semi-supervised settings. By using fewer neighbours in low density regions, we are able to achieve a better balance in the local bias-variance trade-off. In particular, we initially study an oracle, local choice of k that depends on $\bar{f}$, and under regularity conditions, we show that the excess risk over $\mathbb{R}^d$ is $O(n^{-4/(d+4)})$ provided that the feature vectors have $\rho > 4$ finite moments. By contrast, our theory for the standard k-nearest neighbour classifier with a global choice of k requires that $d \geq 5$ and the feature vectors have $\rho > 4d/(d-4)$ finite moments. 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.

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 et al. (2016) recently obtained upper bounds on the supremum excess risk of the k-nearest neighbour classifier, when η is Lipschitz, the well-known margin assumption of Mammen and Tsybakov (1999) is satisfied, and a tail condition on the rate of decay of $\mathbb{P}\{\bar{f}(X) < \delta\}$ as $\delta \searrow 0$ is imposed. They also derived minimax lower bounds (in general, of different order) in the same problem. Our assumptions and conclusions are not directly comparable, but allow us to obtain the same rates of convergence as in situations where the marginal distribution of X is compactly supported and bounded away from zero on its support, as well as to provide the leading constants in the asymptotic expansion for the excess risk in such cases.

The remainder of this chapter is organised as follows. After introducing our setting in Section 4.2, we present in Section 4.3 our main results for the standard k-nearest neighbour classifier. This leads on, in Section 4.4, to our study of the semi-supervised setting, where we derive asymptotic results of the excess risk of our local k-nearest neighbour classifier. We illustrate the finite-sample benefits of the semi-supervised classifier over the standard k-nearest neighbour classifier in a simulation study in Section 4.5. Proofs are given in Section 4.6, while in section 4.7 we present an introduction to the ideas of differential geometry that underpin much of our analysis.

Finally, we fix here some notation used throughout this chapter. 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 $\ddot{g}(x) = \big(g_{jk}(x)\big)$ for its gradient vector and Hessian matrix at x, and let $\|g\|_\infty = \sup_{x \in A} |g(x)|$. Let $\|\cdot\|_{\mathrm{op}}$ denote the operator norm of a matrix.

4.2 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 $\mathbb{R}^d \times \{1, 2\}$. Let $\pi_r := \mathbb{P}(Y = r)$, for $r = 1, 2$, and $X|Y = r \sim P_r$, for $r = 1, 2$, where P_r is a probability measure on $\mathbb{R}^d$. Let $\eta(x) := \mathbb{P}(Y = 1|X = x)$ and let $P_X := \pi_1 P_1 + \pi_2 P_2$ 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 1 or 2.

A *classifier* is a Borel measurable function $C : \mathbb{R}^d \to \{1, 2\}$, 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

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

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

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

minimises the risk over any region $\mathcal{R}$ (Devroye et al., 1996, p. 20). Thus, the performance of a classifier C is measured via its (non-negative) *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

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

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

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

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 upper and lower bounds are only slightly stronger requirement than the consistency conditions of Stone (1977), namely that $k = k_n \to \infty$, $k_n/n \to 0$ as $n \to \infty$.

4.3 Global risk of the k-nearest neighbour classifier

Our aim in this section is to provide an asymptotic expansion for the global risk of the standard (non-local) k-nearest neighbour classifier. Our analysis will make use of the following assumptions:

(A.1) The probability measures P_1 and P_2 are absolutely continuous with respect to Lebesgue measure, with Radon–Nikodym derivatives f_1 and f_2, respectively. Moreover, the marginal density of X, given by $\bar{f} := \pi_1 f_1 + \pi_2 f_2$, is continuous P_X-almost everywhere and $\mathcal{X}_{\bar{f}} := \{x \in \mathbb{R}^d : \bar{f} \text{ is continuous at } x\}$ is open.

Let $\mathcal{S} := \{x \in \mathbb{R}^d : \eta(x) = 1/2\}$ and, for $\epsilon > 0$, let $\mathcal{S}^\epsilon := \mathcal{S} + B_\epsilon(0)$.

(A.2) The set $\mathcal{S} \cap \{x \in \mathbb{R}^d : \bar{f}(x) > 0\}$ is non-empty and $\bar{f}$ is bounded on $\mathcal{S}$. There exist $\epsilon_0 > 0$ and a measurable function $g : \mathcal{S} \to [1, \infty)$ with the property that $\bar{f}$ is twice continuously differentiable on $\mathcal{S}^{\epsilon_0}$, and

$$\max\biggl\{ \|\dot{\bar{f}}(x_0)\|, \sup_{u \in B_{\epsilon_0}(0)} \|\ddot{\bar{f}}(x_0 + u)\|_{\mathrm{op}} \biggr\} \leq \bar{f}(x_0) g(x_0), \tag{4.1}$$

for all $x_0 \in \mathcal{S}$, where $\sup_{x_0 \in \mathcal{S}: \bar{f}(x_0) \geq \delta} g(x_0) = o(\delta^{-\tau})$, as $\delta \searrow 0$, for each $\tau > 0$. Furthermore, writing $p_\epsilon(x) := P_X\bigl(B_\epsilon(x)\bigr)$, there exists $\mu_0 \in (0, a_d)$ such that, for all $x \in \mathbb{R}^d \setminus \mathcal{S}^{\epsilon_0}$ and $\epsilon \in (0, \epsilon_0]$, we have

$$p_\epsilon(x) \geq \mu_0 \epsilon^d \bar{f}(x).$$

(A.3) We have $\inf_{x_0 \in \mathcal{S}} \|\dot{\eta}(x_0)\| > 0$, so that $\mathcal{S}$ is a $(d-1)$-dimensional, orientable manifold (cf. Section 4.7.3). Moreover, $\sup_{x \in \mathcal{S}^{2\epsilon_0}} \|\dot{\eta}(x)\| < \infty$ and $\ddot{\eta}$ is uniformly continuous on $\mathcal{S}^{2\epsilon_0}$ with $\sup_{x \in \mathcal{S}^{2\epsilon_0}} \|\ddot{\eta}(x)\|_{\mathrm{op}} < \infty$. Finally, the function η is continuous on $\{x : \bar{f}(x) > 0\}$, and for every $\tau > 0$,

$$\sup_{x \in \mathbb{R}^d \setminus \mathcal{S}^{\epsilon_0} : \bar{f}(x) \geq \delta} |\eta(x) - 1/2|^{-1} = o(\delta^{-\tau})$$

as $\delta \searrow 0$.

(A.4)(ρ) We have that $\int_{\mathbb{R}^d} \|x\|^\rho dP_X(x), \int_{\mathcal{S}} \bar{f}(x_0)^{d/(\rho+d)} d\mathrm{Vol}^{d-1}(x_0) < \infty$, where $d\mathrm{Vol}^{d-1}$ denotes the $(d-1)$-dimensional volume form on $\mathcal{S}$ (cf. Section 4.7.3).

The density assumption in **(A.1)** allows us to define the tail of the distribution as the region where $\bar{f}$ is smaller than some threshold. The second and third parts of **(A.1)** ensure 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 $P_X(\mathcal{R}^c) \leq \mathbb{P}\bigl\{\bar{f}(X) \leq \delta\bigr\}$, where the latter quantity can be bounded straightforwardly using **(A.4)(ρ)**. 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 allows us still to control the error of a Taylor approximation even in this region. Moreover, (4.1) is satisfied by all Gaussian and multivariate-t densities, for example. The second 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 et al., 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 ask for derivatives of $\bar{f}$ in this region.

The first part of **(A.3)** asks for the class conditional densities, when weighted by their respective prior probabilities, to cross at an angle, while the bounds on the first and second derivatives of η ensure that we can estimate η sufficiently well. The last part of this condition asks that η does not approach the critical value of $1/2$ too fast on the complement of $\mathcal{S}^{\epsilon_0}$. Finally, the first condition of **(A.4)**(ρ) is a simple moment condition, while the second ensures the constants B_1 and B_2 in (4.2) below are finite.

Let

$$B_1 := \int_{\mathcal{S}} \frac{\bar{f}(x_0)}{4\|\dot{\eta}(x_0)\|}\, d\mathrm{Vol}^{d-1}(x_0) \text{ and } B_2 := \int_{\mathcal{S}} \frac{\bar{f}(x_0)^{1-4/d}}{\|\dot{\eta}(x_0)\|} a(x_0)^2 \, d\mathrm{Vol}^{d-1}(x_0), \tag{4.2}$$

where

$$a(x) := \frac{\sum_{j=1}^{d}\{\eta_j(x)\bar{f}_j(x) + \frac{1}{2}\eta_{jj}(x)\bar{f}(x)\}}{(d+2)a_d^{2/d}\bar{f}(x)}. \tag{4.3}$$

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

Theorem 4.1. *Assume* ***(A.1)***, ***(A.2)***, ***(A.3)*** *and* ***(A.4)***(ρ).

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

$$R(\hat{C}_n^{k\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) = \frac{B_1}{k} + B_2\Big(\frac{k}{n}\Big)^{4/d} + o\bigg(\frac{1}{k} + \Big(\frac{k}{n}\Big)^{4/d}\bigg)$$

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

$$R(\hat{C}_n^{k\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) = \frac{B_1}{k} + o\Big(\frac{1}{k} + \Big(\frac{k}{n}\Big)^{\frac{\rho}{\rho+d}-\epsilon}\Big)$$

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

Theorem 4.1 reveals an interesting dichotomy: we see from part (i) that, when $d \geq 5$ and **(A.4)**(ρ) holds for sufficiently large ρ (and the other regularity conditions hold), 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) suggests that when either $d \leq 4$ or $d \geq 5$ and we only know that **(A.4)**(ρ) holds for small ρ, 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 4.1 below provides one simple setting where this dominant contribution does come from the tails of the distribution.

The proof of Theorem 4.1, and indeed those of Theorems 4.2 and 4.3 which follow in Section 4.4 below, depend crucially on Theorem 4.4 in Section 4.6. 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 (4.8) below, shrinks to zero at a rate slow enough to ensure that $X_{(k)}(x)$ concentrates around x uniformly over $\mathcal{R}_n$. This enables us to derive asymptotic expansions for the bias and variance of $\hat{S}_n(x)$, uniformly over $\mathcal{R}_n$, and using a normal approximation, we can deduce an asymptotic expansion for the excess risk, uniformly over the relevant set of nearest neighbour classifiers. Having proved Theorem 4.4, the proof of Theorem 4.1 is completed by controlling the remainder terms in Theorem 4.4 appropriately, and bounding $P_X(\mathcal{R}_n^c)$ using **(A.4)**(ρ).

Example 4.1. 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 **(A.1)**, **(A.2)**, **(A.3)** hold, and **(A.4)**(ρ) holds for every $\rho > 0$. We prove in Section 4.6.3 that for every $\beta \in (0, 1/2)$ and $\epsilon > 0$,

$$\liminf_{n\to\infty} \inf_{k \in K_\beta} \Big(\frac{n}{k}\Big)^{1+\epsilon} \big\{R(\hat{C}_n^{k\mathrm{nn}}) - R(C^{\mathrm{Bayes}})\big\} > 0 \tag{4.4}$$

as $n \to \infty$.

4.4 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 4.1. Initially, we consider an oracle choice, where k is allowed to depend on the marginal feature density $\bar{f}$ (Section 4.4.1), but we then relax this to semi-supervised settings, where $\bar{f}$ can be estimated from unlabelled training data (Section 4.4.2).

4.4.1 Oracle classifier

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

$$k_{\mathrm{O}}(x) := \max\Big[\lceil (n-1)^\beta \rceil,\, \min\big\{\big\lfloor B\{\bar{f}(x)(n-1)\}^{4/(d+4)}\big\rfloor,\, \lfloor (n-1)^{1-\beta} \rfloor\big\}\Big], \tag{4.5}$$

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 4.2. *Assume* ***(A.1)***, ***(A.2)***, ***(A.3)*** *and* ***(A.4)(ρ)***. *Then for each* $0 < B_* \leq B^* < \infty$,
(i) if $\rho > 4$ *then for* $\beta < 4d(\rho - 4)/\{\rho(d+4)^2\}$,

$$R(\hat{C}_n^{k_{\mathrm{O}}\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) = B_3 n^{-4/(d+4)}\{1 + o(1)\},$$

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

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

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

$$R(\hat{C}_n^{k_{\mathrm{O}}\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) = o(n^{-\rho/(\rho+d)+\beta+\epsilon}),$$

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

Comparing Theorem 4.2(i) and Theorem 4.1(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 4.2(ii) are also stronger than those provided by Theorem 4.1(ii) for any global choice of k. Examining the proof of Theorem 4.2, we find that the key difference with the proof of Theorem 4.1 is that we can now choose the region $\mathcal{R}_n$ (cf. the discussion following the statement of Theorem 4.1) to be larger.

4.4.2 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

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

Motivated by the oracle local choice of k in (4.5), for $\beta \in (0, 1/2)$ and $B > 0$, let

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

For $\gamma \in (0, 2]$, we will consider the following condition:

(A.5)(γ**)** We have that $\bar{f}$ is bounded and, if $\gamma > 1$, then $\bar{f}$ is differentiable on $\mathbb{R}^d$; moreover, there exists $\lambda > 0$ such that

$$\begin{aligned} \|\bar{f}(y) - \bar{f}(x)\| &\leq \lambda\|y - x\|^{\gamma} \quad \text{for all } x, y \in \mathbb{R}^d, \text{ if } \gamma \in (0, 1] \\ \|\dot{\bar{f}}(y) - \dot{\bar{f}}(x)\| &\leq \lambda\|y - x\|^{\gamma-1} \text{ for all } x, y \in \mathbb{R}^d, \text{ if } \gamma \in (1, 2]. \end{aligned}$$

Theorem 4.3. *Assume* ***(A.1)***, ***(A.2)***, ***(A.3)***, ***(A.4)(***ρ***)*** *and* ***(A.5)(***γ***)*** *for some* $\gamma \in (0, 2]$*. Let* $m_0 > 0$*, let* $0 < A_* \leq A^* < \infty$ *and* $0 < B_* \leq B^* < \infty$*, and let* $h = h_m := Am^{-1/(d+2\gamma)}$ *for some* $A > 0$*.*

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

$$R(\hat{C}_n^{k_{\mathrm{SS}}\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) = B_3 n^{-4/(d+4)}\{1 + o(1)\}$$

uniformly for $A \in [A_*, A^*]$*,* $B \in [B_*, B^*]$ *and* $m = m_n \geq m_0(n-1)^{2+d/\gamma}$*, where* B_3 *was defined in Theorem 4.2(i).*

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

$$R(\hat{C}_n^{k_{\mathrm{SS}}\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) = o(n^{-\rho/(\rho+d)+\beta+\epsilon}),$$

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

Examination of the proof of Theorem 4.3 reveals that the same conclusion holds for any estimator $\tilde{f}_m$ of $\bar{f}$ constructed from $\mathcal{T}'_m$, provided there exists $\alpha > (1 + d/4)\beta$ such that

$$\mathbb{P}\biggl(\|\tilde{f}_m - \bar{f}\|_\infty \geq \frac{1}{(n-1)^{1-\alpha/2}}\biggr) = o(n^{-4/(d+4)}). \tag{4.6}$$

Condition $\mathbf{(A.5)}(\gamma)$ ensures that (4.6) holds for our kernel density estimator.

4.5 Empirical analysis

In this section, we compare the k_{O}nn and k_{SS}nn classifiers, introduced in Section 4.4 above, with the standard $k$nn classifier studied in Section 4.3. We investigate three settings that reflect the differences between the main results in Sections 4.3 and 4.4.

- Setting 1: P_1 is the distribution of d independent $N(0,1)$ components; whereas P_2 is the distribution of d independent $N(1, 1/4)$ components.
- Setting 2: P_1 is the distribution of d independent t_5 components; P_2 is the distribution of d independent components, the first $\lfloor d/2 \rfloor$ having a t_5 distribution and the remaining $d - \lfloor d/2 \rfloor$ having a $N(1,1)$ distribution.
- Setting 3: P_1 is the distribution of d independent standard Cauchy components; P_2 is the distribution of d independent components, the first $\lfloor d/2 \rfloor$ being standard Cauchy and the remaining $d - \lfloor d/2 \rfloor$ standard normal.

The corresponding marginal distribution P_X in Setting 1 has all moments finite. Hence, for the standard k-nearest neighbour classifier when $d \geq 5$, we are in the setting of Theorem 4.1(i), while for $d \leq 4$, we can only appeal to Theorem 4.1(ii). On the other hand, for the local k-nearest neighbour classifiers, the results of Theorems 4.2(i) and 4.3(i) apply for all dimensions, and we can expect the excess risk to converge to zero at rate $O(n^{-4/(d+4)})$. In Setting 2, $\mathbf{(A.4)}(\rho)$ holds for $\rho < 5$, but not for $\rho \geq 5$. Thus, for the standard k-nearest neighbour classifier, we are in the setting of Theorem 4.1(ii) for $d < 20$, whereas Theorems 4.2(i) and 4.3(i) again apply for all dimensions for the local classifiers. Finally, in Setting 3, $\mathbf{(A.4)}(\rho)$ does not hold for any $\rho \geq 1$, and only the conditions of Theorems 4.1(ii), 4.2(ii) and 4.3(ii) apply.

For the standard $k$nn classifier, we use 5-fold cross validation to choose k, based on a sequence of equally-spaced values between 1 and $\lfloor n/4 \rfloor$ of length at most 40. For the oracle classifier, we set

$$\hat{k}_{\mathrm{O}}(x) := \max\Bigl[1, \min\bigl[\lfloor \hat{B}_{\mathrm{O}}\{\bar{f}(x)n/\|\bar{f}\|_\infty\}^{4/(d+4)} \rfloor, n/2\bigr]\Bigr],$$

where $\hat{B}_{\mathrm{O}}$ was again chosen via 5-fold cross validation, but based on a sequence of 40 equally-spaced points between $n^{-4/(d+4)}$ (corresponding to the 1-nearest neighbour classifier) and $n^{d/(d+4)}$. Similarly, for the semi-supervised classifier, we set

$$\hat{k}_{\mathrm{SS}}(x) := \max\Bigl[1, \min\bigl[\lfloor \hat{B}_{\mathrm{SS}}\{\hat{f}_m(x)n/\|\hat{f}_m\|_\infty\}^{4/(d+4)} \rfloor, n/2\bigr]\Bigr],$$

where $\hat{B}_{\mathrm{SS}}$ was chosen analogously to $\hat{B}_{\mathrm{O}}$, and where $\hat{f}_m$ is the d-dimensional kernel density estimator constructed using a truncated normal kernel and bandwidths chosen via the default method in

Table 4.1: Misclassification rates for Settings 1, 2 and 3. In the final two columns we present the regret ratios given in (4.7) (with standard errors calculated via the delta method).

d	Bayes risk	n	$\hat{k}$nn risk	$\hat{k}_{\mathrm{O}}$nn risk	$\hat{k}_{\mathrm{SS}}$nn risk	O RR	SS RR
Setting 1							
1	22.67	50	$26.85_{0.13}$	$25.91_{0.12}$	$25.98_{0.13}$	$0.78_{0.022}$	$0.79_{0.023}$
		200	$24.07_{0.06}$	$23.52_{0.06}$	$23.48_{0.05}$	$0.61_{0.030}$	$0.58_{0.029}$
		1000	$23.20_{0.04}$	$22.93_{0.04}$	$22.94_{0.04}$	$0.48_{0.048}$	$0.50_{0.048}$
2	13.30	50	$17.70_{0.09}$	$16.96_{0.08}$	$16.95_{0.08}$	$0.83_{0.015}$	$0.83_{0.015}$
		200	$15.09_{0.05}$	$14.69_{0.04}$	$14.74_{0.05}$	$0.77_{0.018}$	$0.80_{0.019}$
		1000	$14.04_{0.04}$	$13.78_{0.03}$	$13.80_{0.03}$	$0.65_{0.025}$	$0.67_{0.025}$
5	3.53	50	$9.46_{0.07}$	$8.95_{0.06}$	$8.94_{0.06}$	$0.91_{0.006}$	$0.91_{0.006}$
		200	$6.94_{0.03}$	$6.67_{0.03}$	$6.70_{0.03}$	$0.92_{0.006}$	$0.93_{0.007}$
		1000	$5.49_{0.02}$	$5.18_{0.02}$	$5.23_{0.02}$	$0.84_{0.008}$	$0.87_{0.008}$
Setting 2							
1	31.16	50	$36.55_{0.14}$	$36.07_{0.14}$	$35.93_{0.14}$	$0.91_{0.020}$	$0.88_{0.020}$
		200	$32.93_{0.08}$	$32.38_{0.07}$	$32.42_{0.07}$	$0.69_{0.031}$	$0.71_{0.032}$
		1000	$31.62_{0.05}$	$31.37_{0.05}$	$31.37_{0.05}$	$0.46_{0.065}$	$0.47_{0.066}$
2	31.15	50	$37.79_{0.13}$	$38.02_{0.12}$	$37.90_{0.12}$	$1.02_{0.014}$	$1.01_{0.015}$
		200	$33.64_{0.08}$	$33.63_{0.07}$	$33.54_{0.07}$	$1.00_{0.028}$	$0.96_{0.026}$
		1000	$31.83_{0.05}$	$31.81_{0.05}$	$31.80_{0.05}$	$0.97_{0.039}$	$0.95_{0.038}$
5	20.10	50	$28.74_{0.12}$	$29.16_{0.12}$	$29.13_{0.11}$	$1.05_{0.011}$	$1.05_{0.011}$
		200	$23.60_{0.06}$	$23.75_{0.06}$	$23.93_{0.06}$	$1.04_{0.014}$	$1.09_{0.015}$
		1000	$21.86_{0.04}$	$21.71_{0.04}$	$21.77_{0.04}$	$0.91_{0.014}$	$0.95_{0.014}$
Setting 3							
1	37.44	50	$44.76_{0.10}$	$43.09_{0.12}$	$43.08_{0.12}$	$0.77_{0.013}$	$0.77_{0.013}$
		200	$41.86_{0.08}$	$40.18_{0.09}$	$40.23_{0.09}$	$0.62_{0.017}$	$0.63_{0.017}$
		1000	$38.68_{0.06}$	$37.85_{0.05}$	$37.89_{0.05}$	$0.33_{0.033}$	$0.36_{0.032}$
2	37.45	50	$46.20_{0.09}$	$44.81_{0.10}$	$45.24_{0.10}$	$0.84_{0.009}$	$0.89_{0.009}$
		200	$43.50_{0.07}$	$42.29_{0.08}$	$42.86_{0.08}$	$0.80_{0.011}$	$0.89_{0.011}$
		1000	$40.53_{0.06}$	$39.64_{0.06}$	$39.96_{0.06}$	$0.71_{0.013}$	$0.82_{0.014}$
5	23.23	50	$41.56_{0.11}$	$38.13_{0.11}$	$39.26_{0.12}$	$0.81_{0.005}$	$0.87_{0.005}$
		200	$36.02_{0.07}$	$33.34_{0.06}$	$34.68_{0.07}$	$0.79_{0.004}$	$0.90_{0.004}$
		1000	$31.46_{0.05}$	$29.91_{0.05}$	$30.58_{0.05}$	$0.81_{0.004}$	$0.89_{0.004}$

the `R` package `ks` (Duong, 2015). In practice, we estimated $\|\hat{f}_m\|_\infty$ by the maximum value attained on the unlabelled training set.

In each of the three settings above, we generated a training set of size $n \in \{50, 200, 1000\}$ in dimensions $d \in \{1, 2, 5\}$, an unlabelled training set of size 1000, and a test set of size 1000. In Table 4.1, we present the sample mean and standard error (in subscript) of the risks computed from 1000 repetitions of each experiment. Further, we present estimates of the regret ratios, given by

$$\frac{R(\hat{C}_n^{\hat{k}_{\mathrm{O}}\mathrm{nn}}) - R(C^{\mathrm{Bayes}})}{R(\hat{C}_n^{\hat{k}\mathrm{nn}}) - R(C^{\mathrm{Bayes}})} \quad \text{and} \quad \frac{R(\hat{C}_n^{\hat{k}_{\mathrm{SS}}\mathrm{nn}}) - R(C^{\mathrm{Bayes}})}{R(\hat{C}_n^{\hat{k}\mathrm{nn}}) - R(C^{\mathrm{Bayes}})}, \tag{4.7}$$

for which the standard errors given are estimated via the delta method. From Table 4.1, we saw improvement in performance from the oracle and semi-supervised classifiers in 22 of the 27 experiments, comparable performance in three experiments, and there were two where the standard $k$nn classifier was the best of the three classifiers considered. In those latter two cases, the theoretical improvement expected for the local classifiers is small; for instance, when $d = 5$ in Setting 2, the excess risk for the local classifiers converges at rate $O(n^{-4/9})$, while the standard k-nearest neighbour classifier can attain a rate at least as fast as $o(n^{-1/3+\epsilon})$ for every $\epsilon > 0$. It is therefore perhaps unsurprising that we require the larger sample size of $n = 1000$ for the local classifiers to yield an improvement in this case. The semi-supervised classifier exhibits similar performance to the oracle classifier in all settings, though some deterioration is noticeable in higher dimensions, where it is harder to construct a good estimate of $\bar{f}$ from the unlabelled training data.

4.6 Proofs

In this section, we provide proofs of all of our claimed results, which rely on the general asymptotic expansion presented in Theorem 4.4 below. We begin with some further notation. Define the $d \times n$ matrices $X^n := (X_1 \dots X_n)$ and $x^n := (x_1 \dots x_n)$. 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 $Y_{(1)}, \dots, Y_{(n)}$ 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 $p_r(x) = P_X\bigl(B_r(x)\bigr)$.

4.6.1 A general asymptotic expansion

Let

$$c_n := \sup_{x_0 \in \mathcal{S} : \bar{f}(x_0) \geq k_{\mathrm{L}}(x_0)/(n-1)} g(x_0),$$

where g is defined in assumption **(A.2)**, and for $x \in \mathbb{R}^d$, let

$$\delta_n(x) = \delta_{n,\mathrm{L}}(x) := \frac{k_{\mathrm{L}}(x)}{n-1} c_n^d \log^d\Bigl(\frac{n-1}{k_{\mathrm{L}}(x)}\Bigr). \tag{4.8}$$

Recall that $\mathcal{S} = \{x \in \mathbb{R}^d : \eta(x) = 1/2\}$, and note that by Proposition 4.5 in Section 4.7.2, for $\epsilon > 0$, we can write

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

Let

$$\epsilon_n := \frac{1}{c_n \beta^{1/2} \log^{1/2}(n-1)}, \tag{4.9}$$

and recall the definition of the function $a(\cdot)$ in (4.3).

Theorem 4.4. *Assume **(A.1)**, **(A.2)**, **(A.3)** and **(A.4)(ρ)**, for some $\rho > 0$. For n sufficiently large, let $\mathcal{R}_n \subseteq \bigl\{x \in \mathbb{R}^d : \bar{f}(x) \geq \delta_n(x)\bigr\}$ 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

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

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

$$R_{\mathcal{R}_n}(\hat{C}_n^{k_{\mathrm{L}}\mathrm{nn}}) - R_{\mathcal{R}_n}(C^{\mathrm{Bayes}}) = \int_{\mathcal{S}_n} \frac{\bar{f}(x_0)}{\|\dot{\eta}(x_0)\|} \left\{ \frac{1}{4k_{\mathrm{L}}(x_0)} + \Big(\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\Big)^{4/d} a(x_0)^2 \right\} d\mathrm{Vol}^{d-1}(x_0)$$
$$+ o\big(\gamma_n(k_{\mathrm{L}})\big) + O\big\{P_X\big((\partial\mathcal{R}_n)^{\epsilon_n} \cap \mathcal{S}^{\epsilon_n}\big)\big\}$$

as $n \to \infty$, uniformly for $k_{\mathrm{L}} \in K_{\beta,\tau}$, where

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

Proof of Theorem 4.4. First observe that

$$R_{\mathcal{R}_n}(\hat{C}_n^{k_{\mathrm{L}}\mathrm{nn}}) - R_{\mathcal{R}_n}(C^{\mathrm{Bayes}}) = \int_{\mathcal{R}_n} \big[\mathbb{P}\{\hat{S}_n(x) < 1/2\} - \mathbb{1}_{\{\eta(x)<1/2\}}\big]\{2\eta(x) - 1\}\bar{f}(x)\,dx. \quad (4.10)$$

The proof is presented in seven steps. We will see that the dominant contribution to the integral in (4.10) arises from a small neighbourhood about the Bayes decision boundary, i.e. 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

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

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 (4.10). Finally, we bound the remaining error terms to conclude the proof in Step 7. 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

$$\mu_n(x) - \eta(x) - \Big(\frac{k_{\mathrm{L}}(x)}{n\bar{f}(x)}\Big)^{2/d} a(x) = o\bigg(\Big(\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\Big)^{2/d} g(x_0)\bigg),$$

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

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

where we show in Step 7 that

$$|R_1| = o\biggl\{\biggl(\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\biggr)^{2/d}\biggr\} \tag{4.11}$$

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

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

$$f_{(i)}(u) := n\bar{f}(x+u)\binom{n-1}{i-1}p_{\|u\|}^{i-1}(1-p_{\|u\|})^{n-i} = n\bar{f}(x+u)p_{\|u\|}^{n-1}(i-1), \tag{4.12}$$

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

$$r_n = r_n(x) := \biggl\{\frac{2k_{\mathrm{L}}(x)}{(n-1)\bar{f}(x)a_d}\biggr\}^{1/d}. \tag{4.13}$$

We show in Step 7 that

$$R_2 := \sup_{k_{\mathrm{L}} \in K_{\beta,\tau}} \sup_{x_0 \in \mathcal{S}_n} \sup_{|t| < \epsilon_n} \mathbb{E}\{\|X_{(k_{\mathrm{L}})} - x\|^2 \mathbb{1}_{\{\|X_{(k_{\mathrm{L}})} - x\| \geq r_n\}}\} = O(n^{-M}), \tag{4.14}$$

for each $M > 0$, as $n \to \infty$. It follows from (4.12) and (4.14), together with the assumption on $\|\dot{\eta}(\cdot)\|$ in **(A.3)** that

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

uniformly for $1 \leq i \leq k_{\mathrm{L}}$, $x_0 \in \mathcal{S}_n$ and $|t| < \epsilon_n$. Similarly, using the assumption on $\|\ddot{\eta}(\cdot)\|_{\mathrm{op}}$ in **(A.3)**,

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

uniformly for $1 \leq i \leq k_{\mathrm{L}}$, $x_0 \in \mathcal{S}_n$ and $|t| < \epsilon_n$. Hence, summing over i, we see that

$$\begin{aligned} &\frac{1}{k_{\mathrm{L}}}\sum_{i=1}^{k_{\mathrm{L}}} \mathbb{E}\{(X_{(i)} - x)^T \dot{\eta}(x)\} + \frac{1}{2k_{\mathrm{L}}}\sum_{i=1}^{k_{\mathrm{L}}} \mathbb{E}\{(X_{(i)} - x)^T \ddot{\eta}(x)(X_{(i)} - x)\} \\ &= \int_{B_{r_n}(0)} \Bigl[\dot{\eta}(x)^T u n\{\bar{f}(x+u) - \bar{f}(x)\} + \frac{1}{2}u^T \ddot{\eta}(x) u n \bar{f}(x+u)\Bigr] q_{\|u\|}^{n-1}(k_{\mathrm{L}})\, du + O(n^{-M}), \end{aligned}$$

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 $\epsilon_n < \epsilon_0$,

$$\sup_{k_{\mathrm{L}} \in K_{\beta,\tau}} \sup_{x_0 \in \mathcal{S}_n} \sup_{|t| < \epsilon_n} \max\Bigl\{\Bigl|\frac{k_{\mathrm{L}}(x)}{k_{\mathrm{L}}(x_0)} - 1\Bigr|, \Bigl|\frac{\bar{f}(x)}{\bar{f}(x_0)} - 1\Bigr|\Bigr\} \leq \max\Bigl\{\tau, c_n\epsilon_n + \frac{c_n\epsilon_n^2}{2}\Bigr\} \to 0. \tag{4.15}$$

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 $n \geq n_0$,

$$\left|\bar{f}(x+u) - \bar{f}(x) - u^T \dot{\bar{f}}(x)\right| \leq \frac{\|u\|^2}{2} \sup_{s \in B_{\|u\|}(0)} \|\ddot{\bar{f}}(x+s)\|_{\mathrm{op}} \leq \frac{\|u\|^2}{2} \bar{f}(x_0)g(x_0).$$

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

$$\begin{aligned} |p_r(x) - \bar{f}(x)a_d r^d| &\leq \int_{B_r(0)} |\bar{f}(x+u) - \bar{f}(x) - u^T \dot{\bar{f}}(x)|\, du \\ &\leq \frac{1}{2}\bar{f}(x_0)g(x_0) \int_{B_r(0)} \|u\|^2\, du = \frac{da_d}{2(d+2)} \bar{f}(x_0)g(x_0)r^{d+2}. \end{aligned} \tag{4.16}$$

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

$$\begin{aligned} |R_3| &\leq \frac{da_d(n-1)\bar{f}(x_0)g(x_0)\|v\|^{d+2}r_n^{d+2}}{2(d+2)} \\ &= \frac{dk_{\mathrm{L}}(x)\bar{f}(x_0)g(x_0)\|v\|^{d+2}r_n^2}{(d+2)\bar{f}(x)} \\ &\leq \frac{2^{2/d}dk_{\mathrm{L}}(x)}{a_d^{2/d}(d+2)\log^2\big(\frac{n-1}{k_{\mathrm{L}}(x_0)}\big)} \Big(\frac{\bar{f}(x_0)}{\bar{f}(x)}\Big)^{1+2/d} \Big(\frac{k_{\mathrm{L}}(x)}{k_{\mathrm{L}}(x_0)}\Big)^{2/d}. \end{aligned}$$

It follows from (4.15) 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 $M > 0$,

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

and

$$\sup_{k_{\mathrm{L}} \in K_{\beta,\tau}} \sup_{x_0 \in \mathcal{S}_n} \sup_{|t| < \epsilon_n} \sup_{\|v\|^d \in [1/2 + 1/\log((n-1)/k_{\mathrm{L}}(x_0)), 1)} q_{\|v\|r_n}^{n-1}(k_{\mathrm{L}}(x)) = O(n^{-M}). \tag{4.17}$$

We conclude that

$$\begin{aligned}
&\frac{1}{k_{\mathrm{L}}(x)}\int_{B_{r_n}(0)}\Bigl[\dot\eta(x)^T un\{\bar f(x+u)-\bar f(x)\}+\frac{1}{2}u^T\ddot\eta(x)un\bar f(x+u)\Bigr]q^{n-1}_{\|u\|}(k_{\mathrm{L}}(x))\,du\\
&\quad=\frac{1}{k_{\mathrm{L}}(x)}\int_{B_{2^{-1/d}r_n}(0)}\Bigl[\dot\eta(x)^T un\{\bar f(x+u)-\bar f(x)\}+\frac{1}{2}u^T\ddot\eta(x)un\bar f(x+u)\Bigr]\,du+R_{41}\\
&\quad=\Bigl(\frac{k_{\mathrm{L}}(x)}{n}\Bigr)^{2/d}\frac{\sum_{j=1}^d\{\eta_j(x)\bar f_j(x)+\frac{1}{2}\eta_{jj}(x)\bar f(x)\}}{(d+2)a_d^{2/d}\bar f(x)^{1+2/d}}+R_{41}+R_{42}\\
&\quad=\Bigl(\frac{k_{\mathrm{L}}(x)}{n\bar f(x)}\Bigr)^{2/d}a(x)+R_{41}+R_{42}, \qquad (4.18)
\end{aligned}$$

where

$$|R_{41}|+|R_{42}|=o\biggl(\Bigl(\frac{k_{\mathrm{L}}(x_0)}{n\bar f(x_0)}\Bigr)^{2/d}g(x_0)\biggr),$$

uniformly for $k_{\mathrm{L}}\in K_{\beta,\tau}$, $x_0\in\mathcal{S}_n$ 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

$$\Bigl|\hat\sigma_n^2(x,X^n)-\frac{1}{4k_{\mathrm{L}}}\Bigr|=o_p(1/k_{\mathrm{L}}), \qquad (4.19)$$

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

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

Let $n_2\in\mathbb{N}$ be large enough that $1-c_n\epsilon_n-\frac{d+1}{d+2}c_n\epsilon_n^2\geq\mu_0/a_d$ for $n\geq n_2$. Then for $n\geq\max\{n_0,n_2\}$, $\epsilon<\epsilon_n$, $x_0\in\mathcal{S}_n$ and $|t|<\epsilon_n$, we have by **(A.2)** and a very similar argument to that in (4.16) that

$$p_\epsilon(x)\geq\mu_0\epsilon^d\bar f(x_0)\geq\mu_0\epsilon^d\delta_n(x_0). \qquad (4.20)$$

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

$$1=P_X(\mathbb{R}^d)\geq\sum_{j=1}^N p_{\epsilon_n/12}(z_j)\geq\frac{N\mu_0\beta^{d/2}\log^{d/2}(n-1)}{12^d(n-1)^{1-\beta}}.$$

For each $j=1,\ldots,N$, 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 $j_0:=\mathrm{argmin}_j\|x-z_j\|$, 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 $\{x_1,\ldots,x_n\}$ 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 (4.15), (4.20) and **(A.2)**

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

for all $k_{\mathrm{L}}\in K_{\beta,\tau}$ and $n\geq n_3$, say. Define $A_{k_{\mathrm{L}}}:=\bigl\{\|X_{(k_{\mathrm{L}})}(x)-x\|<\epsilon_n/2 \text{ for all } x\in\mathcal{R}_n\cup\mathcal{S}_n^{\epsilon_n}\bigr\}$. Then by a standard binomial tail bound (Shorack and Wellner, 2009, Equation (6), p. 440), for

$n \geq n_3$ and any $M > 0$,

$$\begin{aligned}
\mathbb{P}(A^c_{k_{\mathrm{L}}}) &= \mathbb{P}\Bigl\{\sup_{x \in \mathcal{R}_n \cup \mathcal{S}_n^{\epsilon_n}} \|X_{(k_{\mathrm{L}}(x))}(x) - x\| \geq \epsilon_n/2\Bigr\} \\
&\leq \mathbb{P}\Bigl\{\max_{j=1,\ldots,N} \|X_{(k_{\mathrm{L}}(z_j))}(z_j') - z_j'\| \geq \epsilon_n/6\Bigr\} \leq \sum_{j=1}^{N} \mathbb{P}\bigl\{\|X_{(k_{\mathrm{L}}(z_j))}(z_j') - z_j'\| \geq \epsilon_n/6\bigr\} \\
&\leq N \max_{j=1,\ldots,N} \exp\Bigl(-\frac{1}{2} n p_{\epsilon_n/6}(z_j') + k_{\mathrm{L}}(z_j')\Bigr) = O(n^{-M}),
\end{aligned} \tag{4.21}$$

uniformly for $k_{\mathrm{L}} \in K_{\beta,\tau}$. Now,

$$\sup_{k_{\mathrm{L}} \in K_{\beta,\tau}} \sup_{x_0 \in \mathcal{S}_n} \sup_{|t| < \epsilon_n} \sup_{x^n \in A_{k_{\mathrm{L}}}} \max_{1 \leq i \leq k_{\mathrm{L}}(x)} |\eta(x_{(i)}(x)) - 1/2| \to 0.$$

It follows that

$$\sup_{x^n \in A_{k_{\mathrm{L}}}} \left| \frac{1}{k_{\mathrm{L}}(x)^2} \sum_{i=1}^{k_{\mathrm{L}}(x)} \eta(x_{(i)}(x))\{1 - \eta(x_{(i)}(x))\} - \frac{1}{4k_{\mathrm{L}}(x)} \right| = o\Bigl(\frac{1}{k_{\mathrm{L}}(x)}\Bigr) \tag{4.22}$$

as $n \to \infty$, uniformly for $x_0 \in \mathcal{S}_n, |t| < \epsilon_n$ and $k_{\mathrm{L}} \in K_{\beta,\tau}$. The claim (4.19) follows from (4.21) and (4.22).

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

$$\mathrm{Var}\{\hat{\mu}_n^{(k_{\mathrm{L}})}(x, X^n)\} = O\biggl\{\frac{1}{k_{\mathrm{L}}}\Bigl(\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\Bigr)^{2/d}\biggr\} \tag{4.23}$$

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

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

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

$$\begin{aligned}
&\mathrm{Var}\{\hat{\mu}_n^{(k_{\mathrm{L}})}(x, X^n)\} \leq \sum_{i=1}^{n} \mathbb{E}\bigl[\{\hat{\mu}_n^{(k_{\mathrm{L}})}(x, X^n) - \hat{\mu}_{n-1}^{(k_{\mathrm{L}})}(x, X^{n,(i)})\}^2\bigr] \\
&\quad = \frac{1}{k_{\mathrm{L}}^2} \sum_{i=1}^{k_{\mathrm{L}}} \mathbb{E}\bigl[\{\eta(X_{(i)}) - \eta(X_{(k_{\mathrm{L}}+1)})\}^2\bigr] \leq \frac{2}{k_{\mathrm{L}}^2} \sum_{i=1}^{k_{\mathrm{L}}} \mathbb{E}\bigl[\{\eta(X_{(i)}) - \eta(x)\}^2 + \{\eta(X_{(k_{\mathrm{L}}+1)}) - \eta(x)\}^2\bigr].
\end{aligned} \tag{4.24}$$

Recall the definition of r_n given in (4.13). Now observe that, for $\max(\epsilon_n, r_n) \leq \epsilon_0$ and all $M > 0$

we have that

$$\begin{aligned}
&\max_{i\in\{1,\ldots,k_{\mathrm{L}}+1\}} \mathbb{E}\bigl[\{\eta(X_{(i)}) - \eta(x)\}^2\bigr] \\
&\qquad\le \max_{i\in\{1,\ldots,k_{\mathrm{L}}+1\}} \mathbb{E}\bigl[\{\eta(X_{(i)}) - \eta(x)\}^2 \mathbb{1}_{\{\|X_{(i)}-x\|\le r_n\}}\bigr] + \mathbb{P}(\|X_{(k_{\mathrm{L}}+1)} - x\| > r_n) \\
&\qquad\le r_n^2 \sup_{z\in\mathcal{S}^{2\epsilon_0}} \|\dot{\eta}(z)\|^2 + O(n^{-M}),
\end{aligned} \tag{4.25}$$

uniformly for $x_0 \in \mathcal{S}_n$, $|t| < \epsilon_n$ and $k_{\mathrm{L}} \in K_{\beta,\tau}$. The final inequality here follows from similar arguments to those used to bound R_1. Now (4.23) follows from (4.24) and (4.25).

Step 4: We show that

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

for each $M > 0$, as $n \to \infty$. First, by **(A.3)** and Proposition 4.5 in Section 4.7.2, there exists $c_0 > 0$ such that for every $\epsilon \in (0, \epsilon_0]$,

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

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

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

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

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

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

$$\psi(x) := \{2\eta(x) - 1\}\bar{f}(x) = \pi_1 f_1(x) - \pi_2 f_2(x).$$

Recalling that $\mathcal{S}_n := \mathcal{S} \cap \mathcal{R}_n$, we show that

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

uniformly for $k_{\mathrm{L}} \in K_{\beta,\tau}$. Now by Proposition 4.6 in Section 4.7.2, for

$$\epsilon_n \le \min\biggl\{\epsilon_0\,,\, \frac{\inf_{x_0\in\mathcal{S}}\|\dot{\eta}(x_0)\|}{\sup_{z\in\mathcal{S}^{\epsilon_0}}\|\ddot{\eta}(z)\|_{\mathrm{op}}}\biggr\},$$

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 $|t| < \epsilon_n$, $\mathrm{sgn}\{\eta(x_0^t) - 1/2\} = \mathrm{sgn}(t)$. 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 $\mathcal{S}_n^{\epsilon_n} \setminus (\mathcal{S}^{\epsilon_n} \cap \mathcal{R}_n) \subseteq (\partial\mathcal{R}_n)^{\epsilon_n} \cap \mathcal{S}^{\epsilon_n}$. It follows from this and (4.56) in Section 4.7.3 that

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

where B is defined in (4.49) in Section 4.7.2, and $\det(I + tB) = 1 + o(1)$ as $n \to \infty$, uniformly for $x_0 \in \mathcal{S}$ and $t \in (-\epsilon_n, \epsilon_n)$.

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

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

as $n \to \infty$, uniformly for $k_{\mathrm{L}} \in K_{\beta,\tau}$. First observe that

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

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

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

We can write

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

where we show in Step 7 that

$$\Big|\int_{\mathcal{S}_n} R_5(x_0) + R_6(x_0)\,d\mathrm{Vol}^{d-1}(x_0)\Big| = o(\gamma_n(k_{\mathrm{L}})). \tag{4.26}$$

Then, substituting $u = 2k_{\mathrm{L}}(x_0)^{1/2}t$, we see that

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

The conclusion follows by integrating with respect to $d\mathrm{Vol}^{d-1}$ over $\mathcal{S}_n$.

Step 7: To complete the proof it remains to bound the error terms R_1, R_2, R_5 and R_6.

To bound R_1: We have

$$R_1 = \frac{1}{k_{\mathrm{L}}}\sum_{i=1}^{k_{\mathrm{L}}}\Big(\mathbb{E}\eta(X_{(i)}) - \eta(x) - \mathbb{E}\{(X_{(i)} - x)^T\dot{\eta}(x)\} - \frac{1}{2}\mathbb{E}\{(X_{(i)} - x)^T\ddot{\eta}(x)(X_{(i)} - x)\}\Big).$$

By a Taylor expansion and the uniform continuity of $\ddot{\eta}$ from **(A.3)**, for all $\epsilon > 0$, there exists $r = r_\epsilon \in (0, \epsilon_0]$, such that for all $x \in \mathcal{S}^{\epsilon_0}$ and $\|z - x\| < r$,

$$\Big|\eta(z) - \eta(x) - (z - x)^T\dot{\eta}(x) - \frac{1}{2}(z - x)^T\ddot{\eta}(x)(z - x)\Big| \leq \epsilon\|z - x\|^2.$$

Hence

$$\begin{aligned}|R_1| \leq \epsilon\frac{1}{k_{\mathrm{L}}}\sum_{i=1}^{k_{\mathrm{L}}}\mathbb{E}\{\|X_{(i)} - x\|^2\mathbb{1}_{\{\|X_{(k_{\mathrm{L}})} - x\| \leq r\}}\} + 2\mathbb{P}\{\|X_{(k_{\mathrm{L}})} - x\| > r\}&\\ + \sup_{z\in\mathcal{S}^{\epsilon_0}}\|\dot{\eta}(z)\|\mathbb{E}\{\|X_{(k_{\mathrm{L}})} - x\|\mathbb{1}_{\{\|X_{(k_{\mathrm{L}})} - x\| > r\}}\}&\\ + \sup_{z\in\mathcal{S}^{\epsilon_0}}\|\ddot{\eta}(z)\|_{\mathrm{op}}\mathbb{E}\{\|X_{(k_{\mathrm{L}})} - x\|^2\mathbb{1}_{\{\|X_{(k_{\mathrm{L}})} - x\| > r\}}\}&.\end{aligned} \tag{4.27}$$

Now, by similar arguments to those leading to (4.18), we have that

$$\frac{\epsilon}{k_{\mathrm{L}}}\sum_{i=1}^{k_{\mathrm{L}}}\mathbb{E}(\|X_{(i)}-x\|^2\mathbb{1}_{\{\|X_{(k_{\mathrm{L}})}-x\|\leq r\}})=\epsilon\Big(\frac{k_{\mathrm{L}}}{na_d\bar{f}(x)}\Big)^{2/d}\frac{d}{d+2}\{1+o(1)\}, \tag{4.28}$$

uniformly for $x_0\in\mathcal{S}_n, |t|<\epsilon_n$ and $k_{\mathrm{L}}\in K_{\beta,\tau}$. Moreover, for every $M>0$,

$$\mathbb{P}\{\|X_{(k_{\mathrm{L}})}-x\|>r\}=q_r^n(k_{\mathrm{L}})=O(n^{-M}), \tag{4.29}$$

uniformly for $x_0\in\mathcal{S}_n, |t|<\epsilon_n$ and $k_{\mathrm{L}}\in K_{\beta,\tau}$, by (4.17) in Step 1. For the remaining terms, note that

$$\begin{aligned}\mathbb{E}\{\|X_{(k_{\mathrm{L}})}-x\|^2\mathbb{1}_{\{\|X_{(k_{\mathrm{L}})}-x\|>r\}}\}&=\mathbb{P}\{\|X_{(k_{\mathrm{L}})}-x\|>r\}+\int_{r^2}^{\infty}\mathbb{P}\{\|X_{(k_{\mathrm{L}})}-x\|>\sqrt{t}\}\,dt\\&=q_r^n(k_{\mathrm{L}})+\int_{r^2}^{\infty}q_{\sqrt{t}}^n(k_{\mathrm{L}})\,dt.\end{aligned} \tag{4.30}$$

Let $t_0=t_0(x):=5^{2/\rho}(1+2^{\rho-1})^{2/\rho}\big\{\mathbb{E}(\|X\|^\rho)+\|x\|^\rho\big\}^{2/\rho}$. Then, for $t\geq t_0$, we have

$$1-p_{\sqrt{t}}\leq(1+2^{\rho-1})\frac{\mathbb{E}(\|X\|^\rho)+\|x\|^\rho}{t^{\rho/2}}\leq\frac{1}{5}.$$

It follows by Bennett's inequality that for $\rho\{n-(n-1)^{1-\beta}\}>4$,

$$\begin{aligned}\int_{t_0}^{\infty}q_{\sqrt{t}}^n(k_{\mathrm{L}})\,dt&\leq e^{k_{\mathrm{L}}}(1+2^{\rho-1})^{(n-k_{\mathrm{L}})/2}\big\{\mathbb{E}(\|X\|^\rho)+\|x\|^\rho\big\}^{(n-k_{\mathrm{L}})/2}\int_{t_0}^{\infty}t^{-\rho(n-k_{\mathrm{L}})/4}\,dt\\&=\frac{4e^{k_{\mathrm{L}}}5^{2/\rho}}{\rho(n-k_{\mathrm{L}})-4}(1+2^{\rho-1})^{2/\rho}\big\{\mathbb{E}(\|X\|^\rho)+\|x\|^\rho\big\}^{2/\rho}5^{-(n-k_{\mathrm{L}})/2}.\end{aligned}$$

But, when $\beta\log(n-1)\geq(d+2)/d$ and $n\geq\max\{n_0,n_2\}$,

$$\sup_{x\in\mathcal{R}_n\cup\mathcal{S}_n^{\epsilon_n}}\|x\|\leq\epsilon_0+\bigg\{\frac{(n-1)^{1-\beta}c_n^d\mathbb{E}(\|X\|^\rho)}{\mu_0\beta^{d/2}\log^{d/2}(n-1)}\bigg\}^{1/\rho}.$$

We deduce that for every $M>0$,

$$\sup_{k\in K_{\beta,\tau}}\sup_{x\in\mathcal{R}_n\cup\mathcal{S}_n^{\epsilon_n}}\int_{t_0}^{\infty}q_{\sqrt{t}}^n(k_{\mathrm{L}})\,dt=O(n^{-M}). \tag{4.31}$$

Moreover, by Bernstein's inequality, for every $M>0$,

$$\sup_{k_{\mathrm{L}}\in K_{\beta,\tau}}\sup_{x\in\mathcal{R}_n\cup\mathcal{S}_n^{\epsilon_n}}\Big\{q_r^n(k_{\mathrm{L}})+\int_{r^2}^{t_0}q_{\sqrt{t}}^n(k_{\mathrm{L}})\,dt\Big\}=O(n^{-M}). \tag{4.32}$$

We conclude from (4.15), (4.27), (4.28), (4.29), (4.30), (4.31) and (4.32), together with Jensen's inequality to deal with the third term on the right-hand side of (4.27), that (4.11) holds. With only simple modifications, we have also shown (4.14), which bounds R_2.

To bound R_5: Write

$$R_5:=\int_{\mathcal{S}_n}R_5(x_0)\,d\mathrm{Vol}^{d-1}(x_0)\int_{\mathcal{S}_n}\int_{-\epsilon_n}^{\epsilon_n}t\|\dot{\psi}(x_0)\|\Big[\mathbb{P}\{\hat{S}_n(x_0^t)<1/2\}-\mathbb{E}\Phi\big(\hat{\theta}(x_0^t)\big)\Big]\,dt\,d\mathrm{Vol}^{d-1}(x_0).$$

Now by a non-uniform version of the Berry–Esseen theorem (Paditz, 1989, Theorem 1), for every $t \in (-\epsilon_n, \epsilon_n)$ and $x_0 \in \mathcal{S}_n$,

$$\bigl|\mathbb{P}\{\hat{S}_n(x_0^t) < 1/2 | X^n\} - \Phi\bigl(\hat{\theta}(x_0^t)\bigr)\bigr| \leq \frac{32}{k_{\mathrm{L}}(x_0^t)\hat{\sigma}_n(x_0^t, X^n)} \frac{1}{1 + |\hat{\theta}(x_0^t)|^3}. \tag{4.33}$$

Let

$$t_n = t_n(x_0) := C \max\biggl\{k_{\mathrm{L}}(x_0)^{-1/2}, \Bigl(\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\Bigr)^{2/d} g(x_0)\biggr\},$$

where

$$C := \frac{4 \max\{2 \sup_{z \in \mathcal{S}^{\epsilon_0}} \|\dot{\eta}(z)\|, d \sup_{z \in \mathcal{S}^{\epsilon_0}} \|\ddot{\eta}(z)\|_{\mathrm{op}}\}}{(d+2) a_d^{2/d} \inf_{z \in \mathcal{S}} \|\dot{\eta}(z)\|}.$$

In the following we integrate the bound in (4.33) over the regions $|t| \leq t_n$ and $|t| \in (t_n, \epsilon_n)$ separately. Define the event

$$B_{k_{\mathrm{L}}} := \biggl\{\hat{\sigma}_n(x_0^t, X^n) \geq \frac{1}{3k_{\mathrm{L}}(x_0^t)^{1/2}} \text{ for all } x_0 \in \mathcal{S}_n,\, t \in (-\epsilon_n, \epsilon_n)\biggr\},$$

so that, by very similar arguments to those used to bound $\mathbb{P}(A_{k_{\mathrm{L}}}^c)$ in Step 2, we have $\mathbb{P}(B_{k_{\mathrm{L}}}^c) = O(n^{-M})$ for every $M > 0$, uniformly for $k_{\mathrm{L}} \in K_{\beta,\tau}$. It follows by (4.33) and Step 2 that there exists $n_4 \in \mathbb{N}$ such that for all $n \geq n_4$, $k_{\mathrm{L}} \in K_{\beta,\tau}$ and $x_0 \in \mathcal{S}_n$,

$$\begin{aligned}\biggl|\int_{-t_n}^{t_n} t \Bigl[\mathbb{P}\{\hat{S}_n(x_0^t) < 1/2\} - \mathbb{E}\Phi\bigl(\hat{\theta}(x_0^t)\bigr)\Bigr]\, dt\biggr| \\ \leq \int_{-t_n}^{t_n} \mathbb{E}\biggl(\frac{32|t| \mathbb{1}_{B_{k_{\mathrm{L}}}}}{k_{\mathrm{L}}(x_0^t)\hat{\sigma}_n(x_0^t, X^n)}\biggr)\, dt + t_n^2 \mathbb{P}(B_{k_{\mathrm{L}}}^c) \leq \frac{128 t_n^2}{k_{\mathrm{L}}(x_0)^{1/2}}.\end{aligned} \tag{4.34}$$

By Step 1, there exists $n_5 \in \mathbb{N}$ such that for $n \geq n_5, |t| \in (t_n, \epsilon_n), x_0 \in \mathcal{S}_n$ and $k_{\mathrm{L}} \in K_{\beta,\tau}$,

$$\begin{aligned}|\mu_n(x_0^t) - 1/2| &\geq |\eta(x_0^t) - 1/2| - |\mu_n(x_0^t) - \eta(x_0^t)| \\ &\geq \frac{1}{2} \inf_{z \in \mathcal{S}} \|\dot{\eta}(z)\||t| - \frac{1}{4} C \inf_{z \in \mathcal{S}} \|\dot{\eta}(z)\| \Bigl(\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\Bigr)^{2/d} g(x_0) > \frac{1}{4} \inf_{z \in \mathcal{S}} \|\dot{\eta}(z)\||t|.\end{aligned} \tag{4.35}$$

Thus for $n \geq n_5, |t| \in (t_n, \epsilon_n), x_0 \in \mathcal{S}_n$ and $k_{\mathrm{L}} \in K_{\beta,\tau}$ we have that

$$\begin{aligned}&\mathbb{P}\Bigl\{|\hat{\theta}(x_0^t)| < \frac{1}{4} \inf_{z \in \mathcal{S}} \|\dot{\eta}(z)\| k_{\mathrm{L}}^{1/2}(x_0)|t|\Bigr\} \\ &\leq \mathbb{P}\Bigl\{|\hat{\mu}_n(x_0^t, X^n) - \mu_n(x_0^t)| > |\mu_n(x_0^t) - 1/2| - \frac{1}{8} \inf_{z \in \mathcal{S}} \|\dot{\eta}(z)\||t|\Bigr\} \\ &\leq \mathbb{P}\Bigl\{|\hat{\mu}_n(x_0^t, X^n) - \mu_n(x_0^t)| > \frac{1}{8} \inf_{z \in \mathcal{S}} \|\dot{\eta}(z)\||t|\Bigr\} \leq \frac{64 \mathrm{Var}\{\hat{\mu}_n(x_0^t, X^n)\}}{\inf_{z \in \mathcal{S}} \|\dot{\eta}(z)\|^2 t^2}.\end{aligned} \tag{4.36}$$

It follows by (4.33), (4.36) and Step 3 that, for $n \geq n_5$,

$$\begin{aligned}
&\biggl|\int_{|t|\in(t_n,\epsilon_n)} t\bigl[\mathbb{P}\{\hat{S}_n(x_0^t) < 1/2\} - \mathbb{E}\Phi\bigl(\hat{\theta}(x_0^t)\bigr)\bigr]\,dt\biggr| \\
&\leq \int_{|t|\in(t_n,\epsilon_n)} |t|\mathbb{E}\biggl(\frac{32\mathbb{1}_{B_{k_\mathrm{L}}}}{k_\mathrm{L}(x_0^t)\hat{\sigma}_n(x_0^t,X^n)}\frac{1}{1+\frac{1}{64}\inf_{z\in\mathcal{S}}\|\dot{\eta}(z)\|^3 k_\mathrm{L}(x_0)^{3/2}|t|^3}\biggr)\,dt \\
&\qquad + \int_{|t|\in(t_n,\epsilon_n)} \frac{64\mathrm{Var}\{\hat{\mu}_n(x_0^t,X^n)\}}{\inf_{z\in\mathcal{S}}\|\dot{\eta}(z)\|^2|t|}\,dt + \epsilon_n^2\mathbb{P}(B_{k_\mathrm{L}}^c) \\
&\leq \frac{192}{k_\mathrm{L}(x_0)^{3/2}}\int_0^\infty \frac{u}{1+\frac{1}{64}\inf_{z\in\mathcal{S}}\|\dot{\eta}(z)\|^3u^3}\,du \\
&\qquad + \frac{128}{\inf_{z\in\mathcal{S}}\|\dot{\eta}(z)\|^2}\sup_{|t|\in(t_n,\epsilon_n)}\mathrm{Var}\{\hat{\mu}_n(x_0^t,X^n)\}\log\Bigl(\frac{\epsilon_n}{t_n}\Bigr) + \epsilon_n^2\mathbb{P}(B_{k_\mathrm{L}}^c) \\
&= o\Bigl(\frac{1}{k_\mathrm{L}(x_0)}\Bigr) \qquad (4.37)
\end{aligned}$$

uniformly for $x_0 \in \mathcal{S}_n$ and $k_\mathrm{L} \in K_{\beta,\tau}$. We conclude from (4.34) and (4.37) that $|R_5| = o(\gamma_n(k_\mathrm{L}))$.

To bound R_6: Let $\theta(x_0^t) := -2k_\mathrm{L}(x_0^t)^{1/2}\{\mu_n(x_0^t) - 1/2\}$. Write

$$R_6 := \int_{\mathcal{S}_n} R_6(x_0)\,d\mathrm{Vol}^{d-1}(x_0) = R_{61} + R_{62},$$

where

$$R_{61} := \int_{\mathcal{S}_n}\int_{-\epsilon_n}^{\epsilon_n} t\|\dot{\psi}(x_0)\|\Bigl[\mathbb{E}\Phi\bigl(\hat{\theta}(x_0^t)\bigr) - \Phi\bigl(\theta(x_0^t)\bigr)\Bigr]\,dt\,d\mathrm{Vol}^{d-1}(x_0)$$

and

$$R_{62} := \int_{\mathcal{S}_n}\int_{-\epsilon_n}^{\epsilon_n} t\|\dot{\psi}(x_0)\|\Bigl[\Phi\bigl(\theta(x_0^t)\bigr) - \Phi\bigl(\bar{\theta}(x_0,t)\bigr)\Bigr]\,dt\,d\mathrm{Vol}^{d-1}(x_0).$$

To bound R_{61}: We again deal with the regions $|t| \leq t_n$ and $|t| \in (t_n, \epsilon_n)$ separately. First let $\tilde{\theta}(x_0^t) := -2k_\mathrm{L}(x_0^t)^{1/2}\{\hat{\mu}_n(x_0^t,X^n) - 1/2\}$. Writing ϕ for the standard normal density, and using the facts that $|\hat{\theta}(x_0^t)| \geq |\tilde{\theta}(x_0^t)|$, that $\hat{\theta}(x_0^t)$ and $\tilde{\theta}(x_0^t)$ have the same sign, and that $|x\phi(x)| \leq 1$, we have

$$\begin{aligned}
&\biggl|\int_{-t_n}^{t_n} t\bigl[\mathbb{E}\Phi\bigl(\hat{\theta}(x_0^t)\bigr) - \Phi\bigl(\theta(x_0^t)\bigr)\bigr]\,dt\biggr| \\
&\leq \int_{-t_n}^{t_n} |t|\mathbb{E}\Bigl\{|\hat{\theta}(x_0^t) - \tilde{\theta}(x_0^t)|\phi\bigl(\tilde{\theta}(x_0^t)\bigr)\mathbb{1}_{A_{k_\mathrm{L}}} + |\tilde{\theta}(x_0^t) - \theta(x_0^t)|\Bigr\}\,dt + t_n^2\mathbb{P}(A_{k_\mathrm{L}}^c) \\
&\leq \int_{-t_n}^{t_n} |t|\biggl[\mathbb{E}\biggl\{\mathbb{1}_{A_{k_\mathrm{L}}}\biggl|\frac{1}{2k_\mathrm{L}(x_0^t)^{1/2}\hat{\sigma}_n(x_0^t,X^n)} - 1\biggr|\biggr\} \\
&\qquad\qquad + 2k_\mathrm{L}(x_0^t)^{1/2}\mathrm{Var}^{1/2}\{\hat{\mu}_n(x_0^t,X^n)\}\biggr]\,dt + t_n^2\mathbb{P}(A_{k_\mathrm{L}}^c) = o(t_n^2)
\end{aligned}$$

uniformly for $x_0 \in \mathcal{S}_n$ and $k_\mathrm{L} \in K_{\beta,\tau}$. Note that for $|t| \in (t_n, \epsilon_n)$ and $x_0 \in \mathcal{S}_n$, we have when

$\epsilon_n < \epsilon_0$ and $n \geq n_5$ that

$$\begin{aligned}\mathbb{E}\bigl\{\mathbb{1}_{A_{k_\mathrm{L}}\cap B_{k_\mathrm{L}}}\bigl|\hat{\theta}(x_0^t) - \theta(x_0^t)\bigr|\bigr\} &\leq \mathbb{E}\biggl\{\frac{\mathbb{1}_{A_{k_\mathrm{L}}\cap B_{k_\mathrm{L}}}}{\hat{\sigma}_n(x_0^t, X^n)}|\hat{\mu}_n(x_0^t, X^n) - \mu_n(x_0^t)| \\ &\qquad + \mathbb{1}_{A_{k_\mathrm{L}}\cap B_{k_\mathrm{L}}}|\theta(x_0^t)|\Bigl|\frac{1}{2k_\mathrm{L}(x_0^t)^{1/2}\hat{\sigma}_n(x_0^t, X^n)} - 1\Bigr|\biggr\} \\ &\leq 3k_\mathrm{L}(x_0)^{1/2}\mathrm{Var}^{1/2}\{\hat{\mu}_n(x_0^t, X^n)\} \\ &\qquad + \frac{5}{2}k_\mathrm{L}(x_0)^{1/2}\sup_{z\in\mathcal{S}^{\epsilon_0}}\|\dot{\eta}(z)\||t|\mathbb{E}\biggl\{\mathbb{1}_{A_{k_\mathrm{L}}\cap B_{k_\mathrm{L}}}\Bigl|\frac{1}{2k_\mathrm{L}(x_0^t)^{1/2}\hat{\sigma}_n(x_0^t, X^n)} - 1\Bigr|\biggr\}. \end{aligned} \tag{4.38}$$

Thus by (4.35), (4.36), (4.38) and Step 3, for $\epsilon_n < \epsilon_0$ and $n \geq n_5$,

$$\begin{aligned}&\int_{|t|\in(t_n,\epsilon_n)}|t|\bigl|\mathbb{E}\Phi\bigl(\hat{\theta}(x_0^t)\bigr) - \Phi\bigl(\theta(x_0^t)\bigr)\bigr|\,dt \\ &\leq \int_{|t|\in(t_n,\epsilon_n)}|t|\mathbb{E}\Bigl\{\mathbb{1}_{A_{k_\mathrm{L}}\cap B_{k_\mathrm{L}}}\bigl|\hat{\theta}(x_0^t) - \theta(x_0^t)\bigr|\Bigr\}\phi\Bigl(\frac{1}{4}\inf_{z\in\mathcal{S}}\|\dot{\eta}(z)\|k_\mathrm{L}^{1/2}(x_0)|t|\Bigr)\,dt \\ &\quad + \mathbb{P}(A_{k_\mathrm{L}}^c \cup B_{k_\mathrm{L}}^c) + \frac{128}{\inf_{z\in\mathcal{S}}\|\dot{\eta}(z)\|^2}\sup_{|t|\in(t_n,\epsilon_n)}\mathrm{Var}\{\hat{\mu}_n(x_0^t, X^n)\}\log\Bigl(\frac{\epsilon_n}{t_n}\Bigr) = o\Bigl(\frac{1}{k_\mathrm{L}(x_0)}\Bigr) \end{aligned} \tag{4.39}$$

uniformly for $x_0 \in \mathcal{S}_n$ and $k_\mathrm{L} \in K_{\beta,\tau}$.

To bound R_{62}: Let

$$u(x) := k_\mathrm{L}(x)^{1/2}\Bigl(\frac{k_\mathrm{L}(x)}{n\bar{f}(x)}\Bigr)^{2/d}.$$

Given $\epsilon > 0$ small enough that $\epsilon^2 + \frac{\epsilon}{2\inf_{x\in\mathcal{S}}\|\dot{\eta}(x_0)\|} < 1/2$, by Step 1 there exists $n_6 \in \mathbb{N}$ such that for $n \geq n_6$, $k_\mathrm{L} \in K_{\beta,\tau}$, $x_0 \in \mathcal{S}_n$ and $|t| < \epsilon_n$,

$$\bigl|\theta(x_0^t) - \bar{\theta}(x_0, t)\bigr| \leq \epsilon^2\bigl\{|t|k_\mathrm{L}(x_0)^{1/2} + u(x_0)g(x_0)\bigr\}.$$

By decreasing ϵ and increasing n_6 if necessary, it follows that

$$\bigl|\Phi\bigl(\theta(x_0^t)\bigr) - \Phi\bigl(\bar{\theta}(x_0, t)\bigr)\bigr| \leq \epsilon^2\bigl\{|t|k_\mathrm{L}(x_0)^{1/2} + u(x_0)g(x_0)\bigr\}\phi\Bigl(\frac{1}{2}\bar{\theta}(x_0, t)\Bigr),$$

for all $n \geq n_6$, $k_\mathrm{L} \in K_{\beta,\tau}$, and $x_0 \in \mathcal{S}_n$, $t \in (-\epsilon_n, \epsilon_n)$ satisfying $2\epsilon u(x_0)g(x_0)\|\dot{\eta}(x_0)\| \leq |\bar{\theta}(x_0, t)|$. Substituting $u = \bar{\theta}(x_0, t)/2$, it follows that there exists $C^* > 0$ such that for all $n \geq n_6$ and all $k_\mathrm{L} \in K_{\beta,\tau}$,

$$\begin{aligned}|R_{62}| &\leq \int_{\mathcal{S}_n}\int_{|u|\leq\epsilon u(x_0)g(x_0)\|\dot{\eta}(x_0)\|}\frac{2\bar{f}(x_0)}{\|\dot{\eta}(x_0)\|k_\mathrm{L}(x_0)}|u + u(x_0)a(x_0)|\,du\,d\mathrm{Vol}^{d-1}(x_0) \\ &\quad + \int_{\mathcal{S}_n}\int_{-\infty}^{\infty}\frac{2\bar{f}(x_0)|u + u(x_0)a(x_0)|}{\|\dot{\eta}(x_0)\|^2k_\mathrm{L}(x_0)}\bigl\{\epsilon^2|u + u(x_0)a(x_0)| \\ &\qquad + \epsilon|u|\bigr\}\phi(u)\,du\,d\mathrm{Vol}^{d-1}(x_0) \leq C^*\epsilon\gamma_n(k_\mathrm{L}). \end{aligned} \tag{4.40}$$

The combination of (4.39) and (4.40) yields the desired error bound on $|R_6|$ in (4.26) and therefore completes the proof.

□

4.6.2 Proof of Theorem 4.1

Proof of Theorem 4.1. Let $k \in K_\beta$, and note that since $k_{\mathrm{L}}(x) = k$ is constant, we have that $c_n = \sup_{x_0 \in \mathcal{S} : \bar{f}(x_0) \geq k/(n-1)} g(x_0)$, 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 let $n_0 \in \mathbb{N}$ be large enough that $\mathcal{R}_n$ is non-empty for $n \geq n_0$, 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. For $n \geq n_0$, we may apply Theorem 4.4 with $k_{\mathrm{L}}(x) = k$ for all $x \in \mathbb{R}^d$ to deduce that

$$R_{\mathcal{R}_n}(\hat{C}_n^{k\mathrm{nn}}) - R_{\mathcal{R}_n}(C^{\mathrm{Bayes}}) = B_{1,n}\frac{1}{k} + B_{2,n}\Big(\frac{k}{n}\Big)^{4/d} + o(\gamma_n(k)) + O\big\{P_X\big((\partial\mathcal{R}_n)^{\epsilon_n} \cap \mathcal{S}^{\epsilon_n}\big)\big\}$$

uniformly for $k \in K_\beta$, where

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

and

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

and $\mathcal{S}_n := \mathcal{S} \cap \mathcal{R}_n$. 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 finite. First, by Assumptions **(A.3)** and **(A.4)**(ρ),

$$B_1 = \int_{\mathcal{S}} \frac{\bar{f}(x_0)}{4\|\dot{\eta}(x_0)\|} \, d\mathrm{Vol}^{d-1}(x_0) \leq \frac{1}{4\inf_{x_0 \in \mathcal{S}} \|\dot{\eta}(x_0)\|} \int_{\mathcal{S}} \bar{f}(x_0) \, d\mathrm{Vol}^{d-1}(x_0) < \infty.$$

Moreover,

$$B_1 - B_{1,n} = \int_{\mathcal{S} \setminus \mathcal{R}_n} \frac{\bar{f}(x_0)}{4\|\dot{\eta}(x_0)\|} \, d\mathrm{Vol}^{d-1}(x_0) \leq \frac{1}{4} \frac{1}{\inf_{x_0 \in \mathcal{S}} \|\dot{\eta}(x_0)\|} \int_{\mathcal{S} \setminus \mathcal{R}_n} \bar{f}(x_0) \, d\mathrm{Vol}^{d-1}(x_0) \to 0,$$

uniformly for $k \in K_\beta$. By Assumptions **(A.2)**, **(A.3)** and **(A.4)**(ρ) and the fact that $\rho/(\rho + d) > 4/d$, we have that

$$\begin{aligned} B_2 &= \int_{\mathcal{S}} \frac{\bar{f}(x_0)^{1-4/d}}{\|\dot{\eta}(x_0)\|} a(x_0)^2 \, d\mathrm{Vol}^{d-1}(x_0) \\ &\leq \sup_{x_0 \in \mathcal{S}} \Big\{ \frac{a(x_0)^2 \bar{f}(x_0)^{\rho/(\rho+d)-4/d}}{\|\dot{\eta}(x_0)\|} \Big\} \int_{\mathcal{S}} \bar{f}(x_0)^{d/(\rho+d)} \, d\mathrm{Vol}^{d-1}(x_0) < \infty. \end{aligned}$$

Similarly,

$$\begin{aligned} B_2 - B_{2,n} &= \int_{\mathcal{S} \setminus \mathcal{R}_n} \frac{\bar{f}(x_0)^{1-4/d}}{\|\dot{\eta}(x_0)\|} a(x_0)^2 \, d\mathrm{Vol}^{d-1}(x_0) \\ &\leq \sup_{x_0 \in \mathcal{S}} \Big\{ \frac{a(x_0)^2 \bar{f}(x_0)^{\rho/(\rho+d)-4/d}}{\|\dot{\eta}(x_0)\|} \Big\} \int_{\mathcal{S} \setminus \mathcal{R}_n} \bar{f}(x_0)^{d/(\rho+d)} \, d\mathrm{Vol}^{d-1}(x) \to 0, \end{aligned}$$

uniformly for $k \in K_\beta$, as $n \to \infty$. A similar argument shows that $\gamma_n(k) = O(1/k + (k/n)^{4/d})$, uniformly for $k \in K_\beta$.

Finally, we bound $P_X\bigl((\partial\mathcal{R}_n)^{\epsilon_n} \cap \mathcal{S}^{\epsilon_n}\bigr)$ 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

$$\Bigl|\frac{\bar{f}(x)}{\bar{f}(z)} - 1\Bigr| \leq g(z)\|x-z\| + \frac{1}{2}g(z)\|x-z\|^2 \leq \frac{1+\epsilon_n/2}{\beta^{1/2}\log^{1/2}(n-1)}. \tag{4.41}$$

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)**(ρ) and Hölder's inequality, observe that for any $\alpha \in (0,1)$, $n \geq n_1$ and $\epsilon > 0$,

$$\begin{aligned} P_X\bigl((\partial\mathcal{R}_n)^{\epsilon_n} \cap \mathcal{S}^{\epsilon_n}\bigr) &\leq \mathbb{P}\{\bar{f}(X) \leq 2\delta_n\} \leq (2\delta_n)^{\frac{\rho(1-\alpha)}{\rho+d}} \int_{x:\bar{f}(x)\leq 2\delta_n} \bar{f}(x)^{1-\frac{\rho(1-\alpha)}{\rho+d}}\,dx \\ &\leq (2\delta_n)^{\frac{\rho(1-\alpha)}{\rho+d}} \biggl\{\int_{\mathbb{R}^d} (1+\|x\|^\rho)\bar{f}(x)\,dx\biggr\}^{1-\frac{\rho(1-\alpha)}{\rho+d}} \biggl\{\int_{\mathbb{R}^d} \frac{1}{(1+\|x\|^\rho)^{\frac{d+\rho\alpha}{\rho(1-\alpha)}}}\,dx\biggr\}^{\frac{\rho(1-\alpha)}{\rho+d}} \\ &= o\biggl(\Bigl(\frac{k}{n}\Bigr)^{\frac{\rho(1-\alpha)}{\rho+d}-\epsilon}\biggr), \end{aligned} \tag{4.42}$$

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

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

so the same bound (4.42) applies for this region. Since $\rho/(\rho+d) > 4/d$, 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 $B_{1,n} \to B_1 < \infty$, uniformly for $k \in K_\beta$. Furthermore, using Assumptions **(A.3)** and **(A.4)**(ρ) and the fact that $4/d > \rho/(\rho+d)$, we see that

$$\begin{aligned} B_{2,n}\Bigl(\frac{k}{n}\Bigr)^{4/d} &\leq \delta_n^{\rho/(\rho+d)} \int_{\mathcal{S}_n} \frac{\delta_n^{4/d-\rho/(\rho+d)}}{c_n^4\log^4((n-1)/k)} \frac{\bar{f}(x_0)^{1-4/d}}{\|\dot{\eta}(x_0)\|} a(x_0)^2\,d\mathrm{Vol}^{d-1}(x_0) \\ &\leq \sup_{x_0\in\mathcal{S}_n}\Bigl\{\frac{a(x_0)^2}{\|\dot{\eta}(x_0)\|}\Bigr\} \frac{\delta_n^{\rho/(\rho+d)}}{c_n^4\log^4((n-1)/k)} \int_{\mathcal{S}} \bar{f}(x_0)^{d/(\rho+d)}\,d\mathrm{Vol}^{d-1}(x_0) = o((k/n)^{\rho/(\rho+d)-\epsilon}), \end{aligned}$$

for every $\epsilon > 0$, uniformly for $k \in K_\beta$, where the final equality follows from the fact that $\sup_{x_0\in\mathcal{S}_n} a^2(x_0)/c_n^2$ 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). □

4.6.3 Proof of claim in Example 4.1

Proof of claim in Example 4.1. Fix $\epsilon > 0$ and $k \in K_\beta$, let

$$\mathcal{T}_n := (0,1/2) \times \bigl((1+\epsilon)\log(n/k), \infty\bigr),$$

and for $\gamma > 0$, let

$$B_{k,\gamma} = \bigcap_{x=(x_1,x_2)\in\mathcal{T}_n} \{\gamma < \|X_{(k+1)}(x) - x\| < x_2 - 1\}.$$

Now, for $\epsilon\beta \log n > 4$ and $\gamma \in [2, \epsilon \log(n/k)/2)$,

$$\mathbb{P}(B^c_{k,\gamma}) \leq \mathbb{P}(T \geq k+1) + \mathbb{P}(T' \leq k),$$

where $T \sim \mathrm{Bin}(n, p^*_\gamma)$, $T' \sim \mathrm{Bin}(n, p_*)$,

$$\begin{aligned} p^*_\gamma &:= \int_0^1 \int_{(1+\epsilon)\log(n/k)-\gamma}^\infty t_1 \exp(-t_2)\, dt_1 dt_2 \leq \frac{1}{2}\Big(\frac{k}{n}\Big)^{1+\epsilon} e^\gamma \leq \frac{1}{2}\Big(\frac{k}{n}\Big)^{1+\epsilon/2}, \\ p_* &:= \int_0^1 \int_{3-3^{1/2}}^{3+3^{1/2}} t_1 \exp(-t_2)\, dt_1 dt_2 \geq \frac{1}{8}. \end{aligned}$$

Therefore, there exists $n_0 \in \mathbb{N}$ such that $np_* - (k+1) \geq k/2$ and $k + 1 - np^*_\gamma \geq k/2$ for all $k \in K_\beta$, $\gamma \in [2, \epsilon \log(n/k)/2)$ and $n \geq n_0$. It follows by an application of Bernstein's inequality that $\sup_{k \in K_\beta} \sup_{\gamma \in [2, \epsilon \log(n/k)/2)} \mathbb{P}(B^c_{k,\gamma}) = O(n^{-M})$ for every $M > 0$.

Now, for $x = (x_1, x_2) \in \mathcal{T}_n$, $\epsilon\beta \log n > 4$ and $\gamma \in [2, x_2 - 1)$, we have that

$$\begin{aligned} \frac{\int_{B_\gamma(x)} \eta(t) \bar{f}(t)\, dt}{\int_{B_\gamma(x)} \bar{f}(t)\, dt} &= \frac{\int_0^1 \int_{x_2 - \{\gamma^2 - (t_1 - x_1)^2\}^{1/2}}^{x_2 + \{\gamma^2 - (t_1 - x_1)^2\}^{1/2}} t_1^2 e^{-t_2}\, dt_2\, dt_1}{\int_0^1 \int_{x_2 - \{\gamma^2 - (t_1 - x_1)^2\}^{1/2}}^{x_2 + \{\gamma^2 - (t_1 - x_1)^2\}^{1/2}} t_1 e^{-t_2}\, dt_2\, dt_1} = \frac{\int_0^1 t_1^2 \sinh(\{\gamma^2 - (t_1 - x_1)^2\}^{1/2})\, dt_1}{\int_0^1 t_1 \sinh(\{\gamma^2 - (t_1 - x_1)^2\}^{1/2})\, dt_1} \\ &\geq \frac{2}{3} \frac{\sinh\big((\gamma^2 - 1)^{1/2}\big)}{\sinh(\gamma)} \geq \frac{2}{3} \frac{\sinh(3^{1/2})}{\sinh(2)} > \frac{1}{2}. \end{aligned}$$

Our next observation is that for $\gamma \in [0, \infty)$ and $x_{(k+1)} \in \mathbb{R}^d$ such that $\|x_{(k+1)} - x\| = \gamma$, we have that $(X_{(1)}, Y_{(1)}, \ldots, X_{(k)}, Y_{(k)}) | (X_{(k+1)} = x_{(k+1)}) \overset{d}{=} (\tilde{X}_{(1)}, \tilde{Y}_{(1)}, \ldots, \tilde{X}_{(k)}, \tilde{Y}_{(k)})$, where $(\tilde{X}_{(1)}, \tilde{Y}_{(1)}), \ldots, (\tilde{X}_{(k)}, \tilde{Y}_{(k)})$ is a reordering of the independent and identically distributed pairs $(\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_k, \tilde{Y}_k)$ such that $\|\tilde{X}_{(1)} - x\| \leq \ldots \leq \|\tilde{X}_{(k)} - x\|$. Here $\tilde{X}_1 \overset{d}{=} X | (\|X - x\| \leq \gamma)$ and $\mathbb{P}(\tilde{Y}_1 = 1 | \tilde{X}_1 = x) = \eta(x)$. Writing $\tilde{S}_n(x) := \frac{1}{k} \sum_{i=1}^k \mathbb{1}_{\{\tilde{Y}_i = 1\}}$ we therefore have by Hoeffding's inequality that, for $x \in \mathcal{T}_n$, $\epsilon\beta \log n > 4$ and $\|x_{(k+1)} - x\| \in [2, x_2 - 1)$,

$$\begin{aligned} \mathbb{P}\{\hat{S}_n(x) < 1/2 \big| X_{(k+1)} = x_{(k+1)}\} &= \mathbb{P}\{\tilde{S}_n(x) < 1/2\} = \mathbb{P}\{\tilde{S}_n(x) - \mathbb{E}\tilde{S}_n(x) < -(\mathbb{E}\eta(\tilde{X}_1) - 1/2)\} \\ &\leq \exp\bigg(-2k\Big(\frac{2}{3}\frac{\sinh(3^{1/2})}{\sinh(2)} - \frac{1}{2}\Big)^2\bigg) = O(n^{-M}) \end{aligned}$$

for all $M > 0$, uniformly for $k \in K_\beta$. Writing $P_{(k+1)}$ for the marginal distribution of $X_{(k+1)}$, we deduce that

$$\begin{aligned} \mathbb{P}\{\hat{S}_n(x) < 1/2\} &\leq \mathbb{P}\{\hat{S}_n(x) < 1/2, \|X_{(k+1)} - x\| \in [2, x_2 - 1)\} + \mathbb{P}(B^c_{k,2}) \\ &= \int_{B_{x_2 - 1}(x) \setminus B_2(x)} \mathbb{P}\{\hat{S}_n(x) < 1/2 \big| X_{(k+1)} = x_{(k+1)}\}\, dP_{(k+1)}(x_{(k+1)}) + O(n^{-M}) = O(n^{-M}) \end{aligned}$$

for all $M > 0$, uniformly for $k \in K_\beta$. We conclude that for every $M > 0$,

$$\begin{aligned} R_{\mathcal{T}_n}(\hat{C}_n^{k\mathrm{nn}}) - R_{\mathcal{T}_n}(C^{\mathrm{Bayes}}) &= \int_{\mathcal{T}_n} \Big[\mathbb{P}\{\hat{S}_n(x) < 1/2\} - \mathbb{1}_{\{\eta(x) < 1/2\}}\Big] \{2\eta(x) - 1\} \bar{f}(x)\, dx \\ &= \int_{(1+\epsilon)\log(n/k)}^\infty \int_0^{1/2} \mathbb{P}\{\hat{S}_n(x) \geq 1/2\}(1 - 2x_1) x_1 \exp(-x_2)\, dx_1\, dx_2 = \frac{1}{24}\Big(\frac{k}{n}\Big)^{1+\epsilon} + O(n^{-M}), \end{aligned}$$

uniformly for $k \in K_\beta$, which establishes the claim (4.4). □

4.6.4 Proofs of results from Section 4.4

Proof of Theorem 4.2. Recall that

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

and define

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

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

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

Then there exists $n_0 \in \mathbb{N}$ such that for $n \geq n_0$ we have $\mathcal{R}_n \subseteq \bigl\{x \in \mathbb{R}^d : \bar{f}(x) \geq \delta_{n,\mathrm{O}}(x)\bigr\}$ and by Assumption **(A.1)** we then have that $\mathcal{R}_n$ is a d-dimensional manifold. There exists $n_1 \in \mathbb{N}$ such that for all $n \geq n_1$ and $x \in \mathcal{R}_n \cap \mathcal{S}^{\epsilon_0}$ we have that $k_{\mathrm{O}}(x) = \bigl\lfloor B\bigl\{\bar{f}(x)(n-1)\bigr\}^{4/(d+4)}\bigr\rfloor$. By **(A.2)**, we therefore have that $k_{\mathrm{O}} \in K_{\beta,\tau}$ for some $\tau = \tau_n$ with $\tau_n \searrow 0$. We deduce from Theorem 4.4 that

$$R(\hat{C}_n^{k_{\mathrm{O}}\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) = B_{3,n} n^{-4/(d+4)} + o(\gamma_n(k_{\mathrm{O}})) + O\bigl\{P_X\bigl((\partial\mathcal{R}_n)^{\epsilon_n} \cap \mathcal{S}^{\epsilon_n}\bigr) + P_X(\mathcal{R}_n^c)\bigr\}$$

as $n \to \infty$, where

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

By a similar argument to that in (4.41) we have that if $x \in (\partial\mathcal{R}_n)^{\epsilon_n} \cap \mathcal{S}^{\epsilon_n}$ then $\bar{f}(x) \leq 2(n-1)^{-(1-\alpha)}$. But, by Markov's inequality and Hölder's inequality, for $\tilde{\alpha} \in (0,1)$,

$$\begin{aligned}\mathbb{P}\{\bar{f}(X) \leq 2(n-1)^{-(1-\alpha)}\} &\leq \{2(n-1)^{-(1-\alpha)}\}^{\frac{\rho(1-\tilde{\alpha})}{\rho+d}} \int_{\mathbb{R}^d} \bar{f}(x)^{1-\frac{\rho(1-\tilde{\alpha})}{\rho+d}}\, dx \\ &\leq \{2(n-1)^{-(1-\alpha)}\}^{\frac{\rho(1-\tilde{\alpha})}{\rho+d}} \Bigl\{\int_{\mathbb{R}^d} (1+\|x\|^{\rho})\bar{f}(x)\, dx\Bigr\}^{1-\frac{\rho(1-\tilde{\alpha})}{\rho+d}} \\ &\qquad \Bigl\{\int_{\mathbb{R}^d} \frac{1}{(1+\|x\|^{\rho})^{(\rho+d)/\{\rho(1-\tilde{\alpha})\}-1}}\, dx\Bigr\}^{\frac{\rho(1-\tilde{\alpha})}{\rho+d}}. \end{aligned} \tag{4.43}$$

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 (4.43) to conclude that

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

Moreover, by very similar arguments to those given in the proof of Theorem 4.1, $\gamma_n(k_{\mathrm{O}}) = O(n^{-4/(d+4)})$ and $B_{3,n} \to B_3$ as $n \to \infty$. This concludes the proof of part (i).

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 (4.43) that

$$B_{3,n} n^{-4/(d+4)} + \gamma_n(k_{\mathrm{O}}) + P_X\bigl((\partial\mathcal{R}_n)^{\epsilon_n} \cap \mathcal{S}^{\epsilon_n}\bigr) + P_X(\mathcal{R}_n^c) = o\Bigl(\Bigl(\frac{1}{n}\Bigr)^{\frac{\rho}{\rho+d}-\beta-\epsilon}\Bigr),$$

for every $\epsilon > 0$. This proves part (ii). □

Proof of Theorem 4.3. We prove parts (i) and (ii) of the theorem simultaneously, by appealing to the corresponding arguments in the proof of Theorem 4.2. First, as in the proof of Theorem 4.2, for $\alpha \in \bigl((1+d/4)\beta, 1\bigr)$, 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

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

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

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

Now

$$\|\hat{f}_m - \bar{f}\|_\infty \leq \|\hat{f}_m - \mathbb{E}\hat{f}_m\|_\infty + \|\mathbb{E}\hat{f}_m - \bar{f}\|_\infty. \tag{4.44}$$

To bound the first term in (4.44), by Giné and Guillou (2002, Theorem 2.1), there exist $C, L > 0$, such that

$$\mathbb{P}(\|\hat{f}_m - \mathbb{E}\hat{f}_m\|_\infty \geq sm^{-\gamma/(d+2\gamma)}) \leq L\exp\Bigl(\frac{-\log(1+C/(4L))A^d s^2}{LC\|\bar{f}\|_\infty R(K)}\Bigr), \tag{4.45}$$

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

Then, by applying the bound in (4.45) with $s = s_0 := (n-1)^{\alpha/2}m_0^{\gamma/(d+2\gamma)}$, since $m \geq m_0(n-1)^{d/\gamma+2}$, we have that, for large n,

$$\begin{aligned}\mathbb{P}\Bigl\{\|\hat{f}_m - \mathbb{E}\hat{f}_m\|_\infty \geq \frac{1}{(n-1)^{1-\alpha/2}}\Bigr\} &= \mathbb{P}\Bigl\{\|\hat{f}_m - \mathbb{E}\hat{f}_m\|_\infty \geq s_0 m^{-\gamma/(d+2\gamma)}\Bigr\} \\ &\leq L\exp\Bigl(\frac{-\log(1+C/(4L))A^d(n-1)^\alpha m_0^{\gamma/(d+2\gamma)}}{LC\|\bar{f}\|_\infty R(K)}\Bigr) = O(n^{-M}),\end{aligned}$$

for all $M > 0$. For the second term in (4.44), by a Taylor expansion and **(A.5)**(γ), we have that, for all n sufficiently large,

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

It follows that $\mathbb{P}(\hat{f}_m \notin \mathcal{H}_{n,\tau}) = O(n^{-M})$ for all $M > 0$, with $\tau = 2(n-1)^{-\alpha/2}$.

Now, for $h \in \mathcal{H}_{n,\tau}$, let

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

Let $c_n := \sup_{x_0 \in \mathcal{S}:\bar{f}(x_0) \geq k_h(x_0)/(n-1)} g(x_0)$, and let

$$\delta_{n,h}(x) := \frac{k_h(x)}{n-1}c_n^d \log^d\Bigl(\frac{n-1}{k_h(x)}\Bigr).$$

Then there exists $n_0 \in \mathbb{N}$ such that for $n \geq n_0$ and $h \in \mathcal{H}_{n,\tau}$, we have $\mathcal{R}_n \subseteq \bigl\{x \in \mathbb{R}^d : \bar{f}(x) \geq \delta_{n,h}(x)\bigr\}$ and $k_h \in K_{\beta,\tau}$. We can therefore apply Theorem 4.4 (similarly to the application in the proof of Theorem 4.2) to conclude that for every $M > 0$,

$$\begin{aligned} R(\hat{C}_n^{k_h\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) =& B_{3,n} n^{-4/(d+4)}\{1 + o(1)\} + o(\gamma_n(k_h)) \\ &+ O\bigl\{P_X\bigl((\partial\mathcal{R}_n)^{\epsilon_n} \cap \mathcal{S}^{\epsilon_n}\bigr) + P_X(\mathcal{R}_n^c)\bigr\} + O(n^{-M}), \end{aligned}$$

uniformly for $h \in \mathcal{H}_{n,\tau}$, where $B_{3,n}$ was defined in the proof of Theorem 4.2. The proof of both parts (i) and (ii) is now completed by following the relevant steps in the proof of Theorem 4.2. □

4.7 Appendix: An introduction to differential geometry, tubular neighbourhoods and integration on manifolds

The purpose of this section is to give a brief introduction to the ideas from differential geometry, specifically tubular neighbourhoods and integration on manifolds, which play an important role in our analysis of misclassification error rates, but which we expect are unfamiliar to many statisticians. For further details and several of the proofs, we refer the reader to the many excellent texts on these topics, e.g. Guillemin and Pollack (1974), Gray (2004).

4.7.1 Manifolds and regular values

Recall that if $\mathcal{X}$ is an arbitrary subset of $\mathbb{R}^M$, we say $\phi : \mathcal{X} \to \mathbb{R}^N$ is *differentiable* if for each $x \in \mathcal{X}$, there exists an open subset $U \subseteq \mathbb{R}^M$ containing x and a differentiable function $F : U \to \mathbb{R}^N$ such that $F(z) = \phi(z)$ for $z \in U \cap \mathcal{X}$. If $\mathcal{Y}$ is also a subset of $\mathbb{R}^M$, we say $\phi : \mathcal{X} \to \mathcal{Y}$ is a *diffeomorphism* if ϕ is bijective and differentiable and if its inverse ϕ^{-1} is also differentiable. We then say $\mathcal{S} \subseteq \mathbb{R}^d$ is an *m-dimensional manifold* if for each $x \in \mathcal{S}$, there exist an open subset $U_x \subseteq \mathbb{R}^m$, a neighbourhood V_x of x in $\mathcal{S}$ and a diffeomorphism $\phi_x : U_x \to V_x$. Such a diffeomorphism ϕ_x is called a *local parametrisation* of $\mathcal{S}$ around x, and we sometimes suppress the dependence of ϕ_x, U_x and V_x on x. It turns out that the specific choice of local parametrisation is usually not important, and properties of the manifold are well-defined regardless of the choice made.

Let $\mathcal{S} \subseteq \mathbb{R}^d$ be an m-dimensional manifold and let $\phi : U \to \mathcal{S}$ be a local parametrisation of $\mathcal{S}$ around $x \in \mathcal{S}$, where U is an open subset of $\mathbb{R}^m$. Assume that $\phi(0) = x$ for convenience. The *tangent space* $T_x(\mathcal{S})$ to $\mathcal{S}$ at x is defined to be the image of the derivative $D\phi_0 : \mathbb{R}^m \to \mathbb{R}^d$ of ϕ at 0. Thus $T_x(\mathcal{S})$ is the m-dimensional subspace of $\mathbb{R}^d$ whose parallel translate $x + T_x(\mathcal{S})$ is the best affine approximation to $\mathcal{S}$ through x, and $(D\phi_0)^{-1}$ is well-defined as a map from $T_x(\mathcal{S})$ to $\mathbb{R}^m$. If $f : \mathcal{S} \to \mathbb{R}$ is differentiable, we define the derivative $Df_x : T_x(\mathcal{S}) \to \mathbb{R}$ of f at x by $Df_x := Dh_0 \circ (D\phi_0)^{-1}$, where $h := f \circ \phi$.

In practice, it is usually rather inefficient to define manifolds through explicit diffeomorphisms. Instead, we can often obtain them as level sets of differentiable functions. Suppose that $\mathcal{R} \subseteq \mathbb{R}^d$ is a manifold and $\eta : \mathcal{R} \to \mathbb{R}$ is differentiable. We say $y \in \mathbb{R}$ is a *regular value* for η if $\mathrm{image}(D\eta_x) = \mathbb{R}$ for every $x \in \mathcal{R}$ for which $\eta(x) = y$. If $y \in \mathbb{R}$ is a regular value of η, then $\eta^{-1}(y)$ is a $(d-1)$-dimensional submanifold of $\mathcal{R}$ (Guillemin and Pollack, 1974, p. 21).

4.7.2 Tubular neighbourhoods of level sets

For any set $\mathcal{S} \subseteq \mathbb{R}^d$ and $\epsilon > 0$, we call $\mathcal{S} + \epsilon B_1(0)$ the *ϵ-neighbourhood* of $\mathcal{S}$. In circumstances where $\mathcal{S}$ is a $(d-1)$-dimensional manifold defined by the level set of a continuously differentiable function

$\eta : \mathbb{R}^d \to \mathbb{R}$ with non-vanishing derivative on $\mathcal{S}$, the set $\mathcal{S}^\epsilon$ is often called a *tubular neighbourhood*, and $\dot{\eta}(x)^T v = 0$ for all $x \in \mathcal{S}$ and $v \in T_x(\mathcal{S})$. We therefore have the following useful representation of the ϵ-neighbourhood of $\mathcal{S}$ in terms of points on $\mathcal{S}$ and a perturbation in a normal direction.

Proposition 4.5. *Let $\eta : \mathbb{R}^d \to [0,1]$, suppose that $\mathcal{S} := \{x \in \mathbb{R}^d : \eta(x) = 1/2\}$ is non-empty, and suppose further that η is continuously differentiable on $\mathcal{S} + \epsilon B_1(0)$ for some $\epsilon > 0$, with $\dot{\eta}(x) \neq 0$ for all $x \in \mathcal{S}$, so that $\mathcal{S}$ is a $(d-1)$-dimensional manifold. Then*

$$\mathcal{S} + \epsilon B_1(0) = \Bigl\{x_0 + \frac{t\dot{\eta}(x_0)}{\|\dot{\eta}(x_0)\|} : x_0 \in \mathcal{S}, |t| < \epsilon\Bigr\} =: \mathcal{S}^\epsilon.$$

Proof. For any $x_0 \in \mathcal{S}$ and $|t| < \epsilon$, we have $x_0 + t\dot{\eta}(x_0)/\|\dot{\eta}(x_0)\| \in \mathcal{S} + \epsilon B_1(0)$. On the other hand, suppose that $x \in \mathcal{S} + \epsilon B_1(0)$. Since $\mathcal{S}$ is closed, there exists $x_0 \in \mathcal{S}$ such that $\|x - x_0\| \leq \|x - y\|$ for all $y \in \mathcal{S}$. Rearranging this inequality yields that, for $y \neq x_0$,

$$2(x - x_0)^T \frac{(y - x_0)}{\|y - x_0\|} \leq \|y - x_0\|. \tag{4.46}$$

Let U be an open subset of $\mathbb{R}^{d-1}$ and $\phi : U \to \mathcal{S}$ be a local parametrisation of $\mathcal{S}$ around x_0, where without loss of generality we assume $\phi(0) = x_0$. Let $v \in T_{x_0}(\mathcal{S}) \setminus \{0\}$ be given and let $h \in \mathbb{R}^{d-1} \setminus \{0\}$ be such that $D\phi_0(h) = v$. Then for $t > 0$ sufficiently small we have $th \in U$, so by (4.46),

$$2(x - x_0)^T \frac{\{\phi(th) - \phi(0)\}}{\|\phi(th) - \phi(0)\|} \leq \|\phi(th) - \phi(0)\|.$$

Letting $t \searrow 0$ we see that $(x - x_0)^T v \leq 0$. Since $v \in T_{x_0}(\mathcal{S}) \setminus \{0\}$ was arbitrary and $-v \in T_{x_0}(\mathcal{S}) \setminus \{0\}$, we therefore have that $(x - x_0)^T v = 0$ for all $v \in T_{x_0}(\mathcal{S})$. Moreover, $\dot{\eta}(x_0)^T v = 0$ for all $v \in T_{x_0}(\mathcal{S})$, so $x - x_0 \propto \dot{\eta}(x_0)$, which yields the result. □

In fact, under a slightly stronger condition on η, we have the following useful result:

Proposition 4.6. *Let $\mathcal{R}$ be a d-dimensional manifold in $\mathbb{R}^d$, suppose that $\eta : \mathcal{R} \to [0,1]$ satisfies the condition that $\mathcal{S} := \{x \in \mathcal{R} : \eta(x) = 1/2\}$ is non-empty. Suppose further that there exists $\epsilon > 0$ such that η is twice continuously differentiable on $\mathcal{S}^\epsilon$. Assume that $\dot{\eta}(x_0) \neq 0$ for all $x_0 \in \mathcal{S}$. Define $g : \mathcal{S} \times (-\epsilon, \epsilon) \to \mathcal{S}^\epsilon$ by*

$$g(x_0, t) := x_0 + \frac{t\dot{\eta}(x_0)}{\|\dot{\eta}(x_0)\|}.$$

If

$$\epsilon \leq \inf_{x_0 \in \mathcal{S}} \frac{\|\dot{\eta}(x_0)\|}{\sup_{z \in B_{2\epsilon}(x_0) \cap \mathcal{S}^\epsilon} \|\ddot{\eta}(z)\|_{\mathrm{op}}}, \tag{4.47}$$

then g is injective. In fact g is a diffeomorphism, with

$$Dg_{(x_0,t)}(v_1, v_2) = (I + tB)\biggl(v_1 + \frac{\dot{\eta}(x_0)}{\|\dot{\eta}(x_0)\|}v_2\biggr), \tag{4.48}$$

for $v_1 \in T_{x_0}(\mathcal{S})$ and $v_2 \in \mathbb{R}$, where

$$B := \frac{1}{\|\dot{\eta}(x_0)\|}\biggl(I - \frac{\dot{\eta}(x_0)\dot{\eta}(x_0)^T}{\|\dot{\eta}(x_0)\|^2}\biggr)\ddot{\eta}(x_0). \tag{4.49}$$

Proof. Assume for a contradiction that there exist distinct points $x_1, x_2 \in \mathcal{S}$ and $t_1, t_2 \in (-\epsilon, \epsilon)$

with $|t_1| \geq |t_2|$ such that

$$x_1 + \frac{t_1\dot{\eta}(x_1)}{\|\dot{\eta}(x_1)\|} = x_2 + \frac{t_2\dot{\eta}(x_2)}{\|\dot{\eta}(x_2)\|}.$$

Then

$$0 < \|x_2 - x_1\|^2 = \frac{2t_1\dot{\eta}(x_1)^T(x_2 - x_1)}{\|\dot{\eta}(x_1)\|} + t_2^2 - t_1^2 \leq \frac{2t_1\dot{\eta}(x_1)^T(x_2 - x_1)}{\|\dot{\eta}(x_1)\|}. \tag{4.50}$$

By Taylor's theorem and (4.50),

$$\begin{aligned}|\dot{\eta}(x_1)^T(x_2 - x_1)| &= |\eta(x_2) - \eta(x_1) - \dot{\eta}(x_1)^T(x_2 - x_1)| \\ &\leq \frac{1}{2}\sup_{z \in B_{2\epsilon}(x_1)\cap\mathcal{S}^\epsilon} \|\ddot{\eta}(z)\|_{\mathrm{op}}\|x_2 - x_1\|^2 < \sup_{z \in B_{2\epsilon}(x_1)\cap\mathcal{S}^\epsilon} \|\ddot{\eta}(z)\|_{\mathrm{op}} \frac{\epsilon|\dot{\eta}(x_1)^T(x_2 - x_1)|}{\|\dot{\eta}(x_1)\|},\end{aligned}$$

contradicting the hypothesis (4.47).

To show that g is a diffeomorphism, let $x_0 \in \mathcal{S}$ be given and let $\phi : U \to \mathcal{S}$ be a local parametrisation around x_0 with $\phi(0) = x_0$. Define $\Phi : U \times (-\epsilon, \epsilon) \to \mathcal{S} \times (-\epsilon, \epsilon)$ by $\Phi(u, t) := (\phi(u), t)$, and $H : U \times (-\epsilon, \epsilon) \to \mathcal{S}^\epsilon$ by $H := g \circ \Phi$. Finally, define the *Gauss map* $n : \mathcal{S} \to \mathbb{R}^d$ by $n(x_0) := \dot{\eta}(x_0)/\|\dot{\eta}(x_0)\|$. Then, for $h = (h_1^T, h_2)^T \in \mathbb{R}^{d-1} \times \mathbb{R}$ and $s \in \mathbb{R} \setminus \{0\}$,

$$\begin{aligned}\lim_{s\to 0} \frac{H(sh_1, t + sh_2) - H(0, t)}{s} &= \lim_{s \to 0}\biggl\{\frac{\phi(sh_1) - \phi(0)}{s} + \frac{t\{n(\phi(sh_1)) - n(\phi(0))\}}{s} + h_2 n\bigl(\phi(sh_1)\bigr)\biggr\} \\ &= D\phi_0(h_1) + tDn_{x_0} \circ D\phi_0(h_1) + h_2 n(x_0) = Dg_{(x_0,t)} \circ D\Phi_{(0,t)}(h_1, h_2),\end{aligned}$$

where $Dg_{(x_0,t)} : T_{x_0}(\mathcal{S}) \times \mathbb{R} \to \mathbb{R}^d$ is given in (4.48).

To show that $Dg_{(x_0,t)}$ is invertible, note that for $v_1 \in T_{x_0}(\mathcal{S})$ and $|t| < \epsilon$,

$$\frac{|t|}{\|\dot{\eta}(x_0)\|}\biggl\|\biggl(I - \frac{\dot{\eta}(x_0)\dot{\eta}(x_0)^T}{\|\dot{\eta}(x_0)\|^2}\biggr)\ddot{\eta}(x_0)v_1\biggr\| \leq \frac{|t|\|\ddot{\eta}(x_0)\|_{\mathrm{op}}}{\|\dot{\eta}(x_0)\|}\|v_1\| < \|v_1\|,$$

where the final inequality follows from (4.47). Then, since $v_1 + \frac{t}{\|\dot{\eta}(x_0)\|}\Bigl(I - \frac{\dot{\eta}(x_0)\dot{\eta}(x_0)^T}{\|\dot{\eta}(x_0)\|^2}\Bigr)\ddot{\eta}(x_0)v_1$ and $n(x_0)v_2$ are orthogonal, it follows that $Dg_{(x_0,t)}$ is indeed invertible. The inverse function theorem (e.g. Guillemin and Pollack, 1974, p. 13) then gives that g is a local diffeomorphism, and moreover, by Guillemin and Pollack (1974, Exercise 5, p. 18) and the fact that g is bijective, we can conclude that g is in fact a diffeomorphism. □

4.7.3 Forms, pullbacks and integration on manifolds

Let V be a (real) vector space of dimension m. We say $T : V^p \to \mathbb{R}$ is a *p-tensor* on V if it is p-linear, and write $\mathcal{F}^p(V^*)$ for the set of p-tensors on V. If $T \in \mathcal{F}^p(V^*)$ and $S \in \mathcal{F}^q(V^*)$, we define their *tensor product* $T \otimes S \in \mathcal{F}^{p+q}(V^*)$ by

$$T \otimes S(v_1, \ldots, v_p, v_{p+1}, \ldots, v_{p+q}) := T(v_1, \ldots, v_p)S(v_{p+1}, \ldots, v_{p+q}).$$

Let S_p denote the set of permutations of $\{1, \ldots, p\}$. If $\pi \in S_p$ and $T \in \mathcal{F}^p(V^*)$, we can define $T^\pi \in \mathcal{F}^p(V^*)$ by $T^\pi(v) := T(v_{\pi(1)}, \ldots, v_{\pi(p)})$ for $v = (v_1, \ldots, v_p) \in V^p$. We say T is *alternating* if $T^\sigma = -T$ for all transpositions $\sigma : \{1, \ldots, p\} \to \{1, \ldots, p\}$. The set of alternating p-tensors on V, denoted $\Lambda^p(V^*)$, is a vector space of dimension $\binom{m}{p}$. The function Alt : $\mathcal{F}^p(V^*) \to \Lambda^p(V^*)$ is defined by

$$\mathrm{Alt}(T) := \frac{1}{p!} \sum_{\pi \in S_p} (-1)^{\mathrm{sgn}(\pi)} T^{\pi},$$

where $\mathrm{sgn}(\pi)$ denotes the sign of the permutation π. If $T \in \Lambda^p(V^*)$ and $S \in \Lambda^q(V^*)$, we define their *wedge product* $T \wedge S \in \Lambda^{p+q}(V^*)$ by

$$T \wedge S := \mathrm{Alt}(T \otimes S).$$

If W is another (real) vector space and $A : V \to W$ is a linear map, we define the *transpose* $A^* : \Lambda^p(W^*) \to \Lambda^p(V^*)$ of A by

$$A^*T(v_1, \ldots, v_p) := T(Av_1, \ldots, Av_p).$$

Let $\mathcal{S}$ be a manifold. A *p-form* ω on $\mathcal{S}$ is a function which assigns to each $x \in \mathcal{S}$ an element $\omega(x) \in \Lambda^p(T_x(\mathcal{S})^*)$. If ω is a p-form on $\mathcal{S}$ and θ is a q-form on $\mathcal{S}$, we can define their wedge product $\omega \wedge \theta$ by $(\omega \wedge \theta)(x) := \omega(x) \wedge \theta(x)$. For $j = 1, \ldots, m$, let $x_j : \mathbb{R}^m \to \mathbb{R}$ denote the coordinate function $x_j(y_1, \ldots, y_m) := y_j$. These functions induce 1-forms dx_j, given by $dx_j(x)(y_1, \ldots, y_m) = y_j$ (so $dx_j(x) = D(x_j)_x$ in our previous notation). Letting $\mathcal{I} := \{(i_1, \ldots, i_p) : 1 \leq i_1 < \ldots < i_p \leq m\}$, for $I = (i_1, \ldots, i_p) \in \mathcal{I}$, we write

$$dx_I := dx_{i_1} \wedge \ldots \wedge dx_{i_p}.$$

It turns out (Guillemin and Pollack, 1974, p. 163) that any p-form on an open subset U of $\mathbb{R}^m$ can be uniquely expressed as

$$\sum_{I \in \mathcal{I}} f_I \, dx_I, \tag{4.51}$$

where each f_I is a real-valued function on U.

Recall that the set of all ordered bases of a vector space V is partitioned into two equivalence classes, and an *orientation* of V is simply an assignment of a positive sign to one equivalence class and a negative sign to the other. If V and W are oriented vector spaces in the sense that an orientation has been specified for each of them, then an isomorphism $A : V \to W$ always either preserves orientation in the sense that for any ordered basis β of V, the ordered basis $A\beta$ has the same sign as β, or it reverses it. We say an m-dimensional manifold $\mathcal{X}$ is *orientable* if for every $x \in \mathcal{X}$, there exist an open subset U of $\mathbb{R}^m$, a neighbourhood V of x in $\mathcal{X}$ and a diffeomorphism $\phi : U \to V$ such that $D\phi_u : \mathbb{R}^m \to T_x(\mathcal{X})$ preserves orientation for every $u \in U$. A map like ϕ above whose derivative at every point preserves orientation is called an *orientation-preserving* map.

If $\mathcal{X}$ and $\mathcal{Y}$ are manifolds, ω is a p-form on $\mathcal{Y}$ and $\psi : \mathcal{X} \to \mathcal{Y}$ is differentiable, we define the *pullback* $\psi^*\omega$ of ω by ψ to be the p-form on $\mathcal{X}$ given by

$$\psi^*\omega(x) := (D\psi_x)^*\omega\bigl(\psi(x)\bigr).$$

If V is an p-dimensional vector space and $A : V \to V$ is linear, then $A^*T = (\det A)T$ for all $T \in \Lambda^p(V)$ (Guillemin and Pollack, 1974, p. 160).

If ω is an m-form on an open subset U of $\mathbb{R}^m$, then by (4.51), we can write $\omega = f \, dx_1 \wedge \ldots \wedge dx_m$. If ω is an integrable form on U (i.e. f is an integrable function on U), we can define the integral of ω over U by

$$\int_U \omega := \int_U f(x_1, \ldots, x_m)\, dx_1 \ldots dx_m,$$

where the integral on the right-hand side is a usual Lebesgue integral. Now let $\mathcal{S}$ be an m-dimensional orientable manifold that can be parametrised with a single chart, in the sense that there exists an open subset U of $\mathbb{R}^m$ and an orientation-preserving diffeomorphism $\phi : U \to \mathcal{S}$. Define the *support* of an m-form ω on $\mathcal{S}$ to be the closure of $\{x \in \mathcal{S} : \omega(x) \neq 0\}$. If ω is compactly supported, then its pullback $\phi^*\omega$ is a compactly supported m-form on U; moreover $\phi^*\omega$ is integrable, and we can define the integral over $\mathcal{S}$ of ω by

$$\int_{\mathcal{S}} \omega := \int_U \phi^*\omega. \tag{4.52}$$

Alternatively, we can suppose that ω is non-negative and measurable in the sense that $\phi^*\omega = f\, dx_1 \wedge \ldots \wedge dx_m$, say, with f non-negative and measurable on U. In this case, we can also define the integral of ω over $\mathcal{S}$ via (4.52).

More generally, integrals of forms over more complicated manifolds can be defined via partitions of unity. Recall (Guillemin and Pollack, 1974, p. 52) that if $\mathcal{X}$ is an arbitrary subset of $\mathbb{R}^M$, and $\{V_\alpha : \alpha \in A\}$ is a (relatively) open cover of $\mathcal{X}$, then there exists a sequence of real-valued, differentiable functions (ρ_n) on $\mathcal{X}$, called a *partition of unity* with respect to $\{V_\alpha : \alpha \in A\}$, with the following properties:

1. $\rho_n(x) \in [0, 1]$ for all $n \in \mathbb{N}$;
2. Each $x \in \mathcal{X}$ has a neighbourhood on which all but finitely many functions ρ_n are identically zero;
3. Each ρ_n is identically zero except on some closed set contained in some V_α;
4. $\sum_{n=1}^{\infty} \rho_n(x) = 1$ for all $x \in \mathcal{X}$.

Now let $\mathcal{S} \subseteq \mathbb{R}^d$ be an m-dimensional, orientable manifold, so for each $x \in \mathcal{S}$, there exist an open subset U_x of $\mathbb{R}^m$, a neighbourhood V_x of x in $\mathcal{S}$ and an orientation-preserving diffeomorphism $\phi_x : U_x \to V_x$. If ω is a compactly supported m-form on $\mathcal{S}$ and (ρ_n) denotes a partition of unity on $\mathcal{S}$ with respect to $\{V_x : x \in \mathcal{S}\}$, we can define the integral of ω over $\mathcal{S}$ by

$$\int_{\mathcal{S}} \omega := \sum_{n=1}^{\infty} \int_{\mathcal{S}} \rho_n \omega. \tag{4.53}$$

In fact, writing Ω for the compact support of ω, we can find a neighbourhood W_x of $x \in \Omega$, $x_1, \ldots, x_N \in \Omega$ and a finite subset N_j of $\mathbb{N}$ such that $\{\rho_n : n \notin N_j\}$ are identically zero on W_{x_j}, and such that

$$\int_{\mathcal{S}} \omega = \sum_{j=1}^{N} \sum_{n \in N_j} \int_{\mathcal{S}} \rho_n \omega.$$

Thus the integral can be written as a finite sum. Similarly, if ω is a non-negative m-form on $\mathcal{S}$, we can again define the integral of ω over $\mathcal{S}$ via (4.53). Finally, if ω is an integrable m-form on $\mathcal{S}$, the integral can be defined by taking positive and negative parts in the usual way.

In our work, we are especially interested in integrals of a particular type of form. Given an m-dimensional, orientable manifold $\mathcal{S}$ in $\mathbb{R}^d$, the *volume form* $d\mathrm{Vol}^m$ is the unique m-form on $\mathcal{S}$ such that at each $x \in \mathcal{S}$, the alternating m-tensor $d\mathrm{Vol}^m(x)$ on $T_x(\mathcal{S})$ gives value $1/m!$ to each positively oriented orthonormal basis for $T_x(\mathcal{S})$. For example, when $\mathcal{S} = \mathbb{R}^m$, we have $d\mathrm{Vol}^m = dx_1 \wedge \ldots \wedge dx_m$, provided we consider the standard basis to be positively oriented. As another example, if $\mathcal{R} \subseteq \mathbb{R}^d$ is a d-dimensional manifold and $\eta : \mathcal{R} \to \mathbb{R}$ is continuously differentiable with $\mathcal{S} = \{x \in \mathcal{R} : \eta(x) = 1/2\}$ non-empty and $\dot{\eta}(x) \neq 0$ for $x \in \mathcal{S}$, then $\mathcal{S}$ is a $(d-1)$-dimensional, orientable manifold (Guillemin and Pollack, 1974, Exercise 18, p. 106). If we say that an ordered, orthonormal basis $e_1, \ldots, e_{d-1}$ for $T_{x_0}(\mathcal{S})$ is positively oriented whenever $\det(e_1, \ldots, e_{d-1}, \dot{\eta}(x_0)) > 0$, we have that

$$d\mathrm{Vol}^{d-1}(x_0) = \sum_{j=1}^{d} (-1)^{j+d} \frac{\eta_j(x_0)}{\|\dot{\eta}(x_0)\|} dx_1 \wedge \ldots \wedge dx_{j-1} \wedge dx_{j+1} \wedge \ldots \wedge dx_d(x_0),$$

where x_j denotes the jth coordinate function. We now define an ordered, orthonormal basis $(e_1, 0), \ldots, (e_{d-1}, 0), (0, 1)$ for $T_{x_0}(\mathcal{S}) \times \mathbb{R}$ to be positively oriented. Further, we define a $(d-1)$-form ω_1 and a 1-form ω_2 on $\mathcal{S} \times (-\epsilon, \epsilon)$ by

$$\begin{aligned} \omega_1(x_0, t)\big((v_1, w_1), \ldots, (v_{d-1}, w_{d-1})\big) &:= d\mathrm{Vol}^{d-1}(x_0)(v_1, \ldots, v_{d-1}) \\ \omega_2(x_0, t)(v_d, w_d) &:= dt(t)(w_d) = w_d. \end{aligned}$$

Then, with g defined as in Proposition 4.6, and under the conditions of that proposition,

$$\begin{aligned} &g^*(dx_1 \wedge \ldots \wedge dx_d)(x_0, t)\big((e_1, 0), \ldots, (e_{d-1}, 0), (0, 1)\big) \\ &\quad = dx_1 \wedge \ldots \wedge dx_d(x_0^t)\big(Dg_{(x_0,t)}(e_1, 0), \ldots, Dg_{(x_0,t)}(e_{d-1}, 0), Dg_{(x_0,t)}(0, 1)\big) \\ &\quad = \frac{1}{d!} \det(I + tB) \\ &\quad = \frac{1}{d} \det(I + tB) d\mathrm{Vol}^{d-1}(x_0)(e_1, \ldots, e_{d-1}) dt(t)(1) \\ &\quad = \det(I + tB)\ (\omega_1 \wedge \omega_2)(x_0, t)\big((e_1, 0), \ldots, (e_{d-1}, 0), (0, 1)\big), \end{aligned}$$

so $g^*(dx_1 \wedge \ldots \wedge dx_d)(x_0, t) = \det(I + tB)\ (\omega_1 \wedge \omega_2)(x_0, t)$. It follows that if $h : \mathcal{S} \times (-\epsilon, \epsilon) \to \mathbb{R}$ is either compactly supported and integrable, or non-negative and measurable, then

$$\int_{\mathcal{S} \times (-\epsilon, \epsilon)} h\, \omega_1 \wedge \omega_2 = \int_{\mathcal{S}} \int_{-\epsilon}^{\epsilon} h(x_0, t)\, dt\, d\mathrm{Vol}^{d-1}(x_0). \tag{4.54}$$

Finally, we require the change of variables formula: if $\mathcal{X}$ and $\mathcal{Y}$ are orientable manifolds and are of dimension m, and if $\psi : \mathcal{X} \to \mathcal{Y}$ is an orientation-preserving diffeomorphism, then

$$\int_{\mathcal{X}} \psi^* \omega = \int_{\mathcal{Y}} \omega \tag{4.55}$$

for every compactly supported, integrable m-form on $\mathcal{Y}$ (Guillemin and Pollack, 1974, p. 168). In particular, if $f : \mathcal{S}^\epsilon \to \mathbb{R}$ is either compactly supported and integrable, or non-negative and measurable, then writing $x_0^t := x_0 + \frac{t\dot{\eta}(x_0)}{\|\dot{\eta}(x_0)\|}$, we have from (4.54) and (4.55) that

$$\begin{aligned}\int_{\mathcal{S}^\epsilon} f(x)\,dx &= \int_{\mathcal{S}\times(-\epsilon,\epsilon)} \det(I+tB)f(x_0^t)\,(\omega_1\wedge\omega_2)(x_0,t)\\ &= \int_{\mathcal{S}}\int_{-\epsilon}^{\epsilon} \det(I+tB)f(x_0^t)\,dt\,d\mathrm{Vol}^{d-1}(x_0). \qquad (4.56)\end{aligned}$$

Bibliography

Albert, M., Bouret, Y., Fromont, M. and Reynaud-Bouret, P. (2015) Bootstrap and permutation tests of independence for point processes. *Ann. Statist.*, **43**, 2537–2564.

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

Bach, F. R. and Jordan, M. I. (2002) Kernel independent component analysis. *J. Mach. Learn. Res.*, **3**, 1–48.

Bai, F., Jiang, D., Yao, J. and Zheng, S. (2009) Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Statist.*, **37**, 3822–3840.

Baryshnikov, Y., Penrose, M. D. and Yukich, J. E. (2009) Gaussian limits for generalised spacings. *Ann. Appl. Probab.*, **19**, 158–185.

Beirlant, J., Dudewicz, E. J., Györfi, L., and Van der Meulen, E. C. (1997) Nonparametric entropy estimation: an overview. *Int. J. Math. Stat. Sci.*, **6**, 17–39.

Berrett, T. B., Samworth, R. J. and Yuan, M. (2017) Efficient multivariate entropy estimation via k-nearest neighbour distances. `https://arxiv.org/abs/1606.00304`.

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.

Biau, G. and Devroye, L. (2015) *Lectures on the Nearest Neighbor Method.* Springer, New York.

Boucheron, S., Bousquet, O. and Lugosi, G. (2005) Theory of classification: a survey of some recent advances. *ESAIM: PS*, **9**, 323–375.

Boucheron, S., Lugosi, G. and Massart, P. (2013) *Concentration Inequalities.* Oxford University Press, Oxford.

Buchen, P. W. and Kelly, M. (1996) The maximum entropy distribution of an asset inferred from option prices. *J. Financ. Quant. Anal.*, **31**, 143–159.

Cai, T. T. and Low, M. G. (2011) Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *Ann. Statist.*, **39**, 1012–1041.

Cannings, T. I., Berrett, T. B. and Samworth, R. J. (2017) Local nearest neighbour classification with applications to semi-supervised learning. `https://arxiv.org/abs/1704.00642`.

Celisse, A. and Mary-Huard, T. (2015) New upper bounds on cross-validation for the k-nearest neighbor classification rule. `https://arxiv.org/abs/1508.04905`.

Chandola, V., Banerjee, A. and Kumar, V. (2009) Anomaly detection: a survey. *ACM Comput. Surv.*, **41**, 15.

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. *NIPS*, **27**, 3437–3445.

Chen, J. and Shao, J. (2000) Nearest neighbor imputation for survey data. *J. Off. Stat.*, **16**, 113–131.

Comon, P. (1994) Independent component analysis, a new concept?. *Signal Process.*, **36**, 287–314.

Costa, J. A. and Hero, A. O. (2004) Manifold learning using Euclidean k-nearest neighbour graphs [image processing examples]. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, **3**, 988–991.

Cover, T. M. and Hart, P. E. (1967) Nearest neighbour pattern classification. *IEEE Trans. Inf. Th.*, **13**, 21–27.

Cover, T. M. and Thomas, J. A. (2012) *Elements of information theory.* John Wiley & Sons.

Cressie, N. (1976) On the logarithms of high-order spacings. *Biometrika*, **63**, 343–355.

Dasgupta, S. and Kpotufe, S. (2014) Optimal rates for k-NN density and mode estimation. *NIPS*, **27**, 2555–2563.

Delattre, S. and Fournier, N. (2017) On the Kozachenko–Leonenko entropy estimator. *J. Statist. Plann. Inf.*, **185**, 69–93.

Devroye, L., Gyorfi, L. and Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition.* Springer, New York.

Duda, R. O. and Hart, P. E. (1973) *Pattern Classification and Scene Analysis.* Wiley, New York.

Dudley, R. M. (1999) *Uniform Central Limit Theorems.* Cambridge University Press, Cambridge.

Duong, T. (2015) `ks`: Kernel smoothing. `R` package version 1.9.4, `https://cran.r-project.org/web/packages/ks`.

Duong, T., Beck, G., Azzag, H and Lebbah, M. (2016) Nearest neighbour estimators of density derivatives, with application to mean shift clustering. *Pattern Recogn. Lett.*, **80**, 224–230.

El Haje Hussein, F. and Golubev, Y. (2009) On entropy estimation by m-spacing method. *J. Math. Sci.*, **163**, 290–309.

Evans, D., Jones, A. J. and Schmidt, W. M. (2002) Asymptotic moments of near-neighbour distance distributions. *Proc. R. Soc. Lond. A*, **458**, 2839–2849.

Fan, J., Feng, Y. and Xia, L. (2017) A projection based conditional dependence measure with applications to high-dimensional undirected graphical models. Available at `arXiv:1501.01617`.

Fix, E. and Hodges, J. L. (1951) Discriminatory analysis – nonparametric discrimination: Consistency properties. Technical Report number 4, USAF School of Aviation Medicine, Randolph Field, Texas.

Fix, E. and Hodges, J. L. (1989) Discriminatory analysis – nonparametric discrimination: Consistency properties. *Internat. Statist. Rev.*, **57**, 238–247.

Friedman, J. H., Bentley, J. L. and Finkel, R. A. (1977) An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Softw.*, **3**, 209–226.

Gadat, S., Klein, T. and Marteau, C. (2016) Classification with the nearest neighbour rule in general finite dimensional spaces. *Ann. Statist.*, **44**, 982–1001.

Gao, W., Oh, S. and Viswanath, P. (2016) Demystifying fixed k-nearest neighbor information estimators. Available at `arXiv:1604.03006`.

Gibbs, A. L. and Su, F. E. (2002) On choosing and bounding probability metrics. *Int. Statist. Review*, **70**, 419–435.

Giné, E. and Guillou, A. (2002) Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, **38**, 907-921.

Goria, M. N., Leonenko, N. N., Mergel, V. V. and Novi Inverardi, P. L. (2005) A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *J. Nonparametr. Stat.*, **17**, 277–297.

Götze, F. (1991) On the rate of convergence in the multivariate CLT. *Ann. Prob.*, **19**, 724–739.

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

Gretton A., Bousquet O., Smola A. and Schölkopf B. (2005) Measuring Statistical Dependence with Hilbert-Schmidt Norms. *Algorithmic Learning Theory*, **16**, 63–78.

Gretton, A. and Györfi, L. (2010) Consistent nonparametric tests of independence. *J. Mach. Learn. Res.*, **11**, 1391–1423.

Guillemin, V. and Pollack, A. (1974) *Differential Geometry.* Prentice-Hall, New Jersey.

Hall, P. and Morton, S. C. (1993) On the estimation of entropy. *Ann. Inst. Statist. Math.*, **45**, 69–88.

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

Hall, P. and Samworth, R. J. (2005) Properties of bagged nearest-neighbour classifiers. *J. Roy. Statist. Soc., Ser. B*, **67**, 363–379.

Hall, P., Park, B. U. and Samworth, R. J. (2008) Choice of neighbour order in nearest-neighbour classification. *Ann. Statist.*, **36**, 2135–2152.

Heckel, R. and Bölcskei, H. (2015) Robust subspace clustering via thresholding. *IEEE Trans. Info. Th.*, **61**, 6320–6342.

Heller, R., Heller, Y., Kaufman, S., Brill, B. and Gorfine, M. (2016) Consistent distribution-free K-sample and independence tests for univariate random variables. *J. Mach. Learn. Res.*, **17**, 1–54.

Hoeffding, W. (1948) A non-parametric test of independence. *Ann. Math. Statist.*, **19**, 546–557.

Ibragimov, I. A. and Khas'minskii, R. Z. (1991) Asymptotically normal families of distributions and efficient estimation. *Ann. Statist.*, **19**, 1681–1724.

Jaynes, E. T. (1968) Prior probabilities. *IEEE Trans. Syst. Sci. Cybern.*, **4**, 227–241.

Jitkrittum, W., Szabó, Z. and Gretton, A. (2016) An adaptive test of independence with analytic kernel embeddings. Available at `arXiv:1610.04782`.

Joe, H. (1989) Relative entropy measures of multivariate dependence. *J. Amer. Statist. Assoc.*, **84**, 157–164.

Josse, J. and Holmes, S. (2014) Tests of independence and beyond. Available at `arXiv:1307.7383`.

Kozachenko, L. F. and Leonenko, N. N. (1987) Sample estimate of the entropy of a random vector. *Probl. Inform. Transm.*, **23**, 95–101.

Kraskov, A., Stögbauer H. and Grassberger, P. (2004) Estimating mutual information. *Phys. Rev. E*, **69**, 066138.

Kulkarni, S. R. and Posner, S. E. (1995) Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Trans. Info. Th.*, **41**, 1028–1039.

Kwak, N. and Choi, C. (2002) Input feature selection by mutual information based on Parzen window. *IEEE Trans. Pattern Anal. Mach. Intell.*, **24**, 1667–1671.

Laurent, B. (1996) Efficient estimation of integral functionals of a density. *Ann. Statist.*, **24**, 659–681.

Lauritzen, S. L. (1996) *Graphical Models.* Clarendon Press.

Law, M. H. C. and Jain, A. K. (2006) Incremental nonlinear dimensionality reduction by manifold learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, **28**, 377–391.

Leonenko, N., Pronzato, L. and Vippal, S. (2008) A class of Rényi information estimators for multidimensional densities. *Ann. Statist.*, **36**, 2153–2182.

Lepski, O., Nemirovski, A. and Spokoiny, V. (1999) On estimation of the L_r norm of a regression function. *Probab. Th. Rel. Fields*, **113**, 221–253.

Mack, Y. P. and Rosenblatt, M. (1979) Multivariate k-nearest neighbor density estimates. *J. Multivar. Anal.*, **9**, 1–15.

Mammen, E. and Tsybakov, A. B. (1999) Smooth discriminant analysis. *Ann. Statist.*, **27**, 1808–1829.

Mari, D. D. and Kotz, S. (2001) *Correlation and Dependence.* World Scientific.

Miller, E. G. and Fisher, J. W. (2003) ICA using spacings estimates of entropy. *J. Mach. Learn. Res.*, **4**, 1271–1295.

Mnatsakanov, R. M., Misra, N., Li, S. and Harner, E. J. (2008) K_n-nearest neighbor estimators of entropy. *Math. Methods Statist.*, **17**, 261–277.

Moon, K. R., Sricharan, K., Greenewald, K. and Hero, A. O. (2016) Nonparametric ensemble estimation of distributional functionals. `https://arxiv.org/abs/1601.06884v2`.

Muja, M. and Lowe, D. G. (2014) Scalable nearest neighbour algorithms for high dimensional data. *IEEE Trans. Pattern Anal. Mach. Intell.*, **36**, 2227-2240.

Nguyen, D. and Eisenstein, J. (2017) A kernel independence test for geographical language variation. *Comput. Ling., to appear.*

Paditz, L. (1989) On the analytical structure of the constant in the nonuniform version of the Esseen inequality. *Statistics*, **20**, 453–464.

Paninski, L. (2003) Estimation of entropy and mutual information. *Neural Comput.*, **15**, 1191–1253.

Paninski, L. and Yajima, M. (2008) Undersmoothed kernel entropy estimators. *IEEE Trans. Inf. Theory*, **54**, 4384–4388.

Peng, H., Long, F. and Ding, C. (2005) Feature selection based on mutual information: criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**, 1226–1238.

Romano, J. P. (1990) On the behavior of randomization tests without a group invariance assumption. J. Amer. Statist. Assoc., **85**, 686–692.

Roweis, S. T. and Saul, L. K. (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**, 2323-2326.

Samworth, R. J. (2012) Optimal weighted nearest neighbour classifiers. *Ann. Statist.*, **40**, 2733–2763.

Samworth, R. J. and Yuan, M. (2012) Independent component analysis via nonparametric maximum likelihood estimation. *Ann. Statist.*, **40**, 2973–3002.

Sejdinovic, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2013) Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.*, **41**, 2263–2291.

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

Schweizer, B. and Wolff, E. F. (1981) On nonparametric measures of dependence for random variables. *Ann. Statist.*, **9**, 879–885.

Shannon, C. E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379-423, 623-656.

Shorack, G. R. and Wellner, J. A. (2009) *Empirical Processes with Applications to Statistics*. SIAM.

Singh, H., Misra, N., Hnizdo, V., Fedorowicz, A. and Demchuk, E. (2003) Nearest neighbor estimates of entropy. *Am. J. Math. Manag. Sci.*, **23**, 301–321.

Singh, S. and Póczos, B. (2016) Analysis of k nearest neighbor distances with application to entropy estimation. *NIPS*, **29**, 1217–1225.

Song, L., Smola, A., Gretton, A., Bedo, J. and Borgwardt, K. (2012) Feature selection via dependence maximization. *J. Mach. Learn. Res.*, **13**, 1393–1434.

Sricharan, K., Raich, R. and Hero, A. O. (2012) Estimation of nonlinear functionals of densities with confidence. *IEEE Trans. Inf. Theory*, **58**, 4135–4159.

Sricharan, K., Wei, D. and Hero, A. O. (2013) Ensemble estimators for multivariate entropy estimation. *IEEE Trans. Inf. Theory*, **59**, 4374–4388.

Steuer, R., Kurths, J., Daub, C. O., Weise, J. and Selbig, J. (2002) The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, **18**, 231–240.

Stigler, S. M. (1989) Francis Galton's account of the invention of correlation. *Stat. Sci.*, **4**, 73–86.

Stone, C. J. (1977) Consistent nonparametric regression. *Ann. Statist.*, **5**, 595–620.

Su, L. and White, H. (2008) A nonparametric hellinger metric test for conditional independence. *Econometric Theory*, **24**, 829–864.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007) Measuring and testing dependence by correlation of distances. *Ann. Statist.*, **35**, 2769–2794.

Székely, G. J. and Rizzo, M. L. (2013) The distance correlation t-test of independence in high dimension. *J. Multivar. Anal.*, **117**, 193–213.

Torkkola, K. (2003) Feature extraction by non-parametric mutual information maximization. *J. Mach. Learn. Res.*, **3**, 1415–1438.

Tsybakov, A. B. and Van der Meulen, E. C. (1996) Root-n consistent estimators of entropy for densities with unbounded support. *Scand. J. Stat.*, **23**, 75–83.

Vaidya, P. M. (1989) An $O(n \log n)$ algorithm for the all-nearest-neighbors problem. *Discrete Comput. Geom.*, **4**, 101–115.

van der Vaart, A. W. (1998) *Asymptotic Statistics.* Cambridge University Press, Cambridge.

Vasicek, O. (1976) A test for normality based on sample entropy. *J. Roy. Statist. Soc., Ser. B.*, **38**, 54–59.

Vinh, N. X., Epps, J. and Bailey, J. (2010) Information theoretic measures for clusterings comparison: variants, properties, normalisation and correction for chance. *J. Mach. Learn. Res.*, **11**, 2837–2854.

Wang, Q., Kulkarni, S. R. and Verdú, S. (2008) Universal estimation of information measures for analog sources. *Found. Trends Commun. Inf. Theory*, **5**, 265–353.

Weinberger, K. Q. and Saul, L. K. (2009) Distance metric learning for large margin nearest neighbour classification. *J. Mach. Learn. Res.*, **10**, 207–244.

Wettschereck, D. and Dietterich, T. G. (1994) Locally adaptive nearest neighbor algorithms. *NIPS*, **7**, 184–191.

Wiener, N. (1948) *Cybernetics: Control and Communication in the Animal and the Machine.* MIT Press, Cambridge MA.

Zhang, K., Peters, J., Janzing, D. and Schölkopf, B. (2011) Kernel-based conditional independence test and application in causal discovery. `https://arxiv.org/abs/1202.3775`.

Zhang, Q., Filippi, S., Gretton, A. and Sejdinovic, D. (2017) Large-scale kernel methods for independence testing. *Stat. Comput.*, **27**, 1–18.

Zhao, M. and Saligrama, V. (2009) Anomaly detection with score functions based on nearest neighbor graphs. *NIPS*, **22**, 2250–2258.