A law of large numbers for nearest neighbour statistics

BY DAFYDD EVANS*

School of Computer Science, Cardiff University, 5 The Parade, Cardiff CF24 3AA, UK

In practical data analysis, methods based on proximity (near-neighbour) relationships between sample points are important because these relations can be computed in time $O(n \log n)$ as the number of points $n \to \infty$. Associated with such methods are a class of random variables defined to be functions of a given point and its nearest neighbours in the sample. If the sample points are independent and identically distributed, the associated random variables will also be identically distributed but not independent. Despite this, we show that random variables of this type satisfy a strong law of large numbers, in the sense that their sample means converge to their expected values almost surely as the number of sample points $n \to \infty$.

Keywords: nearest neighbours; geometric probability; difference-based methods; noise estimation

1. Introduction

Let $\mathcal{X}=X_1, X_2, \ldots$ be a sequence of independent and identically distributed random vectors $X_i \in \mathbb{R}^d$, and let $\mathcal{X}_n = (X_1, \ldots, X_n)$ denote the first $n$ points of the sequence. The common distribution of the $X_i$ will be called the sampling distribution and $\mathcal{X}_n$ will be called a sample (of size $n$). Let $v_i(n, k)$ denote the index of the $k$-nearest neighbour of $X_i$ among the points of $\mathcal{X}_n = (X_1, \ldots, X_n)$, where the distance between any two points is defined with respect to the metric induced by any $l^p$-norm on $\mathbb{R}^d$,

$$
\| (a_1, \ldots, a_d) \|_p = \begin{cases} 
\left( \sum_{j=1}^{d} |a_j|^p \right)^{1/p}, & 1 \leq p < \infty, \\
\max_{j=1, \ldots, d} |a_j|, & p = \infty,
\end{cases}
$$

and equidistant points are ordered by their indices. Let $h : \mathbb{R}^{d \times (k+1)} \to \mathbb{R}$ be a measurable function. We define the identically distributed random variables

$$
h_{i,n}(\mathcal{X}) = h(X_i, X_{v_i(n,1)}, \ldots, X_{v_i(n,k)}) \quad (i = 1, \ldots, n),
$$

so that $h_{i,n}(\mathcal{X})$ is a function of the sample point $X_i$ and its $k$-nearest neighbours among the points of $\mathcal{X}_n$. Let $\mu_n$, $\sigma_n^2$ and $\kappa_n$, respectively, denote the mean, variance and kurtosis of the (identically distributed) random variables $h_{i,n}(\mathcal{X})$, *d.evans@cs.cardiff.ac.uk

Received 5 June 2008
Accepted 11 July 2008
and let \( S_n \) and \( H_n \), respectively, denote their sum and sample mean,

\[
S_n(\mathcal{X}) = \sum_{i=1}^{n} h_{i,n}(\mathcal{X}) \quad \text{and} \quad H_n(\mathcal{X}) = \frac{1}{n} \sum_{i=1}^{n} h_{i,n}(\mathcal{X}). \tag{1.2}
\]

Our main result is that \( H_n \) is a strongly consistent estimator for \( \mu_n \) as \( n \to \infty \).

**Theorem 1.1.** If the sampling distribution is continuous and if the kurtosis \( \kappa_n \) is uniformly bounded for all \( n \in \mathbb{N} \), then

\[
\left| \frac{H_n - \mu_n}{\sigma_n} \right| \to 0 \quad \text{a.s. as} \quad n \to \infty. \tag{1.3}
\]

Theorem 1.1 shows that the rate at which \( H_n \) converges to \( \mu_n \) is at least of the order of \( O(\sigma_n) \) as \( n \to \infty \). Let us define the normalized random variables,

\[
h_{i,n}^*(\mathcal{X}) = \frac{h_{i,n}(\mathcal{X}) - \mu_n}{\sigma_n}, \tag{1.4}
\]

along with their sum \( S_n^* \) and sample mean \( H_n^* \),

\[
S_n^*(\mathcal{X}) = \sum_{i=1}^{n} h_{i,n}^*(\mathcal{X}) \quad \text{and} \quad H_n^*(\mathcal{X}) = \frac{1}{n} \sum_{i=1}^{n} h_{i,n}^*(\mathcal{X}). \tag{1.5}
\]

If the sampling distribution is continuous, we show that the variance of \( S_n^* \) is of the asymptotic order \( O(n) \) as \( n \to \infty \).

**Theorem 1.2.** If the sampling distribution is continuous, then for all \( n \geq 16k \),

\[
\text{var}(S_n^*) \leq 2(n + 1)(3 + 8k\beta)(3 + 64k)(\kappa_n + \kappa_{n+1}), \tag{1.6}
\]

where \( \beta(k, d, p) = k[dv_{d,p}] \) and \( dv_{d,p} \) is the volume of the unit ball in \((\mathbb{R}^d, \| \cdot \|_p)\).

To show that \( |H_n^*| \to 0 \) a.s. as \( n \to \infty \), we first show that the expected squared difference \( \mathcal{E}(S_n^* - S_{n+1}^*)^2 \) between successive terms of the sequence \( S_n^* \) is bounded independently of \( n \). We then use the Efron–Stein inequality (Efron & Stein 1981; Steele 1986) to bound the variance \( \text{var}(S_n^*) \), which by Chebyshev’s inequality yields the weak convergence of \( H_n^* \) to zero as \( n \to \infty \). Finally, we again use the bound on \( \mathcal{E}(S_n^* - S_{n+1}^*)^2 \) to prove strong convergence using a standard argument.

Similar results have previously appeared in the literature. In particular, Penrose & Yukich (2003) obtained a weak law of large numbers for functions of binomial point processes in \( \mathbb{R}^d \), then applied the result to a number of different types of random proximity graphs, and a number of functions including the total edge length (under arbitrary weighting) and the number of components in the graph. Wade (2007) applied the general result of Penrose & Yukich (2003) to the total (power-weighted) edge length of several types of the nearest neighbour graphs on random point sets in \( \mathbb{R}^d \), and gave explicit expressions for the limiting constants. In two recent papers, Penrose (2007) obtained a law of large numbers for a class of marked point processes and gave some applications to noise estimation, while Penrose & Wade (2008) have investigated the asymptotic properties of the random online nearest neighbour graph.

The methods of Penrose & Yukich (2003) depend on the notion of stabilization, which demands that the neighbourhood of a particular vertex is unaffected by changes in the neighbourhood of another vertex located outside some sufficiently large ball (the radius of this ball is called the radius of stabilization). Because we consider only \( k \)-nearest neighbours graphs, we will instead rely on the standard...
geometric fact that there exists a finite number $\beta = \beta(k, d, p)$ such that for any countable set of points in $(\mathbb{R}^d, \|\cdot\|_p)$, any point of the set can be among the first $k$-nearest neighbours of at most $\beta$ other points of the set.

Motivated by the need for a multidimensional goodness-of-fit test, Bickel & Breiman (1983) investigated the asymptotic properties of sums of bounded functions of the (first) nearest neighbour distances, and proved a central limit theorem for the random variables $f(X_i)\|X_i - X_{r_i(n,1)}\|$ where $f$ is the (unknown) sampling density. This was extended to $k$-nearest neighbour distances by Penrose (2000), with $k$ being allowed to increase as a fractional power of the number of points $n$.

Functions defined in terms of proximity relations exhibit 'local' dependence, which can often be represented by dependency graphs, first described by Petrovskaya & Leontovitch (1982). For a set of random variables which can often be represented by dependency graphs, first described by Breiman (1983) investigated the asymptotic properties of sums of bounded functions of the (first) nearest neighbour distances, and proved a central limit theorem for the

Following result is due to Baldi & Rinott (1989).

**Theorem 1.3 (Baldi & Rinott 1989).** Let $\{Z_i : i \in V\}$ be random variables having a dependency graph $G(V, E)$, and define $S = \sum_{i \in V} Z_i$. Let $D$ denote the maximal degree of $G$ and suppose $|Z_i| \leq B$ almost surely. Then

$$\left| P\left( \frac{S - \mathcal{E}(S)}{\sqrt{\text{var}(S)}} \leq x \right) - \Phi(x) \right| \leq 32 \left( 1 + \sqrt{6} \right) \left( \frac{|V| D^2 B^3}{\text{var}(S)^{3/2}} \right)^{1/2},$$

where $|V|$ is the cardinality of $V$ and $\Phi(x)$ is the standard normal distribution $N(0, 1)$.

Avram & Bertsimas (1993) applied theorem 1.3 to the length of the $k$-nearest neighbour graph, the Delaunay triangulation and the Voronoi diagram of random point sets. More recently, Penrose & Yukich (2001) used stabilization properties to prove central limit theorems for functions of various types of (random) proximity graphs, while Chen & Shao (2004) obtained central limit theorems for various types of local dependence, and in particular improved on the result of Baldi & Rinott (1989).

In the present context, vertices in the dependency graph corresponding to the random variables $h_{i,n}$ and $h_{j,n}$ are connected by an edge only if they share a common nearest neighbour. By lemma 4.2, the maximal degree of $G$ is therefore bounded independently of the number of points $n$. Under the additional assumption that the $h_{i,n}$ are bounded a.s., if we could show that there exist constants $c, \epsilon > 0$ such that $\text{var}(S_n) \geq cn^{2/3+\epsilon}$, it would then follow that $S_n^* \rightarrow N(0, 1)$ in distribution as $n \rightarrow \infty$. This is beyond the scope of the present paper.

2. Applications

(a) Nearest neighbour distances

Let $\mathcal{X} = X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables taking values in the unit cube $[0,1]^d$. Let $\delta_{i,n}(\mathcal{X})$ be the distance between the point $X_i$ and its $k$-nearest neighbour in the sample $\mathcal{X}_n$, and let $\Delta_n(\mathcal{X})$
Thus, distance \( m(i) \) denotes the sample mean of the \( \delta_{i,n}(X) \),

\[
\delta_{i,n}(X) = \|X_i - X_{p,(n,k)}\| \quad \text{and} \quad A_n(X) = \frac{1}{n} \sum_{i=1}^{n} \delta_{i,n}(X).
\] (2.1)

Let \( \mu_n^{(\delta)}, \sigma_n^{(\delta)}, \text{and} \kappa_n^{(\delta)} \), respectively, denote the mean, standard deviation and kurtosis of the (identically distributed) random variables \( \delta_{i,n}(X) \).

Let \( F \) denote the sampling distribution, \( B_x(r) \) denote the ball of radius \( r \) centred at \( x \in \mathbb{R}^d \) and \( \omega_x(r) \) denote its probability measure:

\[
\omega_x(r) = \int_{B_x(r)} dF.
\] (2.2)

Suppose that \( F \) satisfies a positive density condition (Gruber 2004), in the sense that there exist constants \( a>1 \) and \( \rho>0 \) such that \( a^{-1}r^d \leq \omega_x(r) \leq ar^d \) for all \( 0 \leq r \leq \rho \). Suppose also that \( F \) satisfies a smooth density condition in the sense that its second partial derivatives are bounded at every point of \([0,1]^d\). Under these conditions, it is known (Evans et al. 2002) that for all \( \epsilon>0 \), the moments of the \( k \)-nearest neighbour distance distribution satisfies

\[
\mathcal{E}(\delta_{i,n}^\alpha) = \frac{\Gamma(k + \alpha/d)}{v_{d,p}^{\alpha/d} n^{\alpha/d} \Gamma(k)} \left( \int f(x)^{(d-\alpha)/d} \, dx \right) \left( 1 + O\left( \frac{1}{n^{1/d-\epsilon}} \right) \right),
\] (2.3)

as \( n \to \infty \), where \( v_{d,p} \) is the volume of the unit ball in \((\mathbb{R}^d, \| \cdot \|_p)\) and \( f(x) = F'(x) \) is the density of the sampling distribution. In particular, \( \sigma_n^{(\delta)} = O(n^{-1/d}) \) and \( \kappa_n^{(\delta)} = O(1) \) as \( n \to \infty \), so by theorem 1.1 it follows that:

\[
n^{1/d} |A_n - \mu_n^{(\delta)}| \to 0 \quad \text{a.s. as} \quad n \to \infty.
\] (2.4)

Thus, \( A_n \) is a strongly consistent estimator for the expected \( k \)-nearest neighbour distance \( \mu_n^{(\delta)} \), and its rate of convergence is of the asymptotic order \( O(n^{1/d}) \) as the number of sample points \( n \to \infty \). In fact, taking \( \alpha=1 \) in (2.3) we have

\[
\lim_{n \to \infty} n^{1/d} \mu_n^{(\delta)} = \frac{\Gamma(k + 1/d)}{v_{d,p}^{1/d} \Gamma(k)} \int f(x)^{(d-1)/d} \, dx,
\] (2.5)

so by theorem 1.1,

\[
\lim_{n \to \infty} n^{1/d} A_n(X) = \frac{\Gamma(k + 1/d)}{v_{d,p}^{1/d} \Gamma(k)} \int f(x)^{(d-1)/d} \, dx \quad \text{a.s.}
\] (2.6)

(i) The Euclidean \( k \)-nearest neighbours graph

For any set of vertices \( \{x_1, \ldots, x_n\} \) in \( \mathbb{R}^d \), the (undirected) Euclidean \( k \)-nearest neighbours graph is constructed by including an edge between every point and its \( k \)-nearest neighbours in the set, where the nearest neighbour relations are defined with respect to the Euclidean metric \( (p=2) \). Let \( L_n(X) \) denote the total length of the Euclidean \( k \)-nearest neighbours graph of the random sample \( X_n \),

\[
L_n(X) = \sum_{i=1}^{n} \sum_{\ell=1}^{k} \|X_i - X_{p,(n,\ell)}\|_2.
\] (2.7)
Nearest neighbour statistics

Because the volume of the unit ball in \((\mathbb{R}^d, \|\cdot\|_2)\) is equal to \(2\pi^{d/2}/\Gamma(1 + d/2)\), then subject to the above conditions on the sampling distribution it follows by (2.6) that:

\[
\lim_{n \to \infty} L_n(\mathcal{X})/n^{(d-1)/d} = \alpha(k, d) \int f(x)^{(d-1)/d} \, dx \quad \text{a.s.,}
\]

where

\[
\alpha(k, d) = \left( \frac{d}{d+1} \right) \frac{\Gamma(k + 1 + 1/d)}{\Gamma(k)},
\]

which agrees with theorem 2(b) of Wade (2007).

(b) Noise estimation

Non-parametric regression attempts to model the behaviour of some observable variable \(Y \in \mathbb{R}\) in terms of another observable variable \(X \in \mathbb{R}^d\), based only on a finite number of observations of the joint variable \((X, Y)\). The relationship between \(X\) and \(Y\) is often assumed to satisfy the additive hypothesis

\[
Y = \mathcal{E}(Y|X) + R,
\]

where \(\mathcal{E}(Y|X)\) is the regression function and \(R\) is the residual variable (noise). Let \(Z = Z_1, Z_2, \ldots\) be a sequence of independent and identically distributed observations \(Z_i = (X_i, Y_i)\) of the joint variable \(Z = (X, Y)\), let \(\mathcal{X} = X_1, X_2, \ldots\) and \(\mathcal{Y} = Y_1, Y_2, \ldots\) denote the marginal sequences, and let \(\nu_i(n, k)\) denote the index of the \(k\)-nearest neighbour of \(X_i\) among the points of the marginal sample \(\mathcal{X}_n = (X_1, \ldots, X_n)\). To estimate the \(k\)-moment \(\mathcal{E}(R^k)\) of the residual distribution \((k \in \mathbb{N})\), we consider the random variables \(\gamma_{i,n}(Z)\) and their sample mean \(\Gamma_n(Z)\), defined by

\[
\gamma_{i,n}(Z) = \prod_{\ell=1}^k (Y_i - Y_{\nu_i(n,\ell)}) \quad \text{and} \quad \Gamma_n(Z) = \frac{1}{n} \sum_{i=1}^n \gamma_{i,n}(Z).
\]

Let \(\mu_n^{(\gamma)}, \sigma_n^{(\gamma)}\) and \(\kappa_n^{(\gamma)}\), respectively, denote the mean, standard deviation and kurtosis of the (identically distributed) random variables \(\gamma_{i,n}(Z)\).

Perhaps the main contribution of this paper is that theorem 1.1 extends to possibly unbounded random variables, the only requirement being that their first four moments are bounded. This allows the residual variable \(R\) in (2.10) to take arbitrarily large values.

Suppose that the sampling distribution of the explanatory variables \(X_i\) satisfies smooth and positive density conditions, the regression function \(\mathcal{E}(Y|X)\) satisfies a Lipschitz condition on \(\mathbb{R}^d\) and the residual variable \(R\) is independent of \(X\) (homoscedastic). Under these conditions, Evans & Jones (2008) have shown that for \(k \in \mathbb{N}\),

\[
\mu_n^{(\gamma)} = \mathcal{E}(R^k) + \mathcal{O}(\mu_n^{(\delta)}) \quad \text{as } n \to \infty,
\]

where \(\mu_n^{(\delta)}\) is the expected \(k\)-nearest neighbour distance in the marginal sample \((X_1, \ldots, X_n)\) and the implied constant depends on the residual moments up to order \(k-1\), the constant implied by the Lipschitz condition on the regression function, and the constant implied by the positive density condition on the sampling distribution.

Proc. R. Soc. A (2008)
Thus, the sample mean inequality to the sum

$$|\Gamma_n - \mu_n^{(\gamma)}| \to 0 \quad \text{a.s. as } n \to \infty.$$  \hspace{1cm} (2.13)

Thus, the sample mean $\Gamma_n$ is a strongly consistent estimator for the $k$th moment $\mathcal{E}(R^k)$ of the residual distribution as $n \to \infty$.

3. The Efron–Stein inequality

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random vectors in $\mathbb{R}^d$, and let $g : \mathbb{R}^{d \times n} \to \mathbb{R}$ be a symmetric function of $n$ vectors in $\mathbb{R}^d$. The Efron–Stein inequality (Efron & Stein 1981) provides an upper bound for the variance of the statistic $Z = g(X_1, \ldots, X_n)$ in terms of the quantities,

$$Z(i) = g(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{n+1}) \quad \text{and} \quad Z(\cdot) = \frac{1}{n+1} \sum_{i=1}^{n+1} Z(i). \hspace{1cm} (3.1)$$

**Theorem 3.1 (Efron & Stein 1981).** Let $X_1, X_2, \ldots, X_{n+1}$ be independent and identically distributed random vectors in $\mathbb{R}^d$, let $g : \mathbb{R}^{d \times n} \to \mathbb{R}$ be a symmetric function, and define the random variable $Z = g(X_1, \ldots, X_n)$.

If $\mathcal{E}(Z^2) < \infty$, then

$$\text{var}(Z) \leq \sum_{i=1}^{n+1} \mathcal{E}(Z(i) - Z(\cdot))^2. \hspace{1cm} (3.2)$$

The Efron–Stein inequality has found a wide range of applications in statistics. Our approach is partly based on the work of Reitzner (2003), who uses the inequality to prove strong laws of large numbers for statistics related to random polytopes.

First we note that, because $X_1, \ldots, X_{n+1}$ are independent and identically distributed, it follows by symmetry on the indices that:

$$\text{var}(Z) \leq (n+1)\mathcal{E}(Z(n+1) - Z(\cdot))^2. \hspace{1cm} (3.3)$$

Furthermore, for any $a \in \mathbb{R}$, we have that

$$\sum_{i=1}^{n+1} \mathcal{E}(Z(i) - Z(\cdot))^2 = \sum_{i=1}^{n+1} \mathcal{E}((Z(i) - a) - (Z(\cdot) - a))^2 \hspace{1cm} (3.4)$$

$$= \sum_{i=1}^{n+1} \mathcal{E}(Z(i) - a)^2 - (n+1)(Z(\cdot) - a)^2, \hspace{1cm} (3.5)$$

so inequalities (3.2) and (3.3) are preserved if we replace $Z(\cdot)$ by any other function of $X_1, \ldots, X_{n+1}$. To prove theorem 1.1, we apply the Efron–Stein inequality to the sum $S_n(X)$,

$$Z(X_1, \ldots, X_n) = \sum_{i=1}^{n} h_{i,n}(\mathcal{X}) \hspace{1cm} (3.6)$$

$$= \sum_{i=1}^{n} h(X_i, X_{r(n,1)}, \ldots, X_{r(n,k)}). \hspace{1cm} (3.7)$$
If \( \mathcal{E}(h_{i,n}^2) < \infty \) then \( \mathcal{E}(Z^2) < \infty \). Furthermore, because the points \( X_1, \ldots, X_n \) are independent and identically distributed, it follows that \( S_n = Z(X_1, \ldots, X_n) \) is invariant under permutations of its arguments. Thus, we may apply the Efron–Stein inequality to \( S_n \). In this case \( Z_{(n+1)} = S_n \), so by replacing \( Z_{(\cdot)} \) by \( S_{n+1} \) in (3.3), we obtain

\[
\text{var}(S_n) \leq (n + 1)\mathcal{E}(S_n - S_{n+1})^2.
\] (3.8)

An identical expression also holds for the normalized sum \( S_n^* \). The lemma 3.2 shows that the expected squared difference between successive values of \( S_n^* \) is bounded independently of the total number of points \( n \). The proof is given in \( \S4 \).

**Lemma 3.2.** If the sampling distribution is continuous, then for all \( n \geq 16k \),

\[
\mathcal{E}(S_n^* - S_{n+1}^*)^2 \leq 2(3 + 8k\beta)(3 + 64k)(\kappa_n + \kappa_{n+1}).
\] (3.9)

The following corollary of lemma 3.2 follows immediately by (3.8) and Chebyshev’s inequality, and asserts the weak consistency of the sample mean \( H_n \), defined in (1.2), as an estimator for true mean \( \mu_n \) as \( n \to \infty \) (provided the kurtosis \( \kappa_n \) is uniformly bounded for all \( n \in \mathbb{N} \)).

**Corollary 3.3.** If the sampling distribution is continuous, then for all \( n \geq 16k \),

\[
P\left( \left| \frac{H_n - \mu_n}{\sigma_n} \right| > \epsilon \right) \leq \frac{4(3 + 8k\beta)(3 + 64k)(\kappa_n + \kappa_{n+1})}{n\epsilon^2}.
\] (3.10)

We prove lemma 3.2 using methods based on the approach of Bickel & Breiman (1983) and Hitczenko et al. (1999). Having established lemma 3.2, the proof of theorem 1.1 is then straightforward.

### 4. Proofs

**a) Standard results**

We need two standard geometric results. The first of these concerns the expected probability measure of \( k \)-nearest neighbour balls. For \( X_i \in \mathcal{X}_n \), let \( B_{i,n}(\mathcal{X}) \subset \mathbb{R}^d \) denote the \( k \)-nearest neighbour ball of \( X_i \) (with respect to the finite sample \( \mathcal{X}_n \)), defined to be the ball centred at \( X_i \) of radius equal to the distance from \( X_i \) to its \( k \)-nearest neighbour in \( \mathcal{X}_n \), and let \( \omega_{i,n}(\mathcal{X}) \) denote its probability measure,

\[
\omega_{i,n}(\mathcal{X}) = \int_{B_{i,n}(\mathcal{X})} dF,
\] (4.1)

where \( F \) is the (common) distribution function of the sample points \( X_1, X_2, \ldots \). It is well known, see for example Percus & Martin (1998) or Evans et al. (2002), that provided the sampling distribution is continuous, the expected probability measure of any \( k \)-nearest neighbour ball (over all sample realizations) is equal to \( k/n \).

**Lemma 4.1.** If the sampling distribution is continuous, then

\[
\mathcal{E}(\omega_{i,n}) = k/n.
\] (4.2)
The second result concerns the maximum degree of any vertex in a \( k \)-nearest neighbour graph. It is well known that this number is bounded independently of the total number of vertices (e.g. Stone 1977; Bickel & Breiman 1983; Zeger & Gersho 1994; Yukich 1998).

**Lemma 4.2.** For every countable set of points in \( \mathbb{R}^d \), any point can be among the first \( k \)-nearest neighbours of at most \( \beta(k, d, p) = k |pv_d,p| \) other points of the set, where \( v_{d,p} \) is the volume of the unit ball in \( (\mathbb{R}^d, \| \cdot \|_p) \).

In particular, \( \beta(k, d, 2) = k \lceil 2\pi^{d/2} / \Gamma(d/2) \rceil \) and \( \beta(k, d, \infty) = kd2^d \).

(b) Adding a sample point

Let \( \lambda_n \) denote the second moment of the random variables \( h_{i,n}(\mathcal{X}) \),

\[
\lambda_n = \sigma_n^2 + \mu_n^2. \tag{4.3}
\]

**Lemma 4.3.** If the sampling distribution is continuous,

\[
\mathcal{E} \left( \sum_{i=1}^{n} (h_{i,n} - h_{i,n+1})^2 \right) \leq 2k(\lambda_n + \lambda_{n+1}). \tag{4.4}
\]

**Proof.** Let \( B_{i,n}(\mathcal{X}) \subset \mathbb{R}^d \) denote the \( k \)-nearest neighbour ball of \( X_i \) with respect to the finite sample \( \mathcal{X}_n = (X_1, \ldots, X_n) \), and suppose we add another (independent and identically distributed) point \( X_{n+1} \) to the ensemble. If \( X_{n+1} \notin B_{i,n}(\mathcal{X}) \) (i.e. if the new point \( X_{n+1} \) falls outside the current \( k \)-nearest neighbour ball of \( X_i \)), then \( X_{r_i(n+1)} = X_{r_i(n+1,\ell)} \) for all \( 1 \leq \ell \leq k \), and hence

\[
h_{i,n} = h(X_i, X_{r_i(n+1)}, \ldots, X_{r_i(n,k)}) \tag{4.5}
\]

\[
= h(X_i, X_{r_i(n+1,1)}, \ldots, X_{r_i(n+1,k)}) \tag{4.6}
\]

\[
= h_{i,n+1}. \tag{4.7}
\]

Thus, we have \( h_{i,n} \neq h_{i,n+1} \) only if \( X_{n+1} \in B_{i,n}(\mathcal{X}) \). Let \( J_{i,n+1} \) denote the indicator variable of the event \( X_{n+1} \in B_{i,n}(\mathcal{X}) \),

\[
J_{i,n+1}(\mathcal{X}) = \begin{cases} 
1, & \text{if } X_{n+1} \in B_{i,n}(\mathcal{X}), \\
0, & \text{otherwise.}
\end{cases} \tag{4.8}
\]

Then

\[
\mathcal{E} \left( \sum_{i=1}^{n} (h_{i,n} - h_{i,n+1})^2 \right) = \mathcal{E} \left( \sum_{i=1}^{n} (h_{i,n} - h_{i,n+1})^2 J_{i,n+1} \right). \tag{4.9}
\]

By lemma 4.1, the probability measure of \( B_{i,n}(\mathcal{X}) \) is equal to \( k/n \). Hence, because the new point \( X_{n+1} \) is assumed to be independent of the previously selected points \( X_1, \ldots, X_n \), we have

\[
P(J_{i,n+1} = 1) = k/n \tag{4.10}
\]
So that we can consider samples of sizes $n$ and $n+1$ separately, we apply the crude bound
\[
E((h_{i,n} - h_{i,n+1})^2 | J_{i,n+1} = 1) \leq 2E(h_{i,n}^2 | J_{i,n+1} = 1) + 2E(h_{i,n+1}^2 | J_{i,n+1} = 1).
\]
(4.12)

For samples of size $n$, because the random variable $h_{i,n}^2$ is determined only by the points $X_1, \ldots, X_n$, it follows that $h_{i,n}^2$ is independent of the event $X_{n+1} \in B_{i,n}$, and therefore:
\[
E(h_{i,n}^2 | J_{i,n+1} = 1) = E(h_{i,n}^2) = \lambda_n.
\]
(4.13)

For samples of size $n+1$, the event $X_{n+1} \in B_{i,n}$ is simply that the ‘last’ point of the sample $X_1, \ldots, X_{n+1}$ is among the first $k$-nearest neighbours of $X_i$, i.e. $X_{n+1} \in B_{i,n}$ if and only if $\nu_i(n+1, \ell) = n+1$ for some $1 \leq \ell \leq k$. Because the points $X_1, \ldots, X_{n+1}$ are independent and identically distributed, the order in which they are indexed is arbitrary. Hence, it follows by symmetry on the indices $j \neq i$ that the condition $\nu_i(n+1, k) = n+1$ cannot affect the expected value of $h_{i,n+1}^2$, so
\[
E(h_{i,n+1}^2 | J_{i,n+1}) = E(h_{i,n+1}^2) = \lambda_{n+1}.
\]
(4.14)

Thus, we have that
\[
E((h_{i,n} - h_{i,n+1})^2 | J_{i,n+1} = 1) = 2(\lambda_n + \lambda_{n+1})
\]
(4.15)

and the result follows by (4.11).

**Lemma 4.4.** If the sampling distribution is continuous, then for all $n \geq 16k$,
\[
\lambda_{n+1} \leq 3\lambda_n.
\]
(4.16)

**Proof.** Because the $h_{i,n}$ are identically distributed,
\[
\lambda_{n+1} = E\left(\frac{1}{n} \sum_{i=1}^{n} h_{i,n}^2 \right)
\]
(4.17)

\[
= \frac{1}{n} E\left(\sum_{i=1}^{n} (h_{i,n} - (h_{i,n} - h_{i,n+1}))^2\right)
\]
(4.18)

\[
\leq \frac{2}{n} E\left(\sum_{i=1}^{n} (h_{i,n}^2 + (h_{i,n} - h_{i,n+1})^2)\right)
\]
(4.19)

\[
\leq 2\lambda_n + \frac{2}{n} E\left(\sum_{i=1}^{n} (h_{i,n} - h_{i,n+1})^2\right).
\]
(4.20)
Hence by lemma 4.3,\[ \lambda_{n+1} \leq 2\lambda_n + \frac{4k}{n}(\lambda_n + \lambda_{n+1}). \] (4.21)
Thus, for \( n \geq 16k \) we have \( 4k/n \leq 1/4 \) and hence \( \lambda_{n+1} \leq 3\lambda_n. \) ■

**Lemma 4.5.** If the sampling distribution is continuous, then for all \( n \geq 16k, \)
\[ \mathcal{E}\left( \sum_{i=1}^{n}(h_{i,n} - h_{i,n+1})^2 \right) \leq 8k\beta\lambda_n. \] (4.22)

**Proof.** Let \( \mathcal{M}_{n+1} \) be the set containing the index of every point in \( X_{n+1} \), which has the new point \( X_{n+1} \) among its first \( k \)-nearest neighbours,
\[ \mathcal{M}_{n+1}(X) = \{ i : v_i(n+1, \ell) = n+1, \text{ for some } 1 \leq \ell \leq k \}. \] (4.23)
By construction, if \( i \notin \mathcal{M}_{n+1} \) then \( X_{r_i(n,\ell)} = X_{r_i(n+1,\ell)} \) for all \( 1 \leq \ell \leq k \), so \( h_{i,n} \neq h_{i,n+1} \) only if \( i \in \mathcal{M}_{n+1} \). Furthermore, by lemma 4.2, the new point \( X_{n+1} \) can become one of the first \( k \)-nearest neighbours of at most \( \beta \) of the existing points \( X_1, \ldots, X_n \), so \( |\mathcal{M}_{n+1}| \leq \beta \). Hence, by the Cauchy inequality,
\[ \left( \sum_{i=1}^{n}(h_{i,n} - h_{i,n+1}) \right)^2 = \left( \sum_{i \in \mathcal{M}_{n+1}}(h_{i,n} - h_{i,n+1}) \right)^2 \leq |\mathcal{M}_{n+1}| \sum_{i \in \mathcal{M}_{n+1}}(h_{i,n} - h_{i,n+1})^2 \] (4.24)
\[ \leq |\mathcal{M}_{n+1}| \sum_{i \in \mathcal{M}_{n+1}}(h_{i,n} - h_{i,n+1})^2 \leq \beta \sum_{i=1}^{n}(h_{i,n} - h_{i,n+1})^2, \] (4.25)
and the result follows by lemmas 4.3 and 4.4. ■

(c) An upper bound on \( \text{var}(S_n) \)

**Lemma 4.6.** If the sampling distribution is continuous, then for all \( n \geq 16k, \)
\[ \text{(i)} \quad \mathcal{E}(S_n - S_{n+1})^2 \leq 2(3 + 8k\beta)\lambda_n \]
and
\[ \text{(ii)} \quad \text{var}(S_n) \leq 2(n + 1)(3 + 8k\beta)\lambda_n. \]

**Proof.** By definition,\[ S_n - S_{n+1} = \sum_{i=1}^{n}(h_{i,n} - h_{i,n+1}) - h_{n+1,n+1}, \] (4.27)
so\[ (S_n - S_{n+1})^2 \leq 2 \left( \sum_{i=1}^{n}(h_{i,n} - h_{i,n+1}) \right)^2 + 2h_{n+1,n+1}^2. \] (4.28)
Taking expectations then applying lemmas 4.4 and 4.5,\[ \mathcal{E}(S_n - S_{n+1})^2 \leq 2(3 + 8k\beta)\lambda_n \] (4.29)

\textit{Proc. R. Soc. A} (2008)
and by (3.8),
\[ \text{var}(S_n) \leq 2(n+1)(3 + 8k\beta)\lambda_n. \quad (4.30) \]

(d) Normalization

To find an upper bound on \( \mathcal{E}(S_{n+1}^* - S_n^*)^2 \), we first need to quantify the squared difference between successive values of \( \mu_n \) and \( \sigma_n \).

**Lemma 4.7.** If the sampling distribution is continuous, then for all \( n \geq 16k \),

(i) \( (\mu_n - \mu_{n+1})^2 \leq \frac{8k\beta}{n^2} \lambda_n \)

and

(ii) \( (\sigma_n - \sigma_{n+1})^2 \leq \frac{16k}{n} \lambda_n \).

**Proof.** To prove (i), because the \( h_{i,n} \) are identically distributed,

\[
(\mu_n - \mu_{n+1})^2 = \left[ \mathcal{E}\left( \frac{1}{n} \sum_{i=1}^{n} (h_{i,n} - h_{i,n+1}) \right) \right]^2 \leq \frac{1}{n^2} \mathcal{E}\left( \sum_{i=1}^{n} (h_{i,n} - h_{i,n+1}) \right)^2 \leq \frac{8k\beta}{n^2} \lambda_n, 
\]

where the last inequality follows by lemma 4.5.

To prove (ii), let \( \tilde{X} = (\tilde{X}_1, \tilde{X}_2, \ldots) \) be an independent copy of \( X = (X_1, X_2, \ldots) \), and let \( \tilde{h}_{i,n}(k) \) denote the index of the \( k \)-nearest neighbour of \( \tilde{X}_i \) among the points of \( \tilde{X}_n = (\tilde{X}_1, \ldots, \tilde{X}_n) \). We consider the random variables,

\[
\tilde{h}_{i,n}(\tilde{X}) = h(\tilde{X}_i, \tilde{X}_{\tilde{h}_{i,n}(1)}, \ldots, \tilde{X}_{\tilde{h}_{i,n}(k)}).
\]

Because \( h_{i,n} \) and \( \tilde{h}_{i,n} \) are independent and identically distributed,

\[
\mathcal{E}((h_{i,n} - \tilde{h}_{i,n})^2) = \mathcal{E}(h_{i,n}^2) - 2\mathcal{E}(h_{i,n}\tilde{h}_{i,n}) + \mathcal{E}(\tilde{h}_{i,n}^2)
\]

\[
= \mathcal{E}(h_{i,n}^2) - 2\mathcal{E}(h_{i,n})\mathcal{E}(\tilde{h}_{i,n}) + \mathcal{E}(\tilde{h}_{i,n}^2)
\]

\[
= \mathcal{E}(h_{i,n}^2) - \mu_n^2 + \mathcal{E}(\tilde{h}_{i,n}^2) - \mu_n^2
\]

\[
= 2\sigma_n^2.
\]

By the Cauchy–Schwarz inequality,

\[
\mathcal{E}((h_{i,n} - \tilde{h}_{i,n})(h_{i,n+1} - \tilde{h}_{i,n+1})) \leq 2\sigma_n\sigma_{n+1},
\]

Proc. R. Soc. A (2008)
Thus, by lemmas 4.4, 4.6 and 4.7 we obtain
\[
2(\sigma_n - \sigma_{n+1})^2 = 2\sigma_n^2 + 2\sigma_{n+1}^2 - 4\sigma_n\sigma_{n+1} \leq \mathcal{E}(h_{i,n} - \tilde{h}_{i,n})^2 + \mathcal{E}(h_{i,n+1} - \tilde{h}_{i,n+1})^2 - 2\mathcal{E}(h_{i,n} - \tilde{h}_{i,n})(h_{i,n+1} - \tilde{h}_{i,n+1}).
\] (4.41)

Thus, we have
\[
2(\sigma_n - \sigma_{n+1})^2 \leq \mathcal{E}(((h_{i,n} - \tilde{h}_{i,n}) - (h_{i,n+1} - \tilde{h}_{i,n+1}))^2)
\leq 2\mathcal{E}((h_{i,n} - h_{i,n+1})^2 + (\tilde{h}_{i,n} - \tilde{h}_{i,n+1})^2)
\]
and because $\mathcal{X}$ and $\tilde{\mathcal{X}}$ are identically distributed,
\[
(\sigma_n - \sigma_{n+1})^2 \leq 2\mathcal{E}(h_{i,n} - h_{i,n+1})^2.
\] (4.44)

Finally, because the $h_{i,n}$ are identically distributed,
\[
(\sigma_n - \sigma_{n+1})^2 \leq \frac{2}{n}\mathcal{E}\left(\sum_{i=1}^{n} (h_{i,n} - h_{i,n+1})^2\right)
\] (4.45)
and (ii) follows by lemmas 4.3 and 4.4.

\[\blackslug\]

(e) An upper bound on $\text{var}(S_n^*)$

**Proof of lemma 3.2.** We aim to show that for all $n \geq 16k$,
\[
\mathcal{E}(S_n^* - S_{n+1}^*)^2 \leq 2(3 + 8k\beta)(3 + 64k)(\kappa_n + \kappa_{n+1}).
\] (4.46)

First, we write
\[
S_n^* - S_{n+1}^* = \left(\frac{S_n - n\mu_n}{\sigma_n}\right) - \left(\frac{S_{n+1} - (n + 1)\mu_{n+1}}{\sigma_{n+1}}\right)
\] (4.47)
\[
= \frac{1}{\sigma_n}\left(S_n - n\mu_n - \frac{\sigma_n}{\sigma_{n+1}}(S_{n+1} - (n + 1)\mu_{n+1})\right)
\] (4.48)
\[
= \frac{1}{\sigma_n}\left((S_n - S_{n+1}) - n(\mu_n - \mu_{n+1}) + \mu_{n+1}
- \frac{1}{\sigma_{n+1}}(\sigma_n - \sigma_{n+1})(S_{n+1} - (n + 1)\mu_{n+1})\right).
\] (4.49)

from which it follows that:
\[
\mathcal{E}(S_n^* - S_{n+1}^*)^2 \leq \frac{4}{\sigma_n^2}\left(\mathcal{E}(S_n - S_{n+1})^2 + n^2(\mu_n - \mu_{n+1})^2 + \mu_{n+1}^2
+ \frac{1}{\sigma_{n+1}^2}(\sigma_n - \sigma_{n+1})^2 \text{var}(S_{n+1})\right).
\] (4.50)

Thus, by lemmas 4.4, 4.6 and 4.7 we obtain
\[
\mathcal{E}(S_n^* - S_{n+1}^*)^2 \leq 4(3 + 8k\beta)\left(\frac{\lambda_n}{\sigma_n^2}\right)\left(3 + 64k\left(\frac{\lambda_{n+1}}{\sigma_{n+1}^2}\right)\right).
\] (4.51)
Finally, because $\lambda_n/\sigma_n^2 \leq \kappa_n^{1/2}$ and $\kappa_{n+1} \geq 1$,
\[ \mathcal{E}(S_n^*-S_{n+1}^*)^2 \leq 4(3+8k\beta)(3+64k)\kappa_n^{1/2}\kappa_{n+1}^{1/2}, \] (4.52)
and the result follows by the fact that $2\kappa_n^{1/2}\kappa_{n+1}^{1/2} \leq \kappa_n + \kappa_{n+1}$.

**Proof of theorem 1.2.** By lemma 3.2 and (3.8),
\[ \text{var}(S_n^*) \leq 2(n+1)(3+8k\beta)(3+64k)(\kappa_n + \kappa_{n+1}). \] (4.53)
\[ \square \]

(f) Strong convergence

**Proof of theorem 1.1.** We aim to show that $|H_n^*| \to 0$ a.s. as $n \to \infty$. Suppose that there exists a finite constant $c > 0$ such that $\kappa_n < c$ for all $n \in \mathbb{N}$. Then by corollary 3.3, the probabilities $P(|H_n^*| > \epsilon)$ are summable for the subsequence $n_m = m^2$,
\[ \sum_{m=1}^{\infty} P(|H_{m^2}^*| > \epsilon) \leq \frac{c'}{\epsilon^2} \sum_{m=1}^{\infty} \frac{1}{m^2} < \infty, \] (4.54)
where $c' = 4(c(3+8k\beta)(3+64k)$. Hence, by the Borel–Cantelli lemma it follows that $H_{m^2}^* \to 0$ a.s. as $m \to \infty$. For $m^2 < n < (m+1)^2$, we write
\[ H_n^* = \frac{1}{n}(S_{m^2}^* + (S_{n}^* - S_{m^2}^*)) = \frac{m^2}{n} \left( \frac{H_{m^2}^*}{m^2} + \frac{1}{m^2}(S_n^* - S_{m^2}^*) \right). \] (4.55)
Let
\[ W_m = \max \left\{ \frac{1}{m^2} \left| S_n^* - S_{m^2}^* \right| : m^2 < n < (m+1)^2 \right\}. \] (4.56)
Then $|H_n^*| \leq |H_{m^2}^*| + W_m$ for all $m^2 < n < (m+1)^2$, so by (4.54) it is sufficient to show that $W_m \to 0$ a.s. as $m \to \infty$. Writing
\[ (S_n^* - S_{m^2}^*) = (S_n^* - S_{n-1}^*) + (S_{n-1}^* - S_{n-2}^*) + \cdots + (S_{m^2+1}^* - S_{m^2}^*), \] (4.57)
we see that
\[ W_m \leq \frac{1}{m^2} \sum_{j=m^2+1}^{(m+1)^2-1} |S_j^* - S_{j-1}^*|. \] (4.58)
Hence by the Cauchy inequality, and using the fact that the sum has exactly $2m$ terms, it follows by lemma 3.2 that:
\[ \mathcal{E}(W_m^2) \leq \frac{1}{m^2} \mathcal{E} \left( \sum_{j=m^2+1}^{(m+1)^2-1} |S_j^* - S_{j-1}^*| \right)^2 \] (4.59)
\[ \leq \frac{2}{m^2} \sum_{j=m^2+1}^{(m+1)^2-1} \mathcal{E}(S_j^* - S_{j-1}^*)^2 \] (4.60)
\[ \leq \frac{16c(3+8k\beta)(3+64k)}{m^2}. \] (4.61)
Thus, by Markov’s inequality and the Borel–Cantelli lemma, \( W_m \to 0 \) a.s. as \( m \to \infty \), which concludes the proof of theorem 1.1.

\[ 5. \text{ Conclusion} \]

For a sequence \( X = X_1, X_2, \ldots \) of independent and identically distributed random vectors in \( \mathbb{R}^d \), we have proved a strong law of large numbers for functions of a point and its nearest neighbours in the sample \( X_n = (X_1, \ldots, X_n) \). We have also used the result to show that certain non-parametric difference-based estimators of residual moments are strongly consistent as the number of points \( n \to \infty \) (provided the distribution of the explanatory vectors satisfies smooth and positive density conditions). Perhaps the most significant advance put forward in this paper is that the random variables need not be bounded, the only requirement being that their first four moments are bounded.

In practical applications, one area of concern is that the bound on \( \text{var}(S_n^*) \) of theorem 1.2 increases exponentially with the dimension of the underlying space \( \mathbb{R}^d \), which negatively affects the rate of convergence when \( d \) is large. The exponential factor \( \beta \) is introduced in the proof of lemma 4.5, where we have used the relatively crude bound of lemma 4.2. By contrast, the proof of lemma 4.3 uses the bound of lemma 4.1, which avoids the exponential factor. If we attempt to prove lemma 4.5 along the same lines as those followed in the proof of lemma 4.3, we arrive at the expression

\[
\mathcal{E} \left( \sum_{i=1}^{n} (h_{i,n} - h_{i,n+1}) \right)^2 = \sum_{i,j=1}^{n} \mathcal{E}((h_{i,n} - h_{i,n+1})(h_{j,n} - h_{j,n+1})^2 J_{ij}),
\]

where \( J_{ij}(X) \) is the indicator variable of the event \( X_{n+1} \in B_{i,n}(X) \cap B_{j,n}(X) \), and \( B_{i,n}(X) \) and \( B_{j,n}(X) \) are, respectively, the \( k \)-nearest neighbour balls of the points \( X_i \) and \( X_j \) in the sample \( X_n \). At this point, we would like to show that \( P(X_{n+1} \in B_{i,n}(X) \cap B_{j,n}(X)) \) is at most equal to \( c(d)/n \), where \( c(d) \) does not increase exponentially with \( d \). However, Dwyer (1995) showed that if \( k = 1 \) and \( d \geq 7 \), then for \( n \gg d \) the expected number of edges \( n_e \) in the random sphere of influence graph, or equivalently the expected number of pairs \( (X_i, X_j) \) for which \( B_{i,n}(X) \cap B_{j,n}(X) \neq \emptyset \), satisfies

\[
(0.324)2^d n \leq \mathcal{E}(n_e) \leq (0.677)2^d n.
\]

This suggests that the curse of dimensionality will not be easily lifted.

The author would like to thank the Royal Society for supporting his research through the University Research Fellowship scheme.

\section*{Appendix A. Standard results}

\textit{Proof of lemma 4.1.} Let \( \omega_{i,n} \) denote the probability measure of the \( k \)-nearest neighbour ball of \( X_i \in X_n \), \( G(\omega) \) be the distribution function of \( \omega_{i,n} \) and \( B_{i}(\omega) \) be the ball centred at \( X_i \) of probability measure \( \omega \). Because the \( X_i \) are identically
Nearest neighbour statistics \( \text{3189} \)
distributed,
\[
G(\omega) = P(\omega_{i,n} \leq \omega) \tag{A1}
\]
\[
= P(B_i(\omega) \text{ contains at least } k \text{ points other than } X_i) \tag{A2}
\]
\[
= 1 - \sum_{j=0}^{k-1} P(B_i(\omega) \text{ contains exactly } j \text{ points other than } X_i) \tag{A3}
\]
\[
= 1 - \sum_{j=0}^{k-1} \binom{n-1}{j} \omega^j (1-\omega)^{n-j-1}. \tag{A4}
\]
Integrating by parts,
\[
\mathcal{E}(\omega_{i,n}^k) = \int_0^1 \omega^\alpha \, dG(\omega) = 1 - \alpha \int_0^1 \omega^{\alpha-1} G(\omega) d\omega \tag{A5}
\]
\[
= \alpha \sum_{j=0}^{k-1} \binom{n-1}{j} \int_0^1 \omega^{j+\alpha-1} (1-\omega)^{n-j-1}. \tag{A6}
\]
The integral in (A6) is the beta function \( B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b) \) with parameters \( a = j+\alpha \) and \( b = n-j \). Thus, we obtain
\[
\mathcal{E}(\omega_{i,n}^k) = \frac{\alpha \Gamma(n)}{\Gamma(n+\alpha)} \sum_{j=1}^{k-1} \frac{\Gamma(j+\alpha)}{\Gamma(j+1)} = \frac{\Gamma(n)\Gamma(k+\alpha)}{\Gamma(n+\alpha)\Gamma(k)}. \tag{A7}
\]
In particular, \( \mathcal{E}(\omega_{i,n}) = k/n. \)

**Lemma A.1.** For every countable set of points in \( \mathbb{R}^d \), any point can be the (first) nearest neighbour of at most \( |dv_{d,p}| \) other points of the set, where \( v_{d,p} \) is the volume (Lebesgue measure) of the unit ball in \( (\mathbb{R}^d, \| \cdot \|_p) \).

**Proof.** Let \( t > dv_{d,p} \) be an integer and suppose that \( x_0 \) is the nearest neighbour of every point in the set \( \{ x_1, \ldots, x_i \} \). We project the points \( x_i \) onto the surface of the unit ball in \( \mathbb{R}^d \), writing
\[
x_i = x_0 + r_i \xi_i \quad \text{where} \quad r_i = \| x_i - x_0 \| \quad \text{and} \quad \| \xi_i \| = 1. \tag{A8}\]
Let \( x_i \) and \( x_j \) be two distinct points of \( \{ x_1, \ldots, x_i \} \) and suppose (without loss of generality) that \( \| x_j - x_0 \| \leq \| x_i - x_0 \| \). The vector \( x_i - x_j \) can be expressed as
\[
(x_i - x_j) = r_i \xi_i - r_j \xi_j = (r_i - r_j) \xi_i + r_j (\xi_i - \xi_j), \tag{A9}\]
so
\[
\| x_i - x_j \| \leq (r_i - r_j) \| \xi_i \| + r_j \| \xi_i - \xi_j \|. \tag{A10}\]
Hence, because \( r_i \leq \| x_i - x_j \| \) and \( \| \xi_i \| = 1 \) we have
\[
r_i \leq (r_i - r_j) + r_j \| \xi_i - \xi_j \|. \tag{A11}\]
and therefore
\[ 1 \leq \|\xi_i - \xi_j\|. \quad (A12) \]

Thus, every point \( \xi_i \) must be located within an otherwise empty region on the surface of the unit ball in \( \mathbb{R}^d \), and the surface area of each of these regions must be at least equal to 1. By hypothesis, there are \( t > d_{v_{d,p}} \) points \( \xi_1, \ldots, \xi_t \), so the total surface area covered by these disjoint regions must be greater than \( d_{v_{d,p}} \). However, because the total surface area of the unit ball is equal to \( d_{v_{d,p}} \) we have a contradiction, and thus we conclude that \( t \leq [d_{v_{d,p}}] \). □

Lemma 8.4 of Yukich (1998) shows that lemma 4.2 follows lemma A.1. An alternative proof is given here.

**Proof of lemma 4.2.** Let \( t > k d_{v_{d,p}} \) be an integer; suppose that \( x_0 \) is one of the \( k \)-nearest neighbours of every point in the set \( \{x_1, \ldots, x_t\} \). First we choose \( x_1 \) to be the point of \( \{x_1, \ldots, x_t\} \) furthest away from \( x_0 \), then eliminate this point \( x_1 \) along with all points that are closer to \( x_1 \) than \( x_0 \) is to \( x_1 \). At least one point is eliminated (namely the point \( x_1 \) itself), and because \( x_0 \) is one of the \( k \)-nearest neighbours of \( x_1 \) there can be at most \( k - 1 \) other points closer to \( x_1 \) than \( x_0 \) is to \( x_1 \), so at most \( k \) points are eliminated in total. Next we repeat the procedure on remaining points, choosing \( x_2 \) to be the furthest point away from \( x_0 \), then eliminating \( x_2 \) along with all points that are closer to \( x_2 \) than \( x_0 \) is to \( x_2 \). Because \( x_2 \) was not eliminated in the first round, it must be closer to \( x_0 \) than it is to \( x_1 \),
\[ \|\xi_2 - x_0\| \leq \|\xi_2 - \xi_1\|. \quad (A13) \]
We continue in this way, at each stage choosing \( \xi_i \) to be the point (among those that remain) furthest away from \( x_0 \), and eliminating \( \xi_i \) along with all points that are closer to \( \xi_i \) than \( x_0 \) is to \( \xi_i \). Because \( \xi_i \) was not eliminated in the previous rounds, it must be closer to \( x_0 \) than it is to any of the points chosen previously,
\[ \|\xi_i - x_0\| \leq \|\xi_i - \xi_j\| \quad \text{for all} \quad j = 1, \ldots, i - 1. \quad (A14) \]
At least one point is eliminated at each stage, so the process must eventually terminate, say after \( T \) steps, and we obtain a set of points \( \{\xi_1, \ldots, \xi_T\} \), each having \( x_0 \) as its nearest neighbour. At most \( k \) points are eliminated at each step, so a minimum of \( t/k \) steps must be performed before the process terminates. By hypothesis, \( t > k d_{v_{d,p}} \) so \( T > d_{v_{d,p}} \). However, by lemma A.1 we know that any point \( x_0 \) can be the nearest neighbour of at most \( [d_{v_{d,p}}] \) other points. Thus we have a contradiction, and conclude that \( t \leq k d_{v_{d,p}} \). □

For Euclidean space, Zeger & Gersho (1994) establish an alternative bound in terms of kissing numbers.

**Corollary A.2.** If \( p = 2 \) then for every countable set of points in \( \mathbb{R}^d \), any point can be among the \( k \)-nearest neighbours of at most \( kK(d) \) other points of the set, where \( K(d) \) is the maximum kissing number in \( \mathbb{R}^d \).

**Proof.** Following the proof of lemma A.1, by (A12) we can place a set of \( t \) non-overlapping spheres of radius \( 1/2 \) at each point \( \xi_i \), and each of these will be tangent to the sphere of radius \( 1/2 \) centred at the origin. This contradicts the fact that there can be at most \( K(d) \) such tangent spheres. □
It is known that $K(d)$ satisfies the following inequalities, where the upper bound is due to Kabatiansky & Levenshtein (1978) and the lower bound is due to Wyner (1965),

$$2^{0.2075(d(1+o(1))} \leq K(d) \leq 2^{0.401d(1+o(1))}.$$ 

References

Avram, F. & Bertsimas, D. 1993 On central limit theorems in geometrical probability. Ann. Appl. Prob. 3, 1033–1046. (doi:10.1214/aoap/1177005271)

Baldi, P. & Rinott, Y. 1989 On normal approximations of distributions in terms of dependency graphs. Ann. Prob. 17, 1646–1650. (doi:10.1214/aop/1176991178)

Bickel, P. J. & Breiman, L. 1983 Sums of functions of nearest neighbour distances, moment bounds, limit theorems and a goodness of fit test. Ann. Prob. 11, 185–214. (doi:10.1214/aop/1176993668)

Chen, L. H. Y. & Shao, Q. 2004 Normal approximation under local dependence. Ann. Prob. 32, 1985–2028. (doi:10.1214/009117904000000450)

Dwyer, R. A. 1995 The expected size of the sphere-of-influence graph. Comput. Geomet. Theory Appl. 5, 155–164. (doi:10.1016/0925-7721(94)00025-Q)

Efron, B. & Stein, C. 1981 The jackknife estimate of variance. Ann. Stat. 9, 586–596. (doi:10.1214/aos/1176345462)

Evans, D. & Jones, A. J. 2008 Non-parametric estimation of residual moments and covariance. Proc. R. Soc. A 464, 2831–2846. (doi:10.1098/rspa.2007.0195)

Evans, D., Jones, A. J. & Schmidt, W. M. 2002 Asymptotic moments of near neighbour distance distributions. Proc. R. Soc. A 458, 2839–2849. (doi:10.1098/rspa.2002.1011)

Gruber, P. M. 2004 Optimum quantization and its applications. Adv. Math. 186, 456–497. (doi:10.1016/j.aim.2003.07.017)

Hitczenko, P., Janson, S. & Yukich, J. E. 1999 On the variance of the random sphere of influence graph. Random Struct. Algor. 14, 139–152. (doi:10.1002/(SICI)1098-2418(199903)14:2

Kabatiansky, G. A. & Levenshtein, V. I. 1978 Bounds for packings on a sphere and in space. Probl. Peredachi Inf. 1, 3–25.

Penrose, M. D. 2000 Central limit theorems for k-nearest neighbour distances. Stoch. Proc. Appl. 85, 295–320. (doi:10.1016/S0304-4149(99)00080-0)

Penrose, M. D. 2007 Laws of large numbers in stochastic geometry with statistical applications. Bernoulli 13, 1124–1150. (doi:10.3150/07-BEJ5167)

Penrose, M. D. & Wade, A. R. 2008 Limit theory for the random on-line nearest-neighbor graph. Random Struct. Algor. 32, 125–156. (doi:10.1002/rsa.20185)

Penrose, M. D. & Yukich, J. E. 2001 Central limit theorems for some graphs in computational geometry. Ann. Appl. Prob. 11, 1005–1041. (doi:10.1214/aoap/1015345393)

Penrose, M. D. & Yukich, J. E. 2003 Weak laws of large numbers in geometric probability. Ann. Appl. Prob. 13, 277–303. (doi:10.1214/aoap/1042765669)

Peres, A. G. & Martin, O. C. 1998 Scaling universalities of kth nearest neighbor distances on closed manifolds. Adv. Appl. Math. 21, 424–436. (doi:10.1006/aama.1998.0607)

Petrovskaya, M. & Leontovich, A. 1982 The central limit theorem for a sequence of random variables with a slowly growing number of dependencies. Theory Probab. Appl. 27, 815–825. (doi:10.1137/1127089)

Reitzner, M. 2003 Random polytopes and the Efron–Stein jackknife inequality. Ann. Prob. 31, 2136–2166. (doi:10.1214/aop/106864381)

Steele, J. M. 1986 An Efron–Stein inequality for nonsymmetric statistics. Ann. Stat. 14, 753–758. (doi:10.1214/aos/1176349952)
Stone, C. J. 1977 Consistent nonparametric regression. *Ann. Stat.* **5**, 595–620. (doi:10.1214/aos/1176343886)

Wade, A. R. 2007 Explicit laws of large numbers for random nearest-neighbour-type graphs. *Adv. Appl. Prob.* **39**, 326–342. (doi:10.1239/aap/1183667613)

Wyner, A. D. 1965 Capabilities of bounded discrepancy decoding. *Bell Syst. Tech. J.* **44**, 1061–1122.

Yukich, J. E. 1998 *Probability theory of classical Euclidean optimization problems*. Springer Lecture Notes in Mathematics, no. 1675. Berlin, Germany: Springer.

Zeger, K. & Gersho, A. 1994 The number of nearest neighbors in a Euclidean code. *IEEE Trans. Inform. Theory* **40**, 1647–1649. (doi:10.1109/18.333884)