The Gamma Test

Data-derived estimates of noise for unknown smooth models using near-neighbour asymptotics

by

Dafydd Evans
Department of Computer Science
Cardiff University
University of Wales

A thesis submitted in partial fulfilment of the requirement
for the degree of Doctor of Philosophy

1 March 2002

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ..... (candidate)

Date .....

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by citations and footnotes giving explicit references. A bibliography is appended.

Signed ..... (candidate)

Date .....

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

Abstract

The Gamma test is a simple technique for assessing the extent to which a given set of M data points can be modelled by an unknown smooth non-linear function f. If the underlying model is of the form $y = f(\boldsymbol{x}) + r$ where r is a random variable representing that part of the data which cannot be accounted for by the smooth function f, the Gamma test produces an estimate Γ_M for the variance $\mathrm{Var}(r)$. This estimate is rapidly computed directly from the data, and since its introduction in 1995 has been used extensively for a variety of different applications in several theses and papers. Thus it is of some interest to provide a formal basis for the method, and this is precisely the problem addressed in this thesis.

The Gamma test is based on the behaviour of certain near neighbour statistics as the number of data points M becomes large. Our analysis involves determining the probabilistic asymptotic behaviour of the mean squared kth nearest neighbour distance in a set of M points, and other related sums. We develop new techniques for near neighbour functions on sets of points sampled from a compact convex body in $\mathbb{R}^m$, the study of what we have chosen to call L-dependent variables, and some quite technical generalisations of earlier ideas of Bickel and Breiman [1983].

Using these techniques we are able to produce some quite interesting incidental results, but the main result is that for sets of points selected from a compact convex body in $\mathbb{R}^m$ according to some smooth positive sampling density then the Gamma statistic Γ_M converges in probability to the noise variance $\mathrm{Var}(r)$ as $M \to \infty$. While we believe that the Gamma test has wider applicability, this result is sufficient to justify the test in a wide variety of practical applications.

Acknowledgements

I would like to thank Professor Antonia J. Jones for her supervision and support towards the success of this research and for her continuous inspiration. I would also like to thank everybody in Professor Jones' research group at Cardiff University, and especially Tina at the Research Farm.

Thanks are also due to Professor Wolfgang Schmidt of the Department of Mathematics at University of Colorado, Professor Andrew Pollington of the Department of Mathematics, Brigham Young University, Utah and Dr Tom Westerdale of Birkbeck College, University of London, each of whom has provided significant assistance at various points of the project.

Most importantly I would like to thank my beautiful wife Emma who supported me through some extremely difficult times and without whom the project would never have been completed. This thesis is dedicated to Emma and to our wonderful sons Mabon and Gethin.

I also acknowledge the support of the EPSRC, studentship number: 98700188.

Dafydd Evans

1 March 2002

Contents

List of Figures

Index of Notation

We list the mathematical symbols whose meaning remain the same throughout the text, along with the page numbers where they first appear.

Chapter 1

Introduction

The Gamma test [Stefánsson *et al.* 1997; Končar 1997] is a data–analysis technique originally developed as a tool to assist in the construction of data-derived models. The method has subsequently been successfully applied to problems in control theory [Končar 1997], feature selection [Chuzhanova *et al.* 1998; Durrant 2001], secure communications [de Oliveira 1999] and controlling chaotic systems [Tsui 1999; Tsui *et al.* 2002; Jones *et al.* 2002].

The aim of this thesis is to establish a rigorous mathematical foundation for the Gamma test and to provide a detailed analysis of the conditions under which it may be shown to be applicable.

1.1 Data–derived modelling

Data–derived modelling techniques seek to construct models of a system directly from a set of measurements of the system's behaviour, without assuming any *a priori* knowledge of the underlying logical rules or equations that determine this behaviour. Without further assumptions the class of potential models is enormous, ranging from logic functions through rule based systems and probabilistic models to parameterised functions. In this thesis we shall concentrate on the case where the underlying system behaviour is an unknown *smooth* function.

Neural networks, trained by some variant of back-propagation, may be considered as the generic example of a non-parametric smooth modelling technique. While the methods discussed in this thesis are entirely independent of the particular modelling technique

employed, they have nevertheless proved to be extremely useful when applied to model construction using neural networks.

We restrict our attention to those systems which may be thought of as smoothly transforming some *input* vector into a corresponding *output*. This is a fairly general representation – in the case of a smooth dynamical system the current state of the system (possibly together with some of its previous states) may be thought of as the input, with the output representing the transformed state of the system after some time interval has elapsed.

Suppose we have a set of input–output observations of the form

$$\{(\boldsymbol{x}_i, y_i) \,|\, 1 \leq i \leq M\} \tag{1.1}$$

where the inputs $\boldsymbol{x} \in \mathbb{R}^m$ are vectors confined to some closed bounded set $C \subset \mathbb{R}^m$ and without loss of generality, the corresponding outputs $y \in \mathbb{R}$ are scalars. In the general case where the outputs are vectors, the algorithms we consider can be applied independently to each component and as we shall see, this involves very little extra computational cost.

For our purposes a *model* is an algorithm which, using a data structure derived from the initial data set (1.1), can be used to predict the output y corresponding to a previously unseen query vector $\boldsymbol{x}$. It is an implicit requirement that the process of model construction and query should be computationally efficient. In practice this means that at worst model construction should have time complexity $O(M \log M)$ and querying the model for a single input vector should at worst have time complexity $O(\log M)$ as $M \to \infty$. A technique such as the Gamma test, which is intended to assist in the model construction process, must therefore also have time complexity of at worst $O(M \log M)$ as $M \to \infty$.

1.1.1 Noise

One problem in constructing models solely on the basis of observation is that measurements are often corrupted by *noise*. We define noise to be any component of an output that cannot be accounted for by a smooth transformation of the corresponding input.

Noise may occur in a set of observations for several reasons.

- Inaccuracy of measurement.
- Not all causative factors that influence the output are included in the input.
- The underlying relationship between input and output is not smooth.

The Gamma test is a technique for estimating the noise level present in a data set. It calculates this estimate *directly from the data* and does not assume anything regarding the parametric form of the equations governing the system. The only requirement in this direction is that the system is *smooth* – the precise conditions under which the Gamma test may be applied will be given later.

1.1.2 A non–parametric approach

The traditional approach to data–derived modelling is to make some specific assumptions regarding the form of the relationship between the input $\boldsymbol{x} \in \mathbb{R}^m$ and the corresponding output $y \in \mathbb{R}$. An attempt is then made to find the 'best fit' for the parameters in the hypothesised relationship, relative to the observed data. This approach leads to to the study of *parametric statistics*, see for example [Bates and Watts 1988].

However, in many cases we have no *a priori* knowledge with which to construct a parametric model. Traditional statistical models are then reduced to studying quantities such as correlations, auto–regressions and co–variances, all of which are likely to be very crude estimators of the 'average' causal relations between the input variables and the outputs we seek to predict.

A new approach to this general problem is presented in [Pi and Peterson 1994]. Letting f represent the optimal model of the system under investigation[1], then rather than presupposing some particular parametric form for f we suppose instead that it simply belongs to some *general class of functions*, in particular those which are *uniformly continuous* over the input space.

The relationship between an input $\boldsymbol{x}$ and the corresponding output y is then expressed as

$$y = f(\boldsymbol{x}) + r \tag{1.2}$$

where

[1]The goal of any modelling technique is to 'identify' this function in some sense.

- f is a smooth function representing the system.
- r is a random variable representing noise.

Without loss of generality, the expected value of the noise variable r may be assumed to be zero, since any constant bias can be absorbed into the unknown function f. Despite the fact that f is unknown, subject to the condition that f has bounded first and second partial derivatives over the input space C, the Gamma test provides an estimate for the variance of the noise variable, $\mathrm{Var}(r)$.

This estimate is called the *Gamma statistic*, denoted by Γ, and may be derived in $O(M \log M)$ time[2] where the implied constant depends only on the dimension m of the input space. The Gamma test is a non-parametric technique and the results apply regardless of the particular methods subsequently used to build a model.

1.2 Applications of a data–derived estimate of noise

Before describing the Gamma test in detail we outline some important applications of having an efficient technique for estimating $\mathrm{Var}(r)$.

- The Gamma test provides a method for assessing the 'quality' of the data.

If the Gamma statistic is small (relative to the variance of the output y) then it is likely that the output is determined from the inputs by some smooth model. The Gamma test can therefore tell us *directly from the data* whether or not we have sufficient data to form a smooth non–linear model, and also indicates how good that model is likely to be. If the error of prediction is too high regardless of how much data we are given, then we might increase the accuracy of our measurements or alternatively investigate whether or not we have included all input variables that are likely to affect the output. In the case of time series analysis we may also choose to increase the rate of sampling.

To illustrate the convergence of Γ to $\mathrm{Var}(r)$ as the number of points M increases we define $f(x) = \sin(x)$, generate 500 uniformly distributed points x in the range $[0, 2\pi]$ and construct the corresponding output values y by adding a uniformly distributed noise component with a variance of 0.075 to each of the $f(x)$ values.

The data points thus obtained are shown in Figure 1.1 and Figure 1.2 shows the convergence of the Gamma statistic Γ to the true noise variance of 0.075 as M increases from 30 to 500 (the dashed line shows the true noise variance).

[2] In practice on a 500MHz PC with $M \approx 1000$ and $m \approx 100$ a single Gamma test computation is complete within a few seconds.

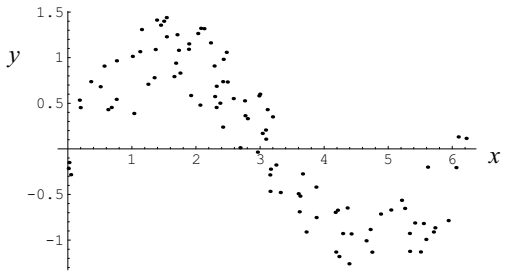

Figure 1.1: The noisy sine function $y = \sin(x) + r$ where $\mathrm{Var}(r) = 0.075$

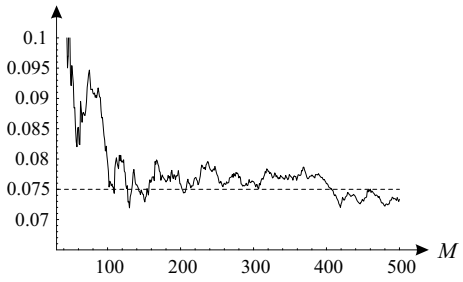

Figure 1.2: The convergence of Γ to $\mathrm{Var}(r)$ for $y = \sin(x) + r$.

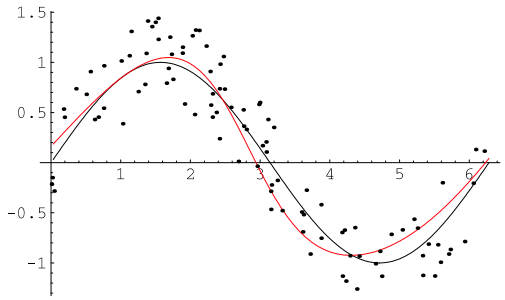

Figure 1.3: Model (red) trained to a mean squared error of 0.0786

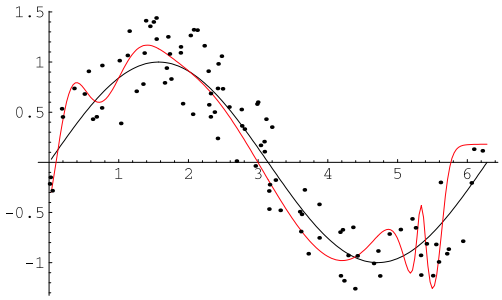

Figure 1.4: Model (red) trained to a mean squared error of 0.056

- The Gamma test provides a method for determining a suitable stage at which to stop adapting a model to fit the data.

Whatever the choice of non–linear modelling tool, one of the central problems of model construction is to determine when to stop adapting the model to fit the data.

The *mean squared error* between the actual (measured) output values and those predicted by the model provides one indicator of how well the model fits the training data. If we fit a model beyond the point at which the mean squared error over the training data falls significantly below the noise level, we will have incorporated some aspect of the noise into the model itself. The model will then perform poorly on previously unseen inputs despite the fact that its performance on the training data may be almost perfect. Such models have effectively memorised the training data and are consequently said to suffer from *overfitting*. If we stop adapting the model when the mean squared error on the training set reaches our estimate Γ for the noise variance $\mathrm{Var}(r)$, and if we have confidence that our estimate Γ is reasonably accurate, then we should obtain a smooth model having the best possible mean squared error when used to predict the output corresponding to an input point not seen during the model construction process.

Figure 1.3 shows the model obtained by training a 1-5-5-1 neural network to a mean squared error of 0.0786 on 100 points of noisy sine data (corresponding to $\Gamma = 0.0795$), while Figure 1.4 shows the model obtained by a training a 1-5-5-1 neural network to a mean squared error of 0.056. In both cases model is plotted in red while the function $f(x) = \sin(x)$ is plotted in black.

- The Gamma test can determine the minimum number of data points required build a good model.

Suppose we compute a sequence of Gamma statistics Γ_M for an increasing number of points M (see Figure 1.2)[3]. The value Γ at which the sequence stabilises serves as our estimate for $\mathrm{Var}(r)$, and if M_0 is the number of points required for Γ_M to stabilise to within some prescribed error of Γ then we will need *at least* this number of data points to build a model that may be expected to predict with a mean squared error of Γ.

- The Gamma test provides a new technique for determining the most significant input variables[4].

In this context, the goal of *model identification* is to choose a selection of input variables that best models the output. That the Gamma test may be useful in this respect derives from the fact that low noise levels will only be encountered when all of the principal causative factors that determine the output have been included in the input. Some input variables may be irrelevant, while others may be subject to high measurement error so that incorporating them into the model will be counter-productive (leading to a higher *effective* noise level on the output). Since performing a single Gamma test is a relatively fast procedure, provided the number m of possible inputs is not too large we may compute a noise estimate for every possible subset of the input variables. The subset for which the associated noise estimate is closest to zero can then be taken as the 'best' selection of inputs.

We can see that this approach to smooth data modelling is extremely general and far reaching. Thus the issue of placing the Gamma test and its associated algorithm into an established framework is of considerable interest.

1.3 Conditions

We now state the conditions under which the Gamma test can be applied. These fall into three categories

- Conditions on the class of models f
- Conditions on the noise distribution Ψ
- Conditions on the sampling distribution Φ

We remark that some of the requirements described below are only necessary for the purpose of obtaining a proof of the algorithm[5]. Experimental results show that the algorithm continues to work under less restrictive conditions.

[3] This is known the M-test.

[4] In conventional statistical analysis this problem is often addressed by a 'principal components analysis' which suffers from the essential limitation that it is based on an 'average' *linear* model.

[5] Although we loosely speak of obtaining a 'proof' of the Gamma test, we acknowledge that this is an abuse of terminology and that in fact we are proving the *consistency* of the algorithm.

1.3.1 The class of models f

The domain of possible models $f : \mathbb{R}^m \to \mathbb{R}$ is restricted to the class of continuous functions having *bounded first and second order partial derivatives* over the input space $C \subset \mathbb{R}^m$. In particular, we remark that the Gamma test is not directly applicable to problems involving categorical data.

Let $\nabla f(\boldsymbol{x})$ and $Hf(\boldsymbol{x})$ denote the first and second partial derivatives of f at the point $\boldsymbol{x} \in C$, defined by

$$\nabla f(\boldsymbol{x}) = \left(\frac{\partial f}{\partial \boldsymbol{x}(i)}\right)_{i=1}^{m} \qquad \text{and} \qquad Hf(\boldsymbol{x}) = \left(\frac{\partial^2 f}{\partial \boldsymbol{x}(i)\partial \boldsymbol{x}(j)}\right)_{i,j=1}^{m} \tag{1.3}$$

where $\boldsymbol{x}(i)$ and $\boldsymbol{x}(j)$ are the ith and jth components of $\boldsymbol{x}$ respectively. Let $\mathcal{H}(C)$ denote the convex hull of C. Then we require that there exist constants $b_1 > 0$ and $b_2 > 0$ such that for all $\boldsymbol{x} \in \mathcal{H}(C)$,

$$\textbf{F.1} \qquad |\nabla f(\boldsymbol{x})| \leq b_1 \quad \text{and} \quad |Hf(\boldsymbol{x})| \leq b_2 \tag{1.4}$$

These conditions are denoted by **F.1** for later reference, and are required in order that we can apply the second mean value theorem to the unknown function f.

1.3.2 The noise distribution Ψ

The noise r is defined to be a random variable representing that component of the output which cannot be accounted for by any smooth model f having bounded first and second partial derivatives. We denote the distribution function of the noise variable by Ψ and the corresponding density function by ψ.

Since any bias within the system can be included in the unknown smooth component f we may suppose without loss of generality that the noise distribution Ψ has mean zero. Using $\mathcal{E}_\psi$ to represent an expectation with respect to Ψ we denote this by $\mathcal{E}_\psi(r) = 0$.

As we are aiming to estimate the *variance* of the noise variable r we also impose the condition that this variance is finite. For technical reasons the third and fourth moments of the noise are also required to be finite.

$$\textbf{N.1} \quad \begin{cases} \mathcal{E}_\psi(r) = 0 \\ \mathcal{E}_\psi(r^2) = \mathrm{Var}(r) < \infty \\ \mathcal{E}_\psi(r^3) < \infty \\ \mathcal{E}_\psi(r^4) < \infty \end{cases} \tag{1.5}$$

We impose two further conditions on the noise distribution as follows.

N.2 The noise is *independent* of the input corresponding to the output on which it is measured. Thus we assume that the noise on an output is *homogeneous* over the input space.

N.3 The noise values r_i and r_j on two different outputs y_i and y_j $(i \neq j)$ are *independent* – in particular they are *uncorrelated* so that $\mathcal{E}_\psi(r_i r_j) = 0$ if $i \neq j$.

If the noise is *not* homogeneous over the input space, this is not necessarily fatal in a practical application – the Gamma test will return an estimate for the *average* noise variance and is therefore still able to provide useful information regarding the selection of relevant inputs (see [Jones *et al.* 2001]).

Non-linear coordinate transformations and their effect on noise

Before using any data–derived modelling technique we can attempt to reduce the dimension of input vectors by applying a transformation to their component variables, for example

$$\begin{array}{rcl} \widetilde{x}_1 & = & g_1(x_1, \ldots, x_m) \\ \widetilde{x}_2 & = & g_2(x_1, \ldots, x_m) \\ \vdots & \vdots & \vdots \\ \widetilde{x}_t & = & g_t(x_1, \ldots, x_m) \end{array}$$

where $t < m$. If the input points $\boldsymbol{x} = (x_1, \ldots, x_m)$ are confined to some lower dimensional manifold then this transformation may be locally invertible. Locally invertible transformations which effect dimensional reduction are extremely useful in simplifying the process of non-linear model construction. However, if the noise on an output is a consequence of measurement error on the inputs then we should be aware that such a transformation will also have an effect on the local noise distribution. The effect may be to invalidate our hypothesis that the noise distribution is homogeneous over the input space, in which case the Gamma test returns the average noise variance over the whole input space.

In fact, the Gamma test has proved useful in the comparative evaluation of alternative non-linear simplifying transformations (see [Končar 1997] for an application to control systems). Finding a simplifying transformation that preserves the essential features of a model is important in the study of non-linear dynamical systems.

An example of non–homogeneous noise - toxin levels in fish

As an example of non-homogeneous noise [Bates and Watts 1988] we might consider the concentration of polychlorinated biphenyls (PCBs) in trout as a function of their age (the data is taken from SCIENCE **117**:1192-1193, 1972). The plot of PCB concentration (ppm) against age (years) is given in Figure 1.5 and reveals a non-linear relationship. Moreover, we see that there is increasing noise in the PCB concentration as the age of the fish increases.

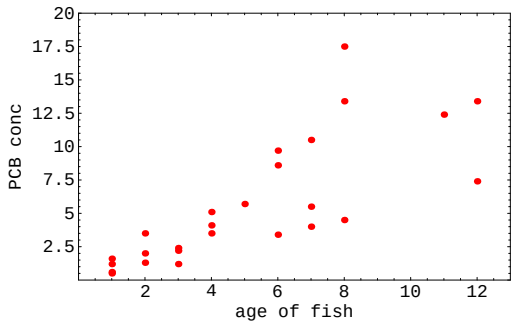

Figure 1.5: PCB concentration against age of fish.

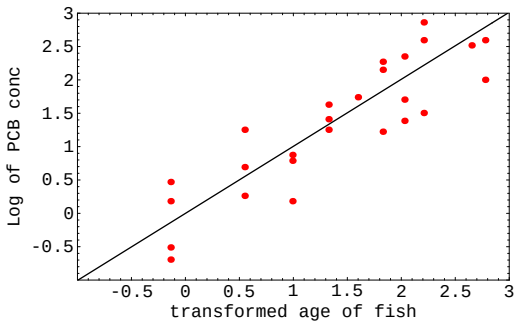

Figure 1.6: Log of PCB concentration against transformed age of fish.

It is suggested in [Bates and Watts 1988] that a suitable transformation for stabilising the noise variance is given by

$$\log y = \alpha + \beta x^{\gamma} \tag{1.6}$$

where x represents the age of a fish and y represents its PCB concentration. We determine the parameters α, β, and γ by a least-squares fit, which yields

$$\alpha = -5.268, \qquad \beta = 5.131, \qquad \gamma = 0.181 \tag{1.7}$$

The transformed data and the associated regression line are shown in Figure 1.6. We see that the noise variance is much more stable in the new co-ordinate system. Although this transformation does not reduce input-space dimension it is nevertheless an interesting illustration of how a non-linear change of co-ordinate system can affect the distribution of the noise.

1.3.3 The sampling distribution Φ

So far we have ignored the question of how the input points $\boldsymbol{x}_i \in C$ are generated. In some situations it might be possible for the experimenter to set values for the inputs and then proceed to measure the corresponding output. However, in most cases of interest the input and output values are generated autonomously by the process we are seeking to model, so their selection is not entirely within our control.

In general we suppose that the input points are selected from $\mathbb{R}^m$ according to some *sampling distribution function* Φ, and denote the corresponding density function by ϕ. The *input space* is then defined to be the set

$$C = \{\boldsymbol{x} \in \mathbb{R}^m \,|\, \phi(\boldsymbol{x}) > 0\} \tag{1.8}$$

As we shall see, the Gamma test computes its estimate for the variance of the noise by considering the distance between *nearest neighbours* in the input space. Moreover, it requires that these distances become progressively smaller as the number of input points increases. Thus we require that the input space C is a *perfect* set – closed and bounded and containing *no isolated points*. We state this for future reference as

C.0 The input space $C \subset \mathbb{R}^m$ is a perfect set of finite diameter.

In practice, the condition that C must contain no isolated points may be a bit restrictive because any closed subset of $\mathbb{R}^m$ can have *at most* countably many isolated points, so the probability of selecting an isolated point is essentially zero.

This follows by the Cantor–Bendixon theorem, which states that every uncountable closed set C in $\mathbb{R}^m$ can be expressed as $C = A \cup B$ where A is perfect and B is countable. To see this, note that every isolated point in C can be centred within a ball of positive radius containing no other point of C. Since these balls are disjoint and because each contains a point with rational coordinates then there can be at most countably many of them.

In a wide class of situations the support C has positive Lebesgue measure (in which case it must have integral dimension equal to the dimension of the imbedding space $\mathbb{R}^m$). Given a reasonably well–behaved sampling distribution we show that this case is manageable theoretically. A simple case of this situation is when the sampling distribution is *uniform* over some closed bounded subset of $\mathbb{R}^m$.

Another case of interest is the analysis of chaotic time series. Here, following [Takens 1981] we seek to model a time series by predicting the next value (the output) based on a number m of previous values (the input vector). For many chaotic time series this relationship is smooth so the Gamma test might reasonably be applied. In this case the sampling distribution is determined by an ergodic process over a set C of zero Lebesgue measure in $\mathbb{R}^m$ but having positive Hausdorff dimension $d < m$. A considerable body of experimental evidence ([Tsui 1999], [de Oliveira 1999]) strongly suggests that whether C is of positive Lebesgue measure or of zero Lebesgue measure (and positive Hausdorff dimension) is largely irrelevant to the estimates of noise variance returned by the Gamma test. Indeed, in the zero measure case the Gamma test may be *more* efficient (in terms of the number of data points required) because the input data is confined to some lower dimensional subset of the full input space.

1.4 The Gamma test

The Gamma test works by exploiting the hypothesised continuity of the unknown function f. If two points $\boldsymbol{x}$ and $\boldsymbol{x}'$ are close together in input space, the continuity of f implies that the points $f(\boldsymbol{x})$ and $f(\boldsymbol{x}')$ will be close together in output space. If the corresponding output values y and y' are *not* close together in output space, this can only be due to the influence of noise on $f(\boldsymbol{x})$ and $f(\boldsymbol{x}')$.

Let $\boldsymbol{x}$ and $\boldsymbol{x}'$ be any two points of the input space C. By (1.2) we have that $y = f(\boldsymbol{x})+r$ and $y' = f(\boldsymbol{x}') + r'$, and hence

$$\frac{1}{2}(y' - y)^2 = \frac{1}{2}\left((r' - r) + (f(\boldsymbol{x}') - f(\boldsymbol{x}))\right)^2 \tag{1.9}$$

The continuity of f implies that

$$|f(\boldsymbol{x}') - f(\boldsymbol{x})| \to 0 \quad \text{as} \quad |\boldsymbol{x}' - \boldsymbol{x}| \to 0 \tag{1.10}$$

so by (1.9) we obtain

$$\frac{1}{2}(y' - y)^2 \to \frac{1}{2}(r' - r)^2 \quad \text{as} \quad |\boldsymbol{x}' - \boldsymbol{x}| \to 0 \tag{1.11}$$

Since the expected value of r is zero and since r and r' are assumed to be independent and identically distributed we have

$$\mathcal{E}\left(\frac{1}{2}(r' - r)^2\right) = \mathrm{Var}(r) \tag{1.12}$$

Hence, taking the expectation of both sides in (1.11) it follows that

$$\mathcal{E}\left(\frac{1}{2}(y' - y)^2\right) \to \mathrm{Var}(r) \quad \text{as} \quad |\boldsymbol{x}' - \boldsymbol{x}| \to 0 \tag{1.13}$$

In fact, if the points $\boldsymbol{x}$ and $\boldsymbol{x}'$ are identical then $\mathcal{E}\left(\frac{1}{2}(y' - y)^2\right) = \mathrm{Var}(r)$. However, given any finite data set we cannot make the distance $|\boldsymbol{x}' - \boldsymbol{x}|$ between two input points $\boldsymbol{x}$ and $\boldsymbol{x}'$ arbitrarily small and thus we cannot evaluate the limit (1.13) directly.

1.4.1 The Delta test

The Gamma test has its origin in the Delta test [Pi and Peterson 1994], where it is observed that the conditional expected value of $\frac{1}{2}(y' - y)^2$, on hypothesis that the associated input points $\boldsymbol{x}'$ and $\boldsymbol{x}$ are located within a distance δ of each other, converges to $\mathrm{Var}(r)$ as δ approaches zero, i.e.

$$\mathcal{E}\left(\frac{1}{2}(y' - y)^2 \,\middle|\, |\boldsymbol{x}' - \boldsymbol{x}| < \delta\right) \to \mathrm{Var}(r) \quad \text{as} \quad \delta \to 0 \tag{1.14}$$

The Delta test is based on the fact that for any particular value of δ, the expectation in (1.14) can be estimated by the sample mean

$$E(\delta) = \frac{1}{|I(\delta)|} \sum_{(i,j) \in I(\delta)} \frac{1}{2}|y_j - y_i|^2 \tag{1.15}$$

where

$$I(\delta) = \{(i,j) \,|\, |\boldsymbol{x}_j - \boldsymbol{x}_i| < \delta, 1 \leq i \neq j \leq M\} \tag{1.16}$$

is the set of index pairs (i,j) for which the associated points $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ are located within distance δ of each other.

At first sight, (1.14) suggests that computing the sample mean $E(\delta)$ for δ sufficiently small will provide us with a reasonably good estimate for $\mathrm{Var}(r)$. A problem arises due to the fact that given only finitely many data points, the number of pairs $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ satisfying $|\boldsymbol{x}_j - \boldsymbol{x}_i| < \delta$ decreases as δ decreases. Choosing a small value of δ therefore implies that the corresponding sample mean $E(\delta)$ is computed over a small number of pairs y_i and y_j, so $E(\delta)$ is subject to significant *sampling error* for small δ. Any δ must therefore be chosen sufficiently large, say $\delta > \delta_0$, to ensure that the sample mean $E(\delta)$ provides a good estimate for the expectation in (1.14). Clearly, this restriction reduces the effectiveness of $E(\delta)$ as an estimate for $\mathrm{Var}(r)$.

If we knew the parametric form of the relationship between δ and the corresponding sample mean $E(\delta)$ we could compute the $E(\delta)$ for a range of values of $\delta > \delta_0$, then estimate the limit as $\delta \to 0$ using some kind of regression technique. However, a parametric form for the relationship between δ and $E(\delta)$ is not apparent. A further problem is that the computational cost of finding the index list (1.16) is of order $O(M^2)$ where M is the number of data points. This restricts the usefulness of the Delta test in practical applications.

The Gamma test addresses these problems by defining quantities analogous to δ and $E(\delta)$, denoted by δ and γ respectively. These are based on the *nearest neighbour* structure of the input points $\boldsymbol{x}_i$ in such a way that

$$\gamma \to \mathrm{Var}(r) \quad \text{as} \quad \delta \to 0 \tag{1.17}$$

The first advantage of this approach is that the computation time is reduced to $O(M \log M)$, a significant improvement over the $O(M^2)$ time required by the Delta test. More importantly, in this case we *are* able to establish a parametric form for the relationship between γ and δ, and a regression technique can therefore be employed to estimate the limit in (1.17).

1.4.2 Nearest neighbours

Returning to (1.9) and (1.13), since f is continuous and has bounded first and second partial derivatives over the input space C, and since C is assumed to be closed and bounded, then by the second mean value theorem of the differential calculus we may write

$$f(\boldsymbol{x}') - f(\boldsymbol{x}) = (\boldsymbol{x}' - \boldsymbol{x}).\nabla f(\boldsymbol{x}) + O(|\boldsymbol{x}' - \boldsymbol{x}|^2) \quad \text{as} \quad |\boldsymbol{x}' - \boldsymbol{x}| \to 0 \tag{1.18}$$

Hence, by (1.13) we see that the error involved in estimating $\mathrm{Var}(r)$ by the expectation $\mathcal{E}(\frac{1}{2}(y' - y)^2)$ essentially depends only on the distance $|\boldsymbol{x}' - \boldsymbol{x}|$ between the points $\boldsymbol{x}$ and $\boldsymbol{x}'$. This error will be minimised if we consider pairs of output values y and y' for which the corresponding inputs $\boldsymbol{x}$ and $\boldsymbol{x}'$ are *nearest neighbours* in the input space C (i.e. those pairs of points $\boldsymbol{x}$ and $\boldsymbol{x}'$ for which the distance $|\boldsymbol{x}' - \boldsymbol{x}|$ is minimised).

Let $\boldsymbol{x}'_i$ denote the nearest neighbour of $\boldsymbol{x}_i$ among the input points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$, defined to be a point that is closer to $\boldsymbol{x}_i$ than any other. The nearest neighbour relations for a set of M points can be found in $O(M \log M)$ time using the *kd*–tree technique developed by J. L. Bentley [Bentley 1975].

By (1.13) we may compute an estimate for $\mathrm{Var}(r)$ by the sample mean

$$\gamma_M = \frac{1}{2M} \sum_{i=1}^{M} (y'_i - y_i)^2 \tag{1.19}$$

where y'_i is the output value[6] associated with the input point $\boldsymbol{x}'_i$. By (1.18), an indication of the error involved in estimating $\mathrm{Var}(r)$ by γ_M is given by the *mean squared distance* between a point $\boldsymbol{x}_i$ and its nearest neighbour $\boldsymbol{x}'_i$. We represent this by

$$\delta_M = \frac{1}{M} \sum_{i=1}^{M} |\boldsymbol{x}'_i - \boldsymbol{x}_i|^2 \tag{1.20}$$

where $|.|$ denotes the Euclidean metric. Intuition then suggests that (1.13) is in some way equivalent to

$$\gamma_M \to \mathrm{Var}(r) \quad \text{as} \quad \delta_M \to 0 \tag{1.21}$$

Since the values of γ_M and δ_M depend on a particular *sample* of input–output data, the notion of convergence must now be weakened to that of *convergence in probability*. Furthermore, by the condition that the input space C contains no isolated points and since the sampling density ϕ is assumed to be strictly positive over C, the distance $|\boldsymbol{x}'_i - \boldsymbol{x}_i|$ between nearest neighbours in input space will converge to zero (in probability) as the number of points $M \to \infty$ so that

$$\delta_M \to 0 \quad \text{as} \quad M \to \infty \tag{1.22}$$

and hence

$$\gamma_M \to \mathrm{Var}(r) \quad \text{as} \quad M \to \infty \tag{1.23}$$

1.5 A more detailed description

Let $\boldsymbol{x}'$ denote the nearest neighbour of $\boldsymbol{x}$ among the input points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$. By the conditions that the input space C contains no isolated points and that the sampling density ϕ is assumed to be strictly positive over C, the distance $|\boldsymbol{x}' - \boldsymbol{x}|$ converges to zero in probability as the number of points $M \to \infty$. In particular

$$\mathcal{E}_\phi(|\boldsymbol{x}' - \boldsymbol{x}|) \to 0 \quad \text{as} \quad M \to \infty \tag{1.24}$$

[6]Note that y'_i is not necessarily the nearest neighbour of y_i in output space.

Replacing $f(\boldsymbol{x}') - f(\boldsymbol{x})$ in (1.9) by the first order approximation $(\boldsymbol{x}' - \boldsymbol{x}).\nabla f(\boldsymbol{x})$ we obtain

$$\frac{1}{2}(y' - y)^2 \approx \frac{1}{2}(r' - r)^2 + (r' - r)(\boldsymbol{x}' - \boldsymbol{x}).\nabla f(\boldsymbol{x}) + \frac{1}{2}\left((\boldsymbol{x}' - \boldsymbol{x}).\nabla f(\boldsymbol{x})\right)^2 \tag{1.25}$$

Let us now take the expectation of (1.25) with respect to both the noise distribution Ψ and the sampling distribution Φ. Since the noise variable r is assumed to be independent of the input $\boldsymbol{x}$ on which it is measured, and since $\boldsymbol{x}$ is independent of its associated noise value r, it may be argued that any expression involving only r and r' is constant relative to Φ and that any expression involving only $\boldsymbol{x}$ and $\boldsymbol{x}'$ is constant relative to Ψ. Thus we write

$$\begin{aligned}\mathcal{E}\left(\frac{1}{2}(y' - y)^2\right) \approx \mathcal{E}_\psi\left(\frac{1}{2}(r' - r)^2\right) + \mathcal{E}_\psi(r' - r)\mathcal{E}_\phi\left((\boldsymbol{x}' - \boldsymbol{x}).\nabla f(\boldsymbol{x})\right) \\ + \frac{1}{2}\mathcal{E}_\phi\left(\left((\boldsymbol{x}' - \boldsymbol{x}).\nabla f(\boldsymbol{x})\right)^2\right)\end{aligned} \tag{1.26}$$

By (1.12) the first expectation on the right hand side of (1.26) is equal to $\mathrm{Var}(r)$. By hypothesis, r and r' are independent and $\mathcal{E}(r) = 0$ so it follows that $\mathcal{E}_\psi(r' - r) = 0$. Thus, since $|(\boldsymbol{x}' - \boldsymbol{x}).\nabla f(\boldsymbol{x})| \leq c_1 b_1 < \infty$ where c_1 is the diameter of C and b_1 is the bound on the first partial derivative of f over C, we see that the second expectation on the right hand side of (1.26) is equal to zero and hence

$$\mathcal{E}\left(\frac{1}{2}(y' - y)^2\right) \approx \mathrm{Var}(r) + \frac{1}{2}\mathcal{E}_\phi\left(\left((\boldsymbol{x}' - \boldsymbol{x}).\nabla f(\boldsymbol{x})\right)^2\right) \tag{1.27}$$

Let

$$A(M) = \frac{\mathcal{E}_\phi\left(\left((\boldsymbol{x}' - \boldsymbol{x}) \cdot \nabla f(\boldsymbol{x})\right)^2\right)}{2\mathcal{E}_\phi(|\boldsymbol{x}' - \boldsymbol{x}|^2)} \tag{1.28}$$

so that (1.27) becomes

$$\mathcal{E}\left(\frac{1}{2}(y' - y)^2\right) \approx \mathrm{Var}(r) + A(M)\mathcal{E}_\phi\left(|\boldsymbol{x}' - \boldsymbol{x}|^2\right) \tag{1.29}$$

Since $|(\boldsymbol{x}' - \boldsymbol{x}).\nabla f(\boldsymbol{x})| \leq |\boldsymbol{x}' - \boldsymbol{x}||\nabla f(\boldsymbol{x})|$ and $|\nabla f(\boldsymbol{x})| \leq b_1$ it follows that $|A(M)| \leq \frac{1}{2}b_1^2 < \infty$ and hence

$$\mathcal{E}\left(\frac{1}{2}(y' - y)^2\right) \to \mathrm{Var}(r) \quad \text{as} \quad \mathcal{E}_\phi\left(|\boldsymbol{x}' - \boldsymbol{x}|^2\right) \to 0 \tag{1.30}$$

From our data set $\{(\boldsymbol{x}_i, y_i) : 1 \leq i \leq M\}$ we compute estimates for the expected values $\mathcal{E}\left(\frac{1}{2}(y' - y)^2\right)$ and $\mathcal{E}_\phi\left(|\boldsymbol{x}' - \boldsymbol{x}|^2\right)$ via the sample means

$$\gamma_M = \frac{1}{2M}\sum_{i=1}^{M} |y_i' - y_i|^2 \tag{1.31}$$

and

$$\delta_M = \frac{1}{M}\sum_{i=1}^{M} |\boldsymbol{x}'_i - \boldsymbol{x}_i|^2 \tag{1.32}$$

respectively.

If the sample mean of a set of identically distributed random variables converges to its expected value (as determined by the associated distribution function), the sample mean is said to satisfy the *law of large numbers*. It is relatively straightforward to prove this in the case of independent random variables. The difficulty in our case is that γ_M and δ_M are sample means of *dependent* random variables.

One step in establishing a rigorous mathematical foundation for the Gamma test will be to show that the law of large numbers holds for both γ_M and δ_M, and also to determine the error incurred by replacing the expected values in (1.30) with the sample means γ_M and δ_M.

Supposing for the moment that γ_M and δ_M do in fact satisfy the law of large numbers so that

$$\gamma_M \to \mathcal{E}\left(\frac{1}{2}(y'_i - y_i)^2\right) \qquad \text{and} \qquad \delta_M \to \mathcal{E}\left((\boldsymbol{x}'_i - \boldsymbol{x}_i)^2\right) \tag{1.33}$$

as $M \to \infty$, then (1.30) is equivalent to

$$\gamma_M \to \mathrm{Var}(r) \quad \text{as} \quad \delta_M \to 0 \tag{1.34}$$

and (1.24) becomes

$$\delta_M \to 0 \quad \text{as} \quad M \to \infty \tag{1.35}$$

so that

$$\gamma_M \to \mathrm{Var}(r) \quad \text{as} \quad M \to \infty \tag{1.36}$$

1.6 The crude method

Let us now perform a *Gedanken* experiment in which we plot successive pairs (δ_M, γ_M) as the number of points M increases in significant steps. By (1.34) and (1.36) the resulting curve should intersect the $\delta_M = 0$ axis at the noise variance $\mathrm{Var}(r)$, as illustrated in Figure 1.7. If we knew the parametric form of this curve we could extrapolate the result of our successive computations to determine the required intersection, providing us with an estimate Γ for $\mathrm{Var}(r)$.

The Gamma test is based on the assertion that this curve is approximately *linear* in probability near $\delta_M = 0$.

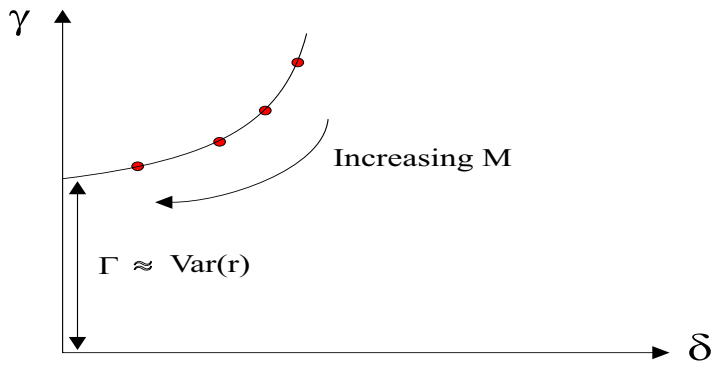

Figure 1.7: Gedanken Experiment

Thus we assert that there there exists some 'constant' $A(M)$, independent of the particular sample $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$, such that

$$\gamma_M \approx \mathrm{Var}(r) + A(M)\delta_M + o(\delta_M) \quad \text{as} \quad \delta_M \to 0 \tag{1.37}$$

or by (1.35),

$$\gamma_M \approx \mathrm{Var}(r) + A(M)\delta_M + o(\delta_M) \quad \text{as} \quad M \to \infty \tag{1.38}$$

where the convergence is in probability.

This assertion forms the basis of the *crude* Gamma test algorithm, in which we compute the pairs (δ_M, γ_M) as M increases in significant steps then perform linear regression on the resulting points. The constant term of the regression line is then returned as our estimate Γ for the variance of the noise, $\mathrm{Var}(r)$. We remark that the 'constant' $A(M)$ in (1.38) may depend on the number of points M, so if the regression technique is to be effective we require that $A(M)$ converges to some fixed value as the number of points M increases.

1.7 The refined method

We now present a refinement of the crude method, and indicate its advantages.

1.7.1 kth nearest neighbours

For any ordered set of points $T = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ the *kth nearest neighbour* of any point $\boldsymbol{x}_i \in T$ is uniquely defined as follows.

The first nearest neighbour of $\boldsymbol{x}_i$ is that point $\boldsymbol{x}_{j_1} \in T \setminus \{\boldsymbol{x}_i\}$ having minimal distance from $\boldsymbol{x}_i$ and minimal index j_1. The second nearest neighbour of $\boldsymbol{x}_i$ is that point $\boldsymbol{x}_{j_2} \in T \setminus \{\boldsymbol{x}_i, \boldsymbol{x}_{j_1}\}$ having minimal distance from $\boldsymbol{x}_i$ and minimal index j_2. In general,

the kth nearest neighbour of $\boldsymbol{x}_i$ is that point $\boldsymbol{x}_{j_k} \in T \setminus \{\boldsymbol{x}_i, \boldsymbol{x}_{j_1}, \ldots, \boldsymbol{x}_{j_{k-1}}\}$ having minimal distance from $\boldsymbol{x}_i$ and minimal index j_k.

Ordering equidistant points by their indices means that every point $\boldsymbol{x}_i$ has a uniquely defined kth nearest neighbour. More importantly, if $\boldsymbol{x}'_i$ is the kth nearest neighbour of $\boldsymbol{x}_i$ then there are *at most* k points $\boldsymbol{x}_j \in T$ with $|\boldsymbol{x}_j - \boldsymbol{x}_i| < |\boldsymbol{x}'_i - \boldsymbol{x}_i|$. This property will be required in (4.17) in order to construct a rigorous proof of the algorithm.

The kth nearest neighbour lists for a set of M points can be found in $O(M \log M)$ time using kd–trees [Bentley 1975] and as we shall see, this is the most computationally expensive aspect of the Gamma test algorithm.

1.7.2 The kth nearest neighbour modification

Let $\boldsymbol{x}_{N[i,k]}$ denote the kth nearest neighbour of the point $\boldsymbol{x}_i$ in the set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ and define the sample means

$$\gamma_M(k) = \frac{1}{2M} \sum_{i=1}^{M} |y_{N[i,k]} - y_i|^2 \tag{1.39}$$

and

$$\delta_M(k) = \frac{1}{M} \sum_{i=1}^{M} |\boldsymbol{x}_{N[i,k]} - \boldsymbol{x}_i|^2 \tag{1.40}$$

where $|.|$ denotes the Euclidean metric and $y_{N[i,k]}$ is the output value[7] associated with $\boldsymbol{x}_{N[i,k]}$.

The *refined* Gamma test algorithm (which we refer to simply as the Gamma test) is based on an extension of the asymptotic linearity assertion (1.38) to the kth nearest neighbour statistics $\gamma_M(k)$ and $\delta_M(k)$.

In this instance, we claim that there exists some 'constant' $A(M,k)$ (defined in (1.46)) which is independent of the particular sample $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ such that

$$\gamma_M(k) \approx \mathrm{Var}(r) + A(M,k)\delta_M(k) + o(\delta_M(k)) \quad \text{as} \quad M \to \infty \tag{1.41}$$

where the convergence is in probability.

On the basis of this assertion, we compute the pairs $(\delta_M(k), \gamma_M(k))$ for $1 \le k \le p$ (where p is typically taken in the range 10-50) and perform linear regression on these points as illustrated in Figure 1.8. The constant term Γ of this regression line is returned as our estimate for the variance of the noise, $\mathrm{Var}(r)$.

In order to establish a rigorous mathematical foundation for the Gamma test we must provide a proof of the asymptotic linearity relation (1.41), along with a formal justification of the associated linear regression technique.

[7] Note that $y_{N[i,k]}$ is not necessarily the kth nearest neighbour of y_i in the output space $\mathbb{R}$.

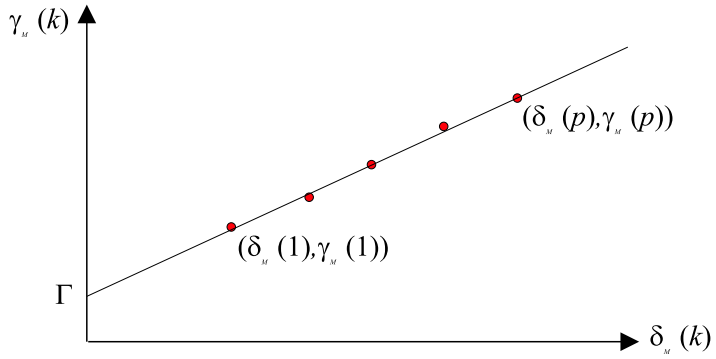

Figure 1.8: The Gamma test regression plot

1.7.3 The gradient of the Gamma test regression line

The gradient $A(M,k)$ of the regression line computed by the Gamma test may depend both on the number of points M and the near neighbour index k.

In the crude Gamma test algorithm the gradient $A = A(M)$ is required to converge to some fixed value as $M \to \infty$. However, in the refined Gamma test algorithm the question of whether or not $A(M,k)$ depends on M is not relevant, since the pairs $(\delta_M(k), \gamma_M(k))$ for $1 \leq k \leq p$ are each computed with respect to the same number of points.

In many cases $A(M,k)$ is also independent of k although simulation results (see section 7.8.3) suggest that this is *not* the case for some sets of fractional dimension. However, subject to a fairly weak condition on the asymptotic behaviour of $\delta_M(k)$ as M increases we can show that even if $A(M,k)$ depends on k, the intercept of the regression line computed by the Gamma test still converges in probability to $\mathrm{Var}(r)$ as $M \to \infty$.

1.7.4 Advantages of the refined method

In the crude Gamma test algorithm we compute the pairs (δ_M, γ_M) as the number of points M increases in significant steps. Since the nearest neighbour structure of the input points will change as M increases in this way, we will need to construct a separate kd–tree [Bentley 1975] for every pair (δ_M, γ_M), and this incurs a computational cost of order $O(M \log M)$ each time.

However, provided p is small relative to the number of points M a single kd–tree is sufficient to determine the kth nearest neighbours of the input points for all $1 \leq k \leq p$. Thus, in order to compute the pairs $(\delta_M(k), \gamma_M(k))$ the refined algorithm requires only a *single* computation of order $O(M \log M)$.

Furthermore, some of the sample means δ_M and γ_M computed by the crude algorithm are necessarily computed over a *small* number of points (relative to the total number

available), and will therefore involve a relatively large sampling error. In contrast, each of the sample means $\delta_M(k)$ and $\gamma_M(k)$ computed by the refined algorithm is computed over the *complete* set of M data points, and the sampling error can therefore be considered as the best possible in each case.

1.8 Statement of the main theorem

The following theorem, proved in Chapter 7, reiterates the conditions and definitions and is best regarded as a formal statement that provided the gradients $A(M,k)$ are independent of k, then with probability approaching one as $M \to \infty$ the relationship between the points $(\delta_M(k), \gamma_M(k))$ for $1 \leq k \leq p$ is approximately linear for M sufficiently large.

Theorem 1.1. *Let C be a closed bounded subset of $\mathbb{R}^m$ containing no isolated points. Let $f : \mathbb{R}^m \to \mathbb{R}$ be continuous and have bounded first and second partial derivatives on the convex hull $\mathcal{H}(C)$ of C. Let the points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$ in $C \subset \mathbb{R}^m$ and $r_1, \ldots, r_M$ in $\mathbb{R}$ be independently selected according to the probability distributions Φ and Ψ respectively, where the first four moments of Ψ are finite, and write*

$$y_i = f(\boldsymbol{x}_i) + r_i \qquad 1 \leq i \leq M \tag{1.42}$$

Let $\boldsymbol{x}_{N[i,k]}$ denote the kth nearest neighbour of $\boldsymbol{x}_i$ and define

$$\delta_M(k) = \frac{1}{M} \sum_{i=1}^{M} |\boldsymbol{x}_{N[i,k]} - \boldsymbol{x}_i|^2 \tag{1.43}$$

$$\gamma_M(k) = \frac{1}{2M} \sum_{i=1}^{M} |y_{N[i,k]} - y_i|^2 \tag{1.44}$$

Then for every $\kappa > 0$,

$$\gamma_M(k) = \mathrm{Var}(r) + A(M,k)\delta_M(k) + o(\delta_M(k)) + O\left(\frac{1}{M^{1/2-\kappa}}\right) \tag{1.45}$$

with probability greater than $1 - O(1/M^{2\kappa})$ as $M \to \infty$, where

$$A(M,k) = \frac{\mathcal{E}_\phi\left(\left((\boldsymbol{x}_{N[i,k]} - \boldsymbol{x}_i) \cdot \nabla f(\boldsymbol{x}_i)\right)^2\right)}{2\mathcal{E}_\phi(|\boldsymbol{x}_{N[i,k]} - \boldsymbol{x}_i|^2)} \tag{1.46}$$

and satisfies

$$0 \leq A(M,k) \leq \frac{1}{2} b_1^2 < \infty \tag{1.47}$$

where b_1 is the upper bound on the gradient of f over C.

Remark: In Chapter 3 we show that the expected value of $\delta_M(k)$ is of order $1/M^{2/m}$. Of the two error terms in (1.45) it thus follows that $O(1/M^{1/2-\kappa})$ dominates $o(\delta_M(k)$ for $m \leq 4$, while $o(\delta_M(k)$ dominates $O(1/M^{1/2-\kappa})$ for $m \geq 5$.

1.9 Statement of the algorithm.

Procedure *Gamma Test* (data)
{data is a set of points $\{(\boldsymbol{x}_i, y_i) \,|\, 1 \le i \le M\}$ where $\boldsymbol{x} \in \mathbb{R}^m$ and $y \in \mathbb{R}$}
for $i = 1$ to M **do**
  for $k = 1$ to p **do**
    compute the index $N[i,k]$ of the kth nearest neighbour of $\boldsymbol{x}_i$.
  end for
end for
{If multiple outputs do the remainder for each output}
for $k = 1$ to p **do**
  compute $\delta_M(k)$ as in (1.40)
  compute $\gamma_M(k)$ as in (1.39)
end for
Perform linear regression on $\{(\delta_M(k), \gamma_M(k)), 1 \le k \le p\}$ obtaining $\gamma = \Gamma + A\delta$
Return (Γ, A)

Algorithm 1: The Gamma test algorithm.

The Gamma test algorithm is given in Algorithm 1. We shall see that Theorem 1.1 allows us to infer that the Gamma statistic Γ returned by Algorithm 1 converges in probability to $\mathrm{Var}(r)$ as $M \to \infty$.

Theorem 1.2. *Subject to the condition that for some fixed $p \ge 1$ there exists a positive constant $c < 1$ such that*

$$\delta_M(1) \le c\delta_M(p) \tag{1.48}$$

for all sufficiently large M, the number Γ returned by Algorithm 1 converges in probability to $\mathrm{Var}(r)$ as $M \to \infty$.

Using the results of Chapter 3 we can then infer

Theorem 1.3. *Let C be a compact convex body in $\mathbb{R}^m$ and let Φ be a sampling distribution having smooth positive density over C. Then the number Γ returned by Algorithm 1 converges in probability to $\mathrm{Var}(r)$ as $M \to \infty$.*

In addition to Γ, Algorithm 1 also returns the gradient A of the regression line. As we shall see in section 7.7, this often provides an indicator of the complexity of the surface defined by the unknown function f.

1.10 Historical remark

The original version of the Gamma test described in [Stefánsson *et al.* 1997] used smoothed versions of the statistics $\gamma_M(k)$ and $\delta_M(k)$ defined by

$$\gamma^*_M(k) = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{2M}\sum_{i=1}^{M}(y_{N[i,j]} - y_i)^2 \tag{1.49}$$

and

$$\delta_M^*(k) = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{M}\sum_{i=1}^{M}|\boldsymbol{x}_{N[i,j]} - \boldsymbol{x}_i|^2 \tag{1.50}$$

The idea behind this was to decrease the significance of more distant near neighbours. Taking p large in such an implementation often does not significantly alter the resulting value of Γ. Later experience showed that provided p is kept small, the extra complication of computing $\gamma_M^*(k)$ and $\delta_M^*(k)$ is largely unnecessary (although this form of the Gamma test can sometimes produce better estimates when the number of data points M is small) and the later implementations of the technique are based on the definitions (1.39) and (1.40) presented here.

1.11 Summary

In this chapter we have given a brief introduction to the ideas behind the Gamma test and indicated why one should be interested in a noise estimation technique. We have also attempted to clarify the line of reasoning leading from the crude Gamma test to the more efficient form of Algorithm 1.

At the time the work of this thesis began there was already a substantial body of experimental results demonstrating the utility of the technique. Moreover, there had also been substantial software development to provide easy-to-use Windows (*winGamma*™) and Unix based software for the analysis of non-linear systems using the Gamma test. Thus there was a clear need to provide a formal mathematical basis for the method, and the goal of this thesis is to address this need.

1.12 Thesis outline

Following Chapter 1 the structure of this thesis is as follows.

In Chapter 2 we outline the general strategy motivating the proof of the Gamma test. We introduce some notation and show how the expression $\gamma_M(k)$ can be decomposed into three separate sums of dependent random variables. The chapter concludes by identifying the need to establish upper bounds on the variance of each of these sums. Figure 1.9 outlines the logical structure of the proof.

In Chapter 3 we give an account of near neighbour distance distributions. The results of this chapter represent extensive generalizations of the little that had previously appeared in the literature. In particular, using a novel *boundary shrinking* technique suggested to us by W. M. Schmidt, we are able to find asymptotic expressions for all moments of the kth nearest neighbour distance distribution on M points as $M \to \infty$,

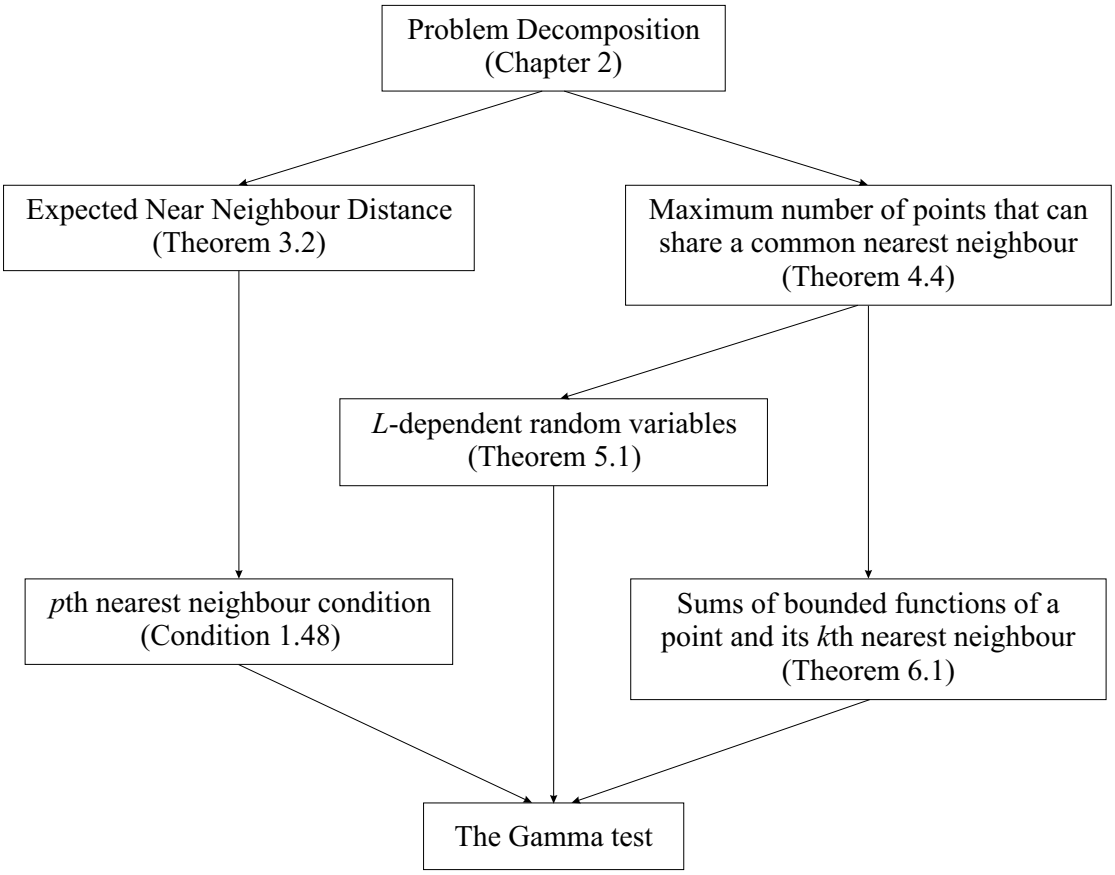

Figure 1.9: Logical structure of the proof.

where the points may be selected from any compact convex body $C \subset \mathbb{R}^m$ according to any sampling distribution having smooth positive density over C. We conjecture that these results may generalize to sets of zero Lebesgue measure but having positive Hausdorff dimension, and provide some preliminary experimental evidence in this direction. These investigations are prompted by the need to quantify the term $\delta_M(k)$ in as wide variety of circumstances as possible and the results will subsequently be used in Chapter 7.

In Chapter 4 we study kth near neighbour graphs. These are defined to be the directed graphs obtained by joining each point in a set of points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\} \subset \mathbb{R}^m$ to its kth nearest neighbour in the set. The main result of this chapter establishes an upper bound on the number of points in the set that can share a common kth nearest neighbour, the crucial fact being that the bound is *independent* of the total number of points M. We also digress to mention an interesting observation on first near neighbour graphs which follows from the investigations of Chapter 3.

It emerges that two quite separate techniques are needed to establish the required upper bounds on the variance of the three sums identified in Chapter 2, both of which depend on the main result of Chapter 4. In Chapter 5 we develop the theory of L-dependent variables to establish the required bounds for the first two sums. Although not required for a proof of the Gamma test algorithm we also prove a Central Limit

theorem for this class of random variables. In Chapter 6 we extend some results of [Bickel and Breiman 1983] regarding functions of a point and its kth nearest neighbour to obtain a bound on the variance of the third sum.

In Chapter 7 we assemble the results obtained in Chapter 4, 5 and 6 to construct a proof of Theorem 1.1. Following this we identify a simple condition that allows us to infer the validity of Algorithm 1 from Theorem 1.1, then use the results of Chapter 3 to show that this condition is satisfied by the class of sampling distributions having smooth positive density over a compact convex body in $\mathbb{R}^m$.

The Gamma test is a technique for estimating the second moment of the noise distribution Ψ, despite the fact that both the underlying function f and the noise distribution itself are unknown. In [Durrant 2001] an interesting and potentially useful generalisation of the Gamma test is described that enables one to estimate *all* higher moments of a *symmetric* noise distribution. This idea is based on a conjectured system of easily computed equations that constrain both odd and even moments. If Ψ happens to be symmetric then there are precisely as many equations as there are even moments and we can therefore solve for the moments. [Durrant 2001] provides experimental evidence supporting this system of equations and in Chapter 8, using techniques similar to those used in Chapter 7, we are able to provide a full proof.

The concluding Chapter 9 provides an overview of what has been accomplished and summarises the extent to which these results are new and original. We also indicate some unresolved questions which would be natural to pursue as a result of the present study.

Chapter 2

Decomposition of the problem and proof strategy

2.1 Introduction

In this chapter we give a representation of the problem in terms of *random samples*. Following this we provide a decomposition of the problem from which emerges three distinct sums of dependent random variables. As we shall see, we will need to show that each of these sums converges to zero in probability as the number of points $M \to \infty$.

The general line of attack to prove the convergence of each sum is based on Chebyshev's inequality. Using Fubini's theorem to separate the noise terms from the near neighbour distance terms, we show that the expectation of each sum is zero and hence that the main problem of proving convergence in probability becomes that of suitably bounding their variances.

2.2 Random samples

Let $C \subset \mathbb{R}^m$ and let $X = (X_1, \ldots X_M)$ be a vector of independent and identically distributed random variables X_i, each taking values $\boldsymbol{x}_i \in C$ according to the sampling density function ϕ. We think of X as a *random point sample* of size M taking values in the *sample space* C^M.

Associated with each random point sample X in C^M we define a *random noise sample* $R = (R_1, \ldots, R_M)$ of identically distributed random variables R_i, each taking values $r_i \in \mathbb{R}$ according to the sampling density function ψ. By hypothesis (**N.1**) we have that $\mathcal{E}_\psi(R_i) = 0$, $\mathcal{E}_\psi(R_i^2) < \infty$, $\mathcal{E}_\psi(R_i^3) < \infty$ and $\mathcal{E}_\psi(R_i^4) < \infty$.

Let $X_{N[i,k]}$ denote the kth nearest neighbour of X_i in the random point sample $X = (X_1, \ldots X_M)$. For every such point sample X we have a corresponding *indexing structure* $\mathcal{N}(X) = \{N[1,k], \ldots N[M,k]\}$, and the associated noise sample R inherits this indexing structure from the point sample X.

Any expression defined in terms of the noise sample $R = (R_1, \ldots, R_M)$ and the indexing structure $\mathcal{N}(X)$ might *not* be independent of the corresponding point sample X. On the other hand, by hypothesis **N.2** we see that any expression defined in terms of X and its kth nearest neighbour indexing structure $\mathcal{N}(X)$ is clearly independent of the associated noise sample R.

Finally, corresponding to each pair (X, R) in the product space $C^M \times \mathbb{R}^M$ we define a random sample $Y = (Y_1, \ldots, Y_M) \in \mathbb{R}^M$ where $Y_i = f(X_i) + R_i$ for each $1 \leq i \leq M$, and this also inherits the indexing structure $\mathcal{N}(X)$ from its associated point sample X.

Reformulating our problem in terms of random samples we obtain

$$\delta_M(k) = \frac{1}{M}\sum_{i=1}^{M} |X_{N[i,k]} - X_i|^2 \tag{2.1}$$

$$\gamma_M(k) = \frac{1}{2M}\sum_{i=1}^{M} (Y_{N[i,k]} - Y_i)^2 \tag{2.2}$$

and think of $\delta_M(k) : C^M \to \mathbb{R}$ and $\gamma_M(k) : C^M \times \mathbb{R}^M \to \mathbb{R}$ as *random variables* on the sample spaces C^M and $C^M \times \mathbb{R}^M$ respectively.

2.3 Decomposition

Consider

$$\frac{1}{2}(Y_{N[i,k]} - Y_i)^2 = \frac{1}{2}\big((R_{N[i,k]} - R_i) + (f(X_{N[i,k]}) - f(X_i))\big)^2 \tag{2.3}$$

By hypothesis (1.4) f has bounded first and second partial derivatives at each point in the convex hull $\mathcal{H}(C)$ of C so it follows that we can apply the second mean value theorem to the function f. Thinking of X_i as a column vector, this means that there exists some point $\Xi_i \in \mathcal{H}(C)$ on the line segment joining the points $X_{N[i,k]}$ and X_i such that

$$f(X_{N[i,k]}) - f(X_i) = (X_{N[i,k]} - X_i) \cdot \nabla f(X_i) + T_f(X_i, X_{N[i,k]}) \tag{2.4}$$

where

$$T_f(X_i, X_{N[i,k]}) = \frac{1}{2}(X_{N[i,k]} - X_i)^{\mathrm{T}} Hf(\Xi_i)(X_{N[i,k]} - X_i) \tag{2.5}$$

Note that the value of $Hf(\Xi_i)$ is uniquely determined by the points $X_{N[i,k]}$ and X_i, even though the intermediate point Ξ_i may not be. Substituting (2.4) back into (2.3) we get

$$\frac{1}{2}(Y_{N[i,k]} - Y_i)^2 = \frac{1}{2}\Big((R_{N[i,k]} - R_i) + (X_{N[i,k]} - X_i) \cdot \nabla f(X_i) + T_f(X_i, X_{N[i,k]})\Big)^2 \tag{2.6}$$

Expanding the square on the right hand side of (2.6) we obtain

$$\begin{aligned}\frac{1}{2}(Y_{N[i,k]} - Y_i)^2 = \frac{1}{2}\left(R_{N[i,k]} - R_i\right)^2 \\ + \left(R_{N[i,k]} - R_i\right)\Big((X_{N[i,k]} - X_i)\cdot\nabla f(X_i) + T_f(X_i, X_{N[i,k]})\Big) \\ + \frac{1}{2}\left((X_{N[i,k]} - X_i)\cdot\nabla f(X_i)\right)^2 \\ + \frac{1}{2}T_f(X_i, X_{N[i,k]})\Big(2(X_{N[i,k]} - X_i)\cdot\nabla f(X_i) + T_f(X_i, X_{N[i,k]})\Big)\end{aligned} \tag{2.7}$$

Summing both sides of (2.7) over $1 \le i \le M$ then dividing by M it follows that

$$\gamma_M(k) = \widetilde{A}_M(k,h) + B_M(k,h) + \widetilde{C}_M(k,h) + D_M(k) \tag{2.8}$$

where

$$\widetilde{A}_M(k) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{2}\left(R_{N[i,k]} - R_i\right)^2 \tag{2.9}$$

$$B_M(k) = \frac{1}{M}\sum_{i=1}^{M}\left(R_{N[i,k]} - R_i\right)\Big((X_{N[i,k]} - X_i)\cdot\nabla f(X_i) + T_f(X_i, X_{N[i,k]})\Big) \tag{2.10}$$

$$\widetilde{C}_M(k) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{2}\left((X_{N[i,k]} - X_i)\cdot\nabla f(X_i)\right)^2 \tag{2.11}$$

$$D_M(k) = \frac{1}{M}\sum_{i=1}^{M}T_f(X_i, X_{N[i,k]})\Big((X_{N[i,k]} - X_i)\cdot\nabla f(X_i) + \frac{1}{2}T_f(X_i, X_{N[i,k]})\Big) \tag{2.12}$$

First, by hypothesis (1.4) there exists some constant $b_1 < \infty$ such that

$$|(X_{N[i,k]} - X_i)\cdot\nabla f(X_i)| \le b_1|X_{N[i,k]} - X_i| \tag{2.13}$$

and some constant $b_2 < \infty$ with

$$|T_f(X_i, X_{N[i,k]})| \le b_2|X_{N[i,k]} - X_i|^2 \tag{2.14}$$

By definition of $\delta_M(k)$ it thus follows that $D_M(k) = o(\delta_M(k))$ as $M \to \infty$ and hence

$$\gamma_M(k) = \widetilde{A}_M(k) + B_M(k) + \widetilde{C}_M(k) + o(\delta_M(k)) \quad \text{as} \quad M \to \infty \tag{2.15}$$

Our aim is to prove Theorem 1.1 which states that for every $\kappa > 0$,

$$\gamma_M(k) = \mathrm{Var}(r) + A(M,k)\delta_M(k) + o(\delta_M(k)) + O\left(\frac{1}{M^{1/2-\kappa}}\right) \tag{2.16}$$

in probability as $M \to \infty$ where

$$A(M,k) = \frac{\mathcal{E}_\phi\left(\left((X_{N[i,k]} - X_i)\cdot\nabla f(X_i)\right)^2\right)}{2\mathcal{E}_\phi(|X_{N[i,k]} - X_i|^2)} \tag{2.17}$$

In view of this, we define

$$A_M(k) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{2}\left(R_{N[i,k]} - R_i\right)^2 - \mathrm{Var}(r) \tag{2.18}$$

$$C_M(k) = \frac{1}{2M}\sum_{i=1}^{M}\left(\left((X_{N[i,k]} - X_i)\cdot\nabla f(X_i)\right)^2 - A(M,k)|X_{N[i,k]} - X_i|^2\right) \tag{2.19}$$

so that $\widetilde{A}_M(k) = A_M(k) + \mathrm{Var}(r)$ and $\widetilde{C}_M(k) = C_M(k) + A(M,k)\delta_M(k)$. Substituting (2.18) and (2.19) into (2.15) we thus obtain

$$\gamma_M(k) = \mathrm{Var}(r) + A(M,k)\delta_M(k) + o(\delta_M(k)) + A_M(k) + B_M(k) + C_M(k) \tag{2.20}$$

and think of $A_M(k)$, $B_M(k)$ and $C_M(k)$ as *random variables* on the product space $C^M \times \mathbb{R}^M$.

2.4 Chebyshev's inequality

Let $\mathbf{P}(A)$ denote the probability that event A occurs. A sequence $Z_1, Z_2, \ldots$ of random variables is said to *converge in probability* to the random variable Z as $M \to \infty$ if for any $\epsilon > 0$,

$$\mathbf{P}(|Z_M - Z| > \epsilon) \to 0 \quad \text{as} \quad M \to \infty \tag{2.21}$$

This is equivalent to

$$\mathbf{P}(|Z_M - Z| \leq \epsilon) \to 1 \quad \text{as} \quad M \to \infty \tag{2.22}$$

so the probability that Z_M is within any $\epsilon > 0$ of Z approaches 1 as $M \to \infty$.

By (2.20), in order to prove Theorem 1.1 we need to show that each of the terms $A_M(k)$, $B_M(k)$ and $C_M(k)$ converges to zero in probability as $M \to \infty$. Our starting point is the following fundamental result of Chebyshev.

Lemma 2.1 (Chebyshev's Inequality). *For any random variable Z and any $\epsilon > 0$,*

$$\mathbf{P}\left(|Z - \mathcal{E}(Z)| > \epsilon\right) \leq \frac{\mathrm{Var}(Z)}{\epsilon^2} \tag{2.23}$$

Applying Chebyshev's inequality to each element of the sequence $Z_1, Z_2, \ldots$ of random variables we obtain

$$\mathbf{P}\left(|Z_M - \mathcal{E}(Z_M)| > \epsilon\right) \leq \frac{\mathrm{Var}(Z_M)}{\epsilon^2} \tag{2.24}$$

Hence, if $\mathrm{Var}(Z_M) = o(1)$ as $M \to \infty$ it follows by (2.21) that $Z_M \to \mathcal{E}(Z_M)$ in probability as $M \to \infty$, in which case Z_M is said to satisfy the *weak law of large numbers*. The smaller the upper bound on $\mathrm{Var}(Z_M)$, the faster the rate at which Z_M converges (in probability) to its expected value.

Thus, in order to prove Theorem 1.1 we need to

(1) show that $A_M(k)$, $B_M(k)$ and $C_M(k)$ each has expected value zero

(2) obtain the best possible asymptotic upper bounds on the variance of $A_M(k)$, $B_M(k)$ and $C_M(k)$ as $M \to \infty$.

2.5 Representation of $A_M(k)$, $B_M(k)$ and $C_M(k)$

Let $h : C \times C \to \mathbb{R}$ be any *bounded* function in the sense that

$$||h|| = \sup\{|h(x,y)| \, : \, x,y \in C\} < \infty \tag{2.25}$$

We define a set of random variables $(h_1(X), \ldots, h_M(X))$ on the set of random point samples C^M by

$$h_i(X) = h(X_i, X_{N[i,k]}) \qquad X \in C^M \tag{2.26}$$

Similarly, let $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be any function and define a set of random variables $(g_1(R), \ldots, g_M(R))$ on the set of random noise samples $\mathbb{R}^M$ by

$$g_i(R) = g(R_i, R_{N[i,k]}) \qquad R \in \mathbb{R}^M \tag{2.27}$$

where $N[i,k]$ is the index of the kth nearest neighbour of X_i in the associated point sample $X = (X_1, \ldots X_M)$. For suitable choices of the functions g and h, each of the terms $A_M(k)$, $B_M(k)$ and $C_M(k)$ is a random variable of the form $Z_M = Z_M(X,R)$ defined by

$$Z_M = \frac{1}{M} \sum_{i=1}^{M} g_i(R) h_i(X) \tag{2.28}$$

over the product space $C^M \times \mathbb{R}^M$.

2.6 The expected value of $A_M(k)$, $B_M(k)$ and $C_M(k)$

Lemma 2.2.

$$\mathcal{E}(Z_M) = \mathcal{E}(g_i(R)h_i(X)) \quad \textit{for all} \quad 1 \leq i \leq M \tag{2.29}$$

Proof. Consider the random variables $h_i(X)$ and $h_j(X)$. By hypothesis, the component variables X_i, X_j, $X_{N[i,k]}$ and $X_{N[j,k]}$ are identically distributed over C and since $i \neq N[i,k]$ and $j \neq N[j,k]$ it follows that $h_i(X)$ and $h_j(X)$ are identically distributed over C^M. Similarly it follows that $g_i(R)$ and $g_j(R)$ are identically distributed over $\mathbb{R}^M$ and hence $\mathcal{E}(g_i(R)h_i(X))$ is independent of the index i, as required. □

The expected value $A_M(k)$, $B_M(k)$ and $C_M(k)$ must be computed over every pair of random samples $(X, R) \in C^M \times \mathbb{R}^M$. Let P_ϕ denote the probability measure on C^M such that each component variable X_i is identically distributed in C with common density ϕ. Similarly, let P_ψ denote the probability measure on $\mathbb{R}^M$ such that each component variable R_i is identically distributed in $\mathbb{R}$ with common density ψ. Furthermore, let $\mathcal{E}_\phi$ denote an expectation taken with respect to P_ϕ over C^M, let $\mathcal{E}_\psi$ denote an expectation taken with respect to P_ψ over $\mathbb{R}^M$ and let $\mathcal{E}$ denote an expectation taken with respect to the product measure $P_\phi \otimes P_\psi$ over $C^M \times \mathbb{R}^M$.

Lemma 2.3.

$$\mathcal{E}(Z_M) = \mathcal{E}_\psi(g_i(R))\mathcal{E}_\phi(h_i(X)) \tag{2.30}$$

Proof. By hypothesis, the point sample X is independent of the noise sample R so by Lemma 2.2 and the Law of total probability (see [Billingsley 1979]) we have that

$$\mathcal{E}(Z_M) = \int_{X \in C^M} h_i(X)\mathcal{E}_\psi(g_i(R) \,\big|\, X)\, dP_\phi \tag{2.31}$$

where $\mathcal{E}_\psi(g_i(R) \,\big|\, X)$ is the conditional expected value of $g_i(R)$ over all noise samples $R \in \mathbb{R}^M$ under the dependence structure $\mathcal{N}(X)$ inherited from $X \in C^M$.

Clearly, since the component variables R_i are identically distributed in $\mathbb{R}$ and since $i \neq N[i,k]$, the expected value of $g_i(R) = g(R_i, R_{N[i,k]})$ is independent of any particular $X \in C^M$. Thus it follows that $\mathcal{E}_\psi(g_i(R) \,\big|\, X) = \mathcal{E}_\psi(g_i(R))$ and hence

$$\mathcal{E}(Z_M) = \mathcal{E}_\psi(g_i(R)) \int_{X \in C^M} h_i(X)\, dP_\phi = \mathcal{E}_\psi(g_i(R))\mathcal{E}_\phi(h_i(X)) \tag{2.32}$$

as required. □

Lemma 2.4.

$$\mathcal{E}(A_M(k)) = \mathcal{E}(B_M(k)) = \mathcal{E}(C_M(k)) = 0 \tag{2.33}$$

Proof. (a) Taking $g_i = \frac{1}{2}(R_{N[i,k]} - R_i)^2 - \mathrm{Var}(r)$ and $h_i = 1$ in Lemma 2.3 we obtain

$$\begin{aligned} \mathcal{E}(A_M(k)) &= \frac{1}{2}\mathcal{E}_\psi\left((R_{N[i,k]} - R_i)^2\right) - \mathrm{Var}(r) && (2.34)\\ &= \frac{1}{2}\left(\mathcal{E}_\psi(R^2_{N[i,k]}) + \mathcal{E}_\psi(R_i^2) - 2\mathcal{E}_\psi(R_{N[i,k]}R_i)\right) - \mathrm{Var}(r) && (2.35)\end{aligned}$$

Since $i \neq N[i,k]$, by hypothesis we have that R_i and $R_{N[i,k]}$ are independent so

$$\mathcal{E}_\psi(R_{N[i,k]}R_i) = \mathcal{E}_\psi(R_{N[i,k]})\mathcal{E}_\psi(R_i) = 0 \tag{2.36}$$

and since $\mathcal{E}_\psi(R^2_{N[i,k]}) = \mathcal{E}_\psi(R_i^2) = \mathrm{Var}(r)$ it follows that $\mathcal{E}(A_M(k)) = 0$.

(b) Taking $g_i = R_{N[i,k]} - R_i$ and $h_i = (X_{N[i,k]} - X_i) \cdot \nabla f(X_i) + T_f(X_i, X_{N[i,k]})$ in Lemma 2.3 we obtain

$$\mathcal{E}(B_M(k)) = \mathcal{E}_\psi(R_{N[i,k]} - R_i)\mathcal{E}_\phi\Big((X_{N[i,k]} - X_i) \cdot \nabla f(X_i) + T_f(X_i, X_{N[i,k]})\Big) \tag{2.37}$$

Clearly,

$$\begin{aligned}
&\Big|\mathcal{E}_\phi\Big((X_{N[i,k]} - X_i)\cdot\nabla f(X_i) + T_f(X_i, X_{N[i,k]})\Big)\Big| \\
&\quad\leq\ \mathcal{E}_\phi\Big(\big|(X_{N[i,k]} - X_i)\cdot\nabla f(X_i) + T_f(X_i, X_{N[i,k]})\big|\Big) \qquad (2.38)\\
&\quad\leq\ \mathcal{E}_\phi\Big(\big|X_{N[i,k]} - X_i\big|\big|\nabla f(X_i)\big| + \big|T_f(X_i, X_{N[i,k]})\big|\Big) \qquad (2.39)
\end{aligned}$$

By hypothesis, $|X_{N[i,k]} - X_i| \leq c_1$, $|\nabla f(X_i)| \leq b_1$ and $|Hf(X_j)| \leq b_2$ for all $X_i, X_j \in C$ so by (2.5) we have that

$$\big|\mathcal{E}_\phi\Big((X_{N[i,k]} - X_i)\cdot\nabla f(X_i) + T_f(X_i, X_{N[i,k]})\Big)\big| \leq c_1 b_1 + c_1^2 b_2 < \infty \qquad (2.40)$$

Hence the second expectation in (2.37) is finite and since $\mathcal{E}_\psi(R_{N[i,k]}) = \mathcal{E}_\psi(R_i) = 0$, the first expectation in (2.37) is identically zero and thus it follows that $\mathcal{E}(B_M(k)) = 0$.

(c) Taking $g_i = 1$ and $h_i = \frac{1}{2}((X_{N[i,k]} - X_i)\cdot\nabla f(X_i))^2 - A(M,k)|X_{N[i,k]} - X_i|^2$ in Lemma 2.3 we obtain

$$\mathcal{E}(C_M(k)) = \frac{1}{2}\mathcal{E}_\phi\left(((X_{N[i,k]} - X_i)\cdot\nabla f(X_i))^2\right) - A(M,k)\mathcal{E}_\phi\left(|X_{N[i,k]} - X_i|^2\right) \qquad (2.41)$$

so by definition of $A(M,k)$ it follows that $\mathcal{E}(C_M(k)) = 0$. □

2.7 The variance of $A_M(k)$, $B_M(k)$ and $C_M(k)$

In order to apply Chebyshev's inequality we need to establish the smallest possible upper bounds on the variance of $A_M(k)$, $B_M(k)$ and $C_M(k)$ in terms of the sample size M. By Lemma 2.4, this is equivalent to finding upper bounds on their second moments $\mathcal{E}(A_M(k)^2)$, $\mathcal{E}(B_M(k)^2)$ and $\mathcal{E}(C_M(k)^2)$.

Lemma 2.5.

$$\mathcal{E}(Z_M^2) = \frac{1}{M^2}\sum_{i,j=1}^{M}\mathcal{E}_\phi\Big(h_i(X)h_j(X)\mathcal{E}_\psi(g_i(R)g_j(R)\,\big|\,X)\Big) \qquad (2.42)$$

where $\mathcal{E}_\psi(g_i(R)g_j(R)\,\big|\,X)$ is the conditional expected value of $g_i(R)g_j(R)$ over all noise samples $R \in \mathbb{R}^M$ subject to the dependence structure inherited from $X \in C^M$.

Proof. We write

$$Z_M^2 = \frac{1}{M^2}\sum_{i,j=1}^{M} g_i(R)g_j(R)h_i(X)h_j(X) \qquad (2.43)$$

and take the expectation of Z_M^2 over all pairs $(X,R) \in C^M \times \mathbb{R}^M$ so that

$$\mathcal{E}(Z_M^2) = \frac{1}{M^2}\sum_{i,j=1}^{M}\mathcal{E}\big(g_i(R)g_j(R)h_i(X)h_j(X)\big) \qquad (2.44)$$

Since X and R are distributed according to the probability measures P_ϕ and P_ψ respectively, by Fubini's theorem we have that

$$\mathcal{E}(g_i(R)g_j(R)h_i(X)h_j(X)) = \int_{X\in C^M} \left(\int_{R\in\mathbb{R}^M} g_i(R)g_j(R)h_i(X)h_j(X)\, dP_\psi \right) dP_\phi \tag{2.45}$$

Furthermore, by hypothesis the point sample X is completely independent of the noise sample R so

$$\mathcal{E}(g_i(R)g_j(R)h_i(X)h_j(X)) = \int_{X\in C^M} h_i(X)h_j(X) \left(\int_{R\in\mathbb{R}^M} g_i(R)g_j(R)\, dP_\psi \right) dP_\phi \tag{2.46}$$

which we write as

$$\mathcal{E}(g_i(R)g_j(R)h_i(X)h_j(X)) = \int_{X\in C^M} h_i(X)h_j(X)\mathcal{E}_\psi(g_i(R)g_j(R) \,\big|\, X)\, dP_\phi \tag{2.47}$$

Hence

$$\mathcal{E}(g_i(R)g_j(R)h_i(X)h_j(X)) = \mathcal{E}_\phi(h_i(X)h_j(X)\mathcal{E}_\psi(g_i(R)g_j(R) \,\big|\, X)) \tag{2.48}$$

and the result follows by (2.44). □

Since each $X \in C^M$ imposes a particular dependence structure on the noise samples $R \in \mathbb{R}^M$, the conditional expectation $\mathcal{E}_\psi(g_i(R)g_j(R) \,\big|\, X)$ is likely to depend on X. However, since we only need an *upper bound* on $\mathcal{E}(Z_M^2)$ (i.e. on the variance of the terms $A_M(k)$, $B_M(k)$ and $C_M(k)$) it will be sufficient to establish an upper bound on the conditional expectation $\mathcal{E}_\psi(g_i(R)g_j(R) \,\big|\, X))$ which is *independent* of any particular $X \in C^M$.

2.8 Summary

We have shown how the term $\gamma_M(k)$ can be decomposed into three rather different sums of dependent random variables and seen how the problem of estimating their probabilistic asymptotic behaviour as functions of the sample size M may be reduced to obtaining upper bounds on their variance. We shall address the issues of obtaining such upper bounds in Chapter 5 and Chapter 6.

In the next chapter we turn to the problem of computing the expected value of the mean squared distance $\delta_M(k)$ between kth nearest neighbours in a set of M points.

Chapter 3

Moments of nearest neighbour distance distributions

3.1 Introduction

The Gamma test is based on the mean squared distance $\delta_M(k)$ between kth nearest neighbours in a set of M points selected at random from the set $C \subset \mathbb{R}^m$ according to some sampling distribution function Φ. In Chapter 6 we show that the sample mean $\delta_M(k)$ converges in probability to its expected value as the number of points M increases without bound. The aim of the present chapter is to compute this expected value under the weakest possible restrictions on the sampling distribution Φ and the associated set C. Our method of dealing with boundary effects will subsequently be used in Chapter 4 to prove a result on first nearest neighbour graphs.

Let $X = (X_1, \ldots, X_M)$ be a random sample of independent and identically distributed random variables where each X_i takes a value *uniformly* at random from the *unit hypercube* in $\mathbb{R}^m$. Let $d_{M,k} = d_{M,k}(X)$ denote the distance between X_i and its kth nearest neighbour$X_{N[i,k]}$ in the random sample X. In [Percus and Martin 1998] it is shown that *under periodic boundary conditions*[1], the expected value of $d_{M,k}$ over all such random samples $X \in C^M$ satisfies the asymptotic expression

$$\mathcal{E}(d_{M,k}) = V_m^{-1/m} \frac{\Gamma(k + 1/m)}{\Gamma(k)} \frac{1}{M^{1/m}} + O\left(\frac{1}{M^{1+1/m}}\right) \quad \text{as} \quad M \to \infty \tag{3.1}$$

where Γ is the Euler gamma function and V_m is the Lebesgue measure of the unit ball in $\mathbb{R}^m$.

In this chapter we prove a similar result for the αth moment $\mathcal{E}(d_{M,k}^{\alpha})$ of the kth nearest neighbour distance distribution, with the following important generalisations:

- The sample points X_i may be selected from any compact convex body C in $\mathbb{R}^m$, relaxing the restriction regarding periodic boundary conditions.

[1] Sets having 'periodic boundary conditions' are sets that have no boundary at all, e.g. the surface of a torus or sphere.

- The sample points X_i may be selected according to any sampling distribution whose density is smooth and strictly positive over C, relaxing the uniformity assumption.

The first of these is achieved using a novel method of dealing with boundary effects, suggested by Professor W. M. Schmidt of the Department of Mathematics, Colorado University, USA (private communication to A. J. Jones).

Under these conditions, Theorem 3.2 states that for all $\rho > 0$,

$$\mathcal{E}(d^{\alpha}_{M,k}) = \frac{c}{M^{\alpha/m}} + O\left(\frac{1}{M^{(\alpha+1)/m-\rho}}\right) \quad \text{as} \quad M \to \infty \tag{3.2}$$

where $c = c(m, \alpha, k, \phi)$ is independent of M and given by

$$c = V_m^{-\alpha/m} \frac{\Gamma(k + \alpha/m)}{\Gamma(k)} \int_C \phi(x)^{1-\alpha/m}\, dx \tag{3.3}$$

A related result appears as Theorem 8.3 of [Yukich 1998], which gives an asymptotic expression (in terms of the number of points) for the length of the k-nearest neighbours graph[2] of a set of points selected from the unit cube $[0,1]^m$ ($m \geq 2$) according to a well behaved sampling distribution. A scaling argument is then employed to extend the result to arbitrary compact subsets of $\mathbb{R}^m$. This will be discussed further in section 6.8.

For the proof of the Gamma test given in Chapter 7 the result specifically needed from this chapter is the expected asymptotic behaviour of the second moment ($\alpha = 2$) in Theorem 3.2. However, the Gamma test appears to work very well for a far larger class of sampling distributions than those whose densities are smooth and strictly positive over C. In particular it has been applied to chaotic dynamical systems where the sampling is an ergodic process over some set C which, far from being a compact convex body in $\mathbb{R}^m$, is an attractor of fractional dimension. In section 3.9 we examine some of the issues involved in estimating the asymptotic size of the mean squared kth nearest neighbour distance $\delta_M(k)$ where the sampling is over a chaotic attractor. This appears to raise very difficult questions and although Proposition 3.2 provides a weak result in this direction, we have to be content with Conjecture 3.2.

Nevertheless, in section 3.10 we provide an experimental comparison between Theorem 3.6 for a uniform sampling distribution and Conjecture 3.2 for case where the points lie on the chaotic attractor in $\mathbb{R}^2$ generated by the Hénon map. The case of the Hénon map provides some modest support for Conjecture 3.2.

The following notation will be employed throughout this chapter.

- μ denotes Lebesgue measure in $\mathbb{R}^m$.

[2]The k-nearest neighbours graph of a set of points is constructed by inserting an edge between each point and its k nearest neighbours in the set

- ∂C denotes the boundary of the set $C \subset \mathbb{R}^m$.
- $B_x(r)$ denotes the closed ball of radius r centred at $x \in \mathbb{R}^m$.
- V_m denotes the (Lebesgue) volume of the unit ball $B_0(1)$ in $\mathbb{R}^m$.
- For any $C \subset \mathbb{R}^m$ and $\delta > 0$, the set of points of C that are within a distance δ of its boundary is called the *boundary region of width* δ, and denoted by

$$C(\delta) = \{x \in C : \inf_{y \in \partial C} |x - y| < \delta\} \tag{3.4}$$

3.2 The sampling distribution Φ.

If the sampling distribution function Φ has a well defined density function ϕ, the associated *probability measure* on subsets $A \subseteq \mathbb{R}^m$ is given by

$$\mathbf{P}(A) = \int_A \phi(x)\, dx \tag{3.5}$$

which corresponds to the probability that a point is selected from the set $A \subset \mathbb{R}^m$. Let C denote the set of points $x \in \mathbb{R}^m$ for which the sampling density $\phi(x)$ is strictly positive,

$$C = \{x \in \mathbb{R}^m : \phi(x) > 0\} \tag{3.6}$$

We restrict our attention to those sampling distributions Φ satisfying the following conditions:

P.1 C is closed and bounded

P.2 ϕ is continuous on C.

P.3 ϕ has bounded partial derivatives at each point of C.

Note that by (3.6), condition **P.1** means that C is equivalent to the *support* of ϕ. In the following lemma we show that conditions **P.1** and **P.2** ensure that ϕ is uniformly bounded away from zero for all $x \in C$. Since C is closed, this means that ϕ cannot be continuously extended beyond C. Thus conditions **P.1** and **P.2** eliminate sampling distributions whose densities vanish at the boundary of C.

Lemma 3.1. *If Φ satisfies conditions **P.1** and **P.2** then there exist constants a_1, a_2 such that for all $x \in C$,*

$$0 < a_1 < \phi(x) < a_2 < \infty \tag{3.7}$$

Proof. Since ϕ is continuous and C is compact there exist $x_1, x_2 \in C$ such that $x_1 = \inf\{\phi(x) : x \in C\}$ and $x_2 = \sup\{\phi(x) : x \in C\}$. Taking $a_1 = \phi(x_1)$ and $a_2 = \phi(x_2)$ completes the proof. □

3.2.1 The set C

To deal with boundary effects we need that C satisfies the following geometric conditions.

C.1 There exists some finite constant $c_1 > 0$ such that

$$\max\{|x - y| : x, y \in C\} = c_1 \tag{3.8}$$

so C has diameter c_1.

C.2 There exists some constant $c_2 > 0$ such that for all $x \in C$ and $0 < r < c_1$,

$$\mu(B_x(r) \cap C) > c_2 r^m \tag{3.9}$$

so for every $r > 0$, at least a uniformly constant proportion c_2/V_m of the ball $B_x(r)$ intersects C.

C.3 There exist constants $c_3 > 0$ and $\lambda = \lambda(C) > 0$ such that for all $0 < \delta < \lambda$,

$$\mu(C(\delta)) \leq c_3 \delta \tag{3.10}$$

so for sufficiently small $\delta > 0$, the measure of the boundary region $C(\delta)$ is uniformly bounded by some constant multiple of its width δ.

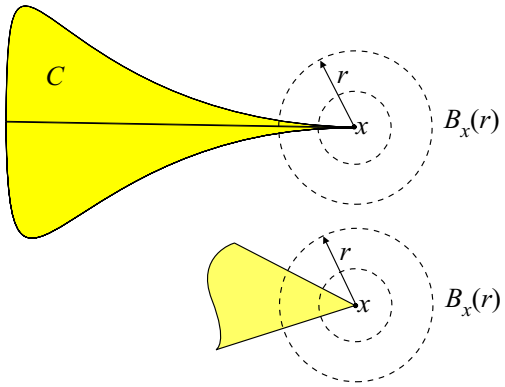

Figure 3.1: Condition **C.2** eliminates certain types of boundary points (top).

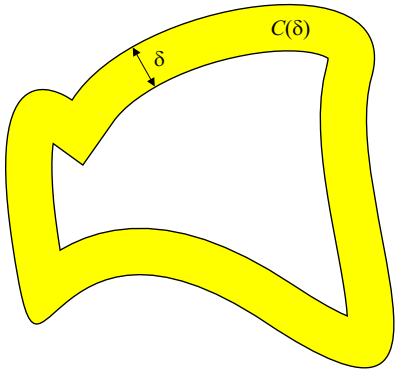

Figure 3.2: Condition **C.3** bounds the measure of $C(\delta)$ in terms of its width.

3.2.2 A probability measure on neighbourhood balls

Let $\omega_x(r)$ denote the probability measure induced by the sampling distribution Φ on the neighbourhood balls $B_x(r)$ of C. This is defined by

$$\omega_x(r) = \int_{B_x(r) \cap C} \phi(t)\, dt \tag{3.11}$$

and corresponds to the probability that a point selected from C according to Φ is contained in the ball $B_x(r)$. Since the sampling density ϕ is strictly positive on C we have that $\omega_x(r)$ is a monotonic increasing function of r. In the following lemma we show that $\omega_x(r)$ satisfies a *positive density* condition on C.

Lemma 3.2. *Subject to conditions **P.1**, **P.2** and **C.2**, there exists some constant $c_4 > 0$ such that*

$$\omega_x(r) \geq c_4 r^m \quad \textit{for all} \quad x \in C \quad \textit{and} \quad 0 \leq r \leq c_1 \tag{3.12}$$

Proof. Since ϕ satisfies **P.1** and **P.2**, by Lemma 3.1 there exists some constant $a_1 > 0$ such that $\phi(x) > a_1 > 0$ for all $x \in C$. By (3.11) it thus follows that $\omega_x(r) \geq a_1 \mu(B_x(r) \cap C)$ and by condition **C.2** we have that $\omega_x(r) \geq a_1 c_2 r^m$ where $c_2 > 0$ is constant. The result follows on taking $c_4 = a_1 c_2 > 0$. □

3.2.3 Compact convex bodies in $\mathbb{R}^m$

We now show that the conditions **C.1**, **C.2** and **C.3** are satisfied by compact convex bodies in $\mathbb{R}^m$.

Proposition 3.1. *Conditions **C.1**, **C.2** and **C.3** are satisfied by compact convex bodies in $\mathbb{R}^m$ having non-empty interior.*

Proof. Let C be a compact convex body in $\mathbb{R}^m$. Condition **C.1** follows easily by compactness.

To prove condition **C.2** we first note that a convex body has non–empty interior by definition, so there exist points $a, b \in C$ and some $r_a, r_b > 0$ such that the balls $B_a(r_a)$ and $B_b(r_b)$ are disjoint and completely contained in C.

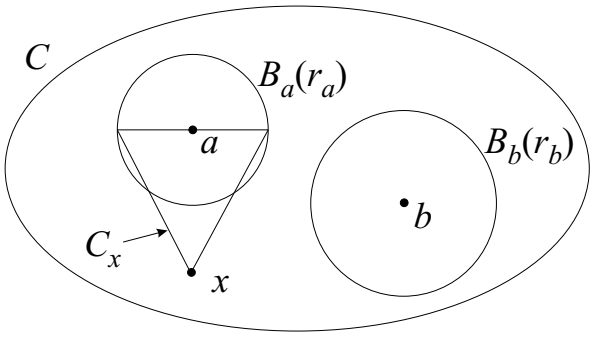

Figure 3.3: The cone C_x is completely contained in C.

First suppose that $x \in C \setminus B_a(r_a)$ and consider the cone C_x having vertex x and base equal to the intersection of $B_a(r_a)$ with the hyperplane through a perpendicular to the line joining x and a (see Figure 3.3). By convexity, C_x is completely contained in C and furthermore, since $x \notin B_a(r_a)$ and $r_a > 0$ it follows that C_x is of positive measure for each $x \in C \setminus B_a(r_a)$.

Let $c_x(r) > 0$ denote the proportion of the ball $B_x(r)$ occupied by the cone C_x, given by

$$c_x(r) = \frac{\mu(C_x \cap B_x(r))}{\mu(B_x(r))} \tag{3.13}$$

As r increases from 0 to c_1, the proportion $c_x(r)$ remains constant while $r \leq |x-a|$ and then decreases monotonically for $r > |x-a|$. Let $c_x = c_x(c_1)$ denote the minimum value of $c_x(r)$. Since $\mu(C_x) > 0$ and $\mu(B_x(c_1)) < \infty$ it follows that $c_x > 0$ and also that $\mu(B_x(r) \cap C) \geq c_x \mu(B_x(r)) > 0$ for all $0 < r \leq c_1$.

Define c_a to be the minimum value of c_x over all points $x \in C \backslash B_a(r_a)$. This corresponds to those points x that lie on the boundary of the ball $B_a(r_a)$, for which the cones C_x are of minimum volume. Since $B_a(r_a)$ is of positive radius we have that $c_a > 0$ and hence $\mu(B_x(r) \cap C) \geq c_a \mu(B_x(r)) > 0$ for all $0 < r \leq c_1$ and $x \in C \setminus B_a(r_a)$.

Now suppose that $x \in B_a(r_a)$ and define C_x and c_x as above but this time relative to the ball $B_b(r_b)$. Define c_b to be the minimum value of c_x over each $x \in B_a(r_a)$. Since $B_a(r_a)$ and $B_b(r_b)$ are disjoint and since $B_b(r_b)$ is of positive radius we have that $c_b > 0$. Hence $\mu(B_x(r) \cap C) \geq c_b \mu(B_x(r)) > 0$ for all $0 < r \leq c_1$ and $x \in B_a(r_a)$. Thus, letting $c = \min\{c_a, c_b\} > 0$ it follows that for all $0 < r \leq c_1$ and for every $x \in C$,

$$\mu(B_x(r) \cap C) \geq c\mu(B_x(r)) \geq cV_m r^m > 0 \tag{3.14}$$

and hence condition **C.2** is satisfied with $c_2 = cV_m > 0$.

Finally we show that C satisfies condition **C.3**. Since C is a convex body then by definition it has non–empty interior. Suppose without loss of generality that the origin $0 \in \operatorname{int} C$. Let $B_0(\lambda)$ denote the ball of maximal radius $\lambda > 0$ centred at the origin and contained in C. We claim that for every $0 < \delta < \lambda$ the sets $C(\delta)$ and $(1 - \delta/\lambda)C$ are disjoint.

Let $0 < \delta < \lambda$ and suppose that $x \in C(\delta)$. Let y be a boundary point of C such that the distance from x to y is strictly less than δ. Such a point exists by definition of $C(\delta)$.

Consider the vector z defined by $x = y + z$. Then $|z| < \delta$ and we write $-z = (\delta/\lambda)a$ for some $a \in \mathbb{R}^m$. From here we see that $a = (\lambda/\delta)(-z)$ and hence $|a| = (\lambda/\delta)|z|$. Furthermore, $|z| < \delta$ implies that $|a| < \lambda$ and hence $a \in \operatorname{int} B_0(\lambda)$. Since $B_0(\lambda) \subseteq C$, this means that $a \in \operatorname{int} C$ and hence $-z \in \operatorname{int} (\delta/\lambda)C$ as illustrated in Figure 3.4.

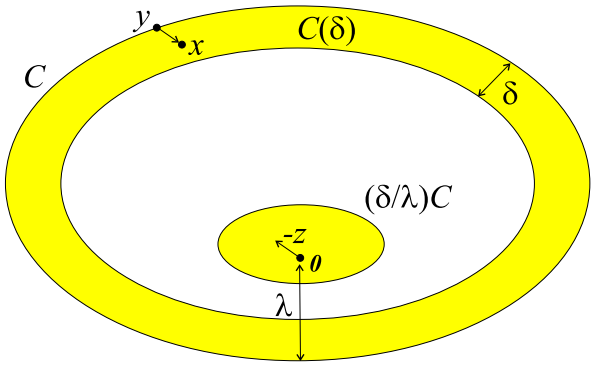

Figure 3.4: The vector $-z = y - x$ is contained in $(\delta/\lambda)C$.

Now suppose that $x \in (1 - \delta/\lambda)C$. Then $y = x - z$ can be expressed as $y = (1 - \delta/\lambda)b + (\delta/\lambda)a$ for some $b \in C$ and $a \in \operatorname{int} C$. Since $0 < \delta < \lambda$ we have $0 < \delta/\lambda < 1$

and $0 < 1 - \delta/\lambda < 1$ and since C is convex and $a \in \operatorname{int} C$ this implies that $y \in \operatorname{int} C$, contradicting the fact that y is a boundary point of C.

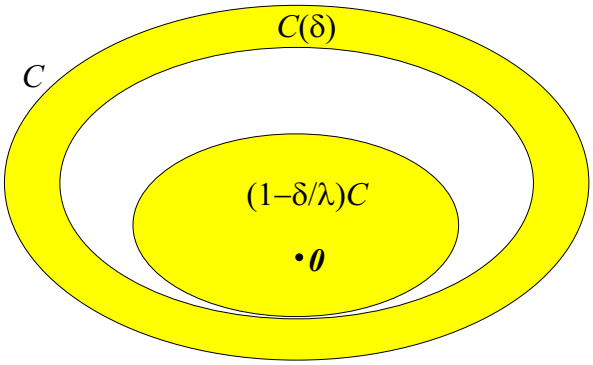

Figure 3.5: The sets $C(\delta)$ and $(1 - \delta/\lambda)C$ are disjoint.

Hence $x \notin (1 - \delta/\lambda)C$ so the sets $C(\delta)$ and $(1 - \delta/\lambda)C$ are disjoint. $C(\delta)$ is therefore contained in $C \setminus (1 - \delta/\lambda)C$ so

$$\mu\left(C(\delta)\right) \leq \mu(C) - \mu\left((1 - \delta/\lambda)C\right) \tag{3.15}$$

Writing $\mu((1 - \delta/\lambda)C) = (1 - \delta/\lambda)^m \mu(C)$ it follows that

$$\mu(C(\delta)) \leq \mu(C)\left(1 - (1 - \delta/\lambda)^m\right) \tag{3.16}$$

which we write as

$$\mu(C(\delta)) \leq \delta\mu(C)\left(\frac{1}{\lambda} - \frac{m\delta}{\lambda^2} + \ldots + (-1)^m \frac{\delta^{m-1}}{\lambda^m}\right) \tag{3.17}$$

By compactness we know that $\mu(C) < \infty$. Hence, for all $0 < \delta < \lambda$ we see that $\mu(C(\delta)) \leq c_3\delta$ for some constant $c_3 > 0$, as required. □

3.3 An integral representation of the moments

Let $C \subset \mathbb{R}^m$ satisfy conditions **C.1**, **C.2** and **C.3** and let $X = (X_1, \ldots X_M)$ be a random sample of independent and identically distributed random variables X_i, each taking values in C according to the probability distribution Φ. Let $X_{N[i,k]}$ denote the kth nearest neighbour of X_i in $X \in C^M$ and consider the function $d_{M,k} : C^M \to [0, c_1]$ defined by

$$d_{M,k}(X) = |X_{N[i,k]} - X_i| \tag{3.18}$$

so that $d_{M,k}(X)$ is the distance between X_i and its kth nearest neighbour in $X \in C^M$. We think of $d_{M,k} = d_{M,k}(X)$ as *random variable* on the sample space C^M and its αth moment is given by

$$\mathcal{E}(d^\alpha_{M,k}) = \int_{X \in C^M} d^\alpha_{M,k}(X)\, dP \tag{3.19}$$

where P is the probability measure on C^M for which each X_i is identically distributed in C with distribution Φ. Note that since the component variables X_i are *independent*, this expectation is independent of the index i.

For each $x \in C$ let

$$C^M(x) = \{X \in C^M \,:\, X_i = x\} \tag{3.20}$$

be the set of samples $X \in C^M$ for which the component variable X_i takes the fixed value $x \in C$ and define the random variable $d_{M,k}(x)$ on $C^M(x)$ by

$$\begin{array}{rccl} d_{M,k}(x): & C^M(x) & \to & [0, c_1] \\ & X & \mapsto & |X_{N[i,k]} - X_i| \end{array} \tag{3.21}$$

By Fubini's theorem

$$\mathcal{E}(d^\alpha_{M,k}) = \int_{x \in C} \mathcal{E}(d^\alpha_{M,k}(x))\phi(x)\, dx \tag{3.22}$$

where

$$\mathcal{E}(d^\alpha_{M,k}(x)) = \mathcal{E}(d^\alpha_{M,k}(X) \,\big|\, X_i = x) \tag{3.23}$$

is the conditional expected value of $|X_{N[i,k]} - X_i|^\alpha$ given that X_i is fixed at x.

3.3.1 A radial density function

Let $q_x(r) = q_x(r, M, k)$ be the distribution function of $d_{M,k}(x)$ over $C^M(x)$, defined by

$$q_x(r) = \mathbf{P}(d_{M,k}(x) \leq r) \tag{3.24}$$

This corresponds to the probability that the distance $d_{M,k}(x)$ from $X_i = x$ to its kth nearest neighbour in the random sample $X \in C^M(x)$ is *at most* equal to r. The conditional expected value of $d^\alpha_{M,k}$ given that $X \in C^M(x)$ is then given by

$$\mathcal{E}(d^\alpha_{M,k}(x)) = \int_0^{c_1} r^\alpha \, dq_x(r) \tag{3.25}$$

where the integral is the Stieltjes integral of r^α with respect to $q_x(r)$ as r ranges over $[0, c_1]$. Following [Percus and Martin 1996], we obtain the following expression for the corresponding density function $dq_x(r)$.

Lemma 3.3. *For fixed $x \in C$ the probability density function of the random variable $d_{M,k}(x)$ is given by*

$$dq_x(r) = k\binom{M-1}{k}\omega_x(r)^{k-1}(1 - \omega_x(r))^{M-k-1}\, d\omega_x(r) \tag{3.26}$$

where $\omega_x(r)$ is the probability measure of the ball $B_x(r)$ centred at x having radius r.

Proof. Let $\epsilon > 0$ and consider the probability that the kth nearest neighbour of $X_i = x$ lies in the spherical shell of radius r and width $\epsilon > 0$ centred at x. By (3.24) this is given by

$$q_x(r+\epsilon) - q_x(r) = \mathbf{P}(r \leq d_{M,k}(x) \leq r+\epsilon) \tag{3.27}$$

Since the sampling density ϕ is assumed to be continuous on C, for ϵ sufficiently small we may suppose that the kth nearest neighbour of x is the *only* point lying in this region. As illustrated in Figure 3.6 we must therefore have

- $k-1$ points in the ball $B_x(r)$, each selected with probability $\omega_x(r)$.
- Exactly one of the remaining $M-k$ points in the shell $B_x(r+\epsilon) \setminus B_x(r)$, selected with probability $\omega_x(r+\epsilon) - \omega_x(r)$.
- The remaining $M-k-1$ points in the region $C \setminus B_x(r+\epsilon)$, each selected with probability $1 - \omega_x(r+\epsilon)$

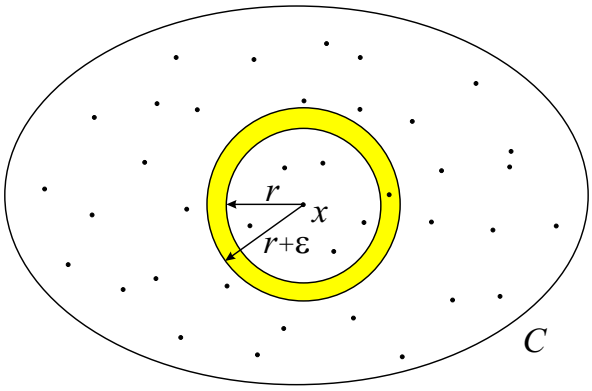

Figure 3.6: Exactly one point falls in the shaded region $B_x(r+\epsilon) \setminus B_x(r)$.

Using elementary combinatorial arguments we obtain

$$q_x(r+\epsilon) - q_x(r) = k\binom{M-1}{k}\omega_x(r)^{k-1}(1-\omega_x(r+\epsilon))^{M-k-1}(\omega_x(r+\epsilon)-\omega_x(r)) \tag{3.28}$$

and letting $\epsilon \to 0$ we obtain (3.26) as required. □

By (3.22), (3.25) and Lemma 3.3 it follows that

$$\mathcal{E}(d^\alpha_{M,k}) = \int_C \mathcal{E}(d^\alpha_{M,k}(x))\phi(x)\,dx \tag{3.29}$$

where

$$\mathcal{E}(d^\alpha_{M,k}(x)) = k\binom{M-1}{k}\int_0^{c_1} r^\alpha \omega_x(r)^{k-1}(1-\omega_x(r))^{M-k-1} d\omega_x(r) \tag{3.30}$$

3.4 Asymptotic expansions

We aim to find an asymptotic expression for (3.29) in terms of the number of points M as $M \to \infty$. We first prove some asymptotic expansions required for this purpose.

Lemma 3.4. *For any fixed $c, d > 0$,*

$$\left(1 - \frac{c}{M}\right)^d = 1 + O\left(\frac{1}{M}\right) \quad \textit{as} \quad M \to \infty \tag{3.31}$$

Proof. Using the series expansion $(1+x)^d = 1 + dx + O(x^2)$ as $x \to 0$ we get

$$\left(1 + \frac{c}{M}\right)^d = 1 - d\left(\frac{c}{M}\right) + O\left(\frac{1}{M^2}\right) = 1 + O\left(\frac{1}{M}\right) \quad \text{as} \quad M \to \infty \tag{3.32}$$

as required. □

Lemma 3.5. *For any fixed $c > 0$,*

$$\left(1 - \frac{c}{M}\right)^M = e^{-c}\left(1 + O\left(\frac{1}{M}\right)\right) \quad \textit{as} \quad M \to \infty \tag{3.33}$$

Proof. Taking $x = c/M$ in the expansion $\log(1-x) = -x + O(x^2)$ as $x \to 0$ we have

$$\log\left(1 - \frac{c}{M}\right)^M = M \log\left(1 - \frac{c}{M}\right) = M\left(\frac{-c}{M} + O\left(\frac{1}{M^2}\right)\right) = -c + O\left(\frac{1}{M}\right) \tag{3.34}$$

as $M \to \infty$, as required. □

Lemma 3.6 (The exponential convergence lemma). *Let $c > 0$ and $0 < \sigma < 1$ be constants. For every $\beta > 0$*

$$\left(1 - \frac{c}{M^\sigma}\right)^M = O\left(\frac{1}{M^\beta}\right) \quad \textit{as} \quad M \to \infty \tag{3.35}$$

Proof. Let $\beta > 0$ and define

$$y = \left(1 - \frac{c}{M^\sigma}\right)^M \tag{3.36}$$

so that

$$\log y = M \log\left(1 - \frac{c}{M^\sigma}\right) \tag{3.37}$$

Since $\log(1-x) = -x + O(x^2)$ as $x \to 0$, we have

$$\begin{aligned} \log y &= M\left(-\frac{c}{M^\sigma} + O\left(\frac{1}{M^{2\sigma}}\right)\right) \\ &= -cM^{1-\sigma}\left(1 + O\left(\frac{1}{M^\sigma}\right)\right) \quad \text{as} \quad M \to \infty \end{aligned} \tag{3.38}$$

and therefore

$$y = \exp(-cM^{1-\sigma}) \left(1 + O\left(\frac{1}{M^\sigma}\right)\right) \quad \text{as} \quad M \to \infty \tag{3.39}$$

For all $x > 0$ the summands in the power series expansion $e^x = \sum_{k=0}^{\infty} x^k/k!$ are all positive. Hence $e^x \geq x^k/k!$ for every integer $k \geq 0$ and thus $e^{-x} \leq k!/x^k$ for each $x > 0$ and $k \geq 0$. Since $c > 0$ we can apply this to $x = cM^{1-\sigma}$ so that for each $k \geq 0$

$$y \leq \frac{k!}{(cM^{1-\sigma})^k} \left(1 + O\left(\frac{1}{M^\sigma}\right)\right) \quad \text{as} \quad M \to \infty \tag{3.40}$$

Since $1 - \sigma > 0$, given any $\beta > 0$ we can choose k sufficiently large to ensure that $k(1-\sigma) > \beta$ and hence

$$y = O\left(\frac{1}{M^\beta}\right) \quad \text{as} \quad M \to \infty \tag{3.41}$$

as required. □

Lemma 3.7. *For any fixed $\sigma > 0$,*

$$\frac{\Gamma(M)}{\Gamma(M+\sigma)} = \frac{1}{M^\sigma} \left(1 + O\left(\frac{1}{M}\right)\right) \quad \textit{as} \quad M \to \infty \tag{3.42}$$

where Γ is the Euler gamma function.

Proof. Stirling's theorem [Artin 1964] states that for $x > 0$,

$$\Gamma(x) = \frac{\sqrt{2\pi} x^x e^{\theta/x}}{x^{1/2} e^x} \tag{3.43}$$

for some $0 < \theta < 12$. Hence

$$\begin{aligned} \frac{\Gamma(M)}{\Gamma(M+\sigma)} &= \left(\frac{M}{e}\right)^M \left(\frac{e}{M+\sigma}\right)^{M+\sigma} \left(\frac{M+\sigma}{M}\right)^{1/2} \left(\frac{e^{\theta_1/M}}{e^{\theta_2/(M+\sigma)}}\right) \\ &= H_{M,\sigma} \left(\frac{M}{e}\right)^{-\sigma} \left(\frac{M}{e}\right)^{M+\sigma} \left(\frac{e}{M+\sigma}\right)^{M+\sigma} \left(\frac{M+\sigma}{M}\right)^{1/2} \\ &= H_{M,\sigma} \left(\frac{e}{M}\right)^{\sigma} \left(1 - \frac{\sigma}{M+\sigma}\right)^{M+\sigma} \left(1 + \frac{\sigma}{M}\right)^{1/2} \end{aligned} \tag{3.44}$$

where

$$H_{M,,\sigma} = \exp\left(\frac{\theta_1 + \theta_2}{M+\sigma} - \frac{\theta_1 \sigma}{M(M+\sigma)}\right) \tag{3.45}$$

By Lemma 3.5

$$\left(1 - \frac{\sigma}{M+\sigma}\right)^{M+\sigma} = e^{-\sigma} \left(1 + O\left(\frac{1}{M+\sigma}\right)\right) \quad \text{as} \quad M \to \infty \tag{3.46}$$

By Lemma 3.4

$$\left(1-\frac{\sigma}{M}\right)^{1/2} = 1 + O\left(\frac{1}{M}\right) \quad \text{as} \quad M \to \infty \tag{3.47}$$

Expanding $H_{M,\sigma}$ via the power series of e^x we get

$$\begin{aligned} H_{M,\sigma} &= \exp\left(\frac{\theta_1+\theta_2}{M+\sigma} - \frac{\theta_1\sigma}{M(M+\sigma)}\right) \\ &= 1 + \frac{\theta_1+\theta_2}{M+\sigma} - \frac{\theta_1\sigma}{M(M+\sigma)} + O\left(\frac{1}{(M+\sigma)^2}\right) \\ &= 1 + O\left(\frac{1}{M+\sigma}\right) \quad \text{as} \quad M \to \infty \end{aligned} \tag{3.48}$$

Thus

$$\frac{\Gamma(M)}{\Gamma(M+\sigma)} = e^{-\sigma}\left(\frac{e}{M}\right)^{\sigma}\left(1+O\left(\frac{1}{M}\right)\right)\left(1+O\left(\frac{1}{M+\sigma}\right)\right) \quad \text{as} \quad M \to \infty \tag{3.49}$$

and since

$$\frac{1}{M+\sigma} = \frac{1}{M}\left(1+\frac{\sigma}{M}\right)^{-1} = \frac{1}{M}\left(1-\frac{\sigma}{M}+\frac{\sigma^2}{2M^2}-\ldots\right) = O\left(\frac{1}{M}\right) \quad \text{as} \quad M \to \infty \tag{3.50}$$

for all $\sigma > 0$ we conclude that

$$\frac{\Gamma(M)}{\Gamma(M+\sigma)} = \frac{1}{M^{\sigma}}\left(1+O\left(\frac{1}{M}\right)\right) \quad \text{as} \quad M \to \infty \tag{3.51}$$

as required. □

3.5 Boundary effects

The *boundary* of a set $C \subset \mathbb{R}^m$ is defined to be the set of points $x \in \mathbb{R}^m$ with the property that for every $r > 0$, the ball $B_x(r)$ has non–empty intersection with both C and $\mathbb{R}^m \backslash C$. If some property of a point $x \in C$ is influenced in some way by its proximity to the boundary of C then the property is said to be subject to *boundary effects*. Such effects are certainly significant when considering the expected kth nearest neighbour distance $\mathcal{E}(d^{\alpha}_{M,k}(x))$. This is because the neighbourhood balls $B_x(r)$ of a point x located near the boundary of C may intersect the boundary of C, in which case the probability measure of $B_x(r)$ (and therefore the probability that at least k points are selected from it) is likely to be *smaller* than that of a similar ball which does not intersect the boundary. The kth nearest neighbour distance for a point near the boundary of C is therefore likely to be *greater* than that for a point located away from the boundary of C.

The problem posed by boundary effects in determining the conditional expectation $\mathcal{E}(d^{\alpha}_{M,k}(x))$ derives from the fact it is defined as an integral with respect to the probability measure $\omega_x(r)$ of the neighbourhood balls $B_x(r)$ for $0 \leq r \leq c_1$. If $B_x(r)$ intersects the boundary of C then it will not be possible to specify its probability measure $\omega_x(r)$ without having exact information regarding the boundary of C in the neighbourhood of x. Thus we are not able to evaluate the integral explicitly.

This is not critical however, as we are only seeking a first order approximation for this integral in terms of the number of points M. We show that boundary effects do not influence this first order approximation, and occur as only a second order phenomenon.

3.5.1 Large balls are insignificant

Let $\delta > 0$ be fixed and let $x \in C$. We make the observation that any ball $B_x(r)$ having radius $r > \delta$ asymptotically contributes nothing to a polynomial expansion of the expectation $\mathcal{E}(d^{\alpha}_{M,k}(x))$ in terms of the number of points M. To see this, suppose that the kth nearest neighbour of x is at least a distance $\delta > 0$ away from it. Then the ball $B_x(\delta)$ must always contain at most $k-1$ points (distinct from x) and since the probability measure of $B_x(\delta)$ is strictly positive for $\delta > 0$ (condition **C.2**), this becomes increasingly unlikely as the total number of points increases without bound.

The following lemma makes this precise by showing that when evaluating (3.30), the error incurred by ignoring balls of radius $r > \delta$ for any fixed $\delta > 0$ is exponentially small in the limit as $M \to \infty$.

Lemma 3.8. *Let $0 < \delta < c_1$ be fixed. Then for every $\beta > 0$,*

$$I(\delta) = k\binom{M-1}{k}\int_{\delta}^{c_1} r^{\alpha}\omega(r)^{k-1}(1-\omega(r))^{M-k-1}\,d\omega(r) = O\left(\frac{1}{M^{\beta}}\right) \quad \textit{as} \quad M \to \infty \tag{3.52}$$

Proof. Since $\omega(r)$ is an increasing function of r we have that $1-\omega(r) \leq 1-\omega(\delta)$ for all $\delta \leq r \leq c_1$. Clearly, $r \leq c_1$ and $|\omega(r)| \leq 1$ so

$$I(\delta) \leq k\binom{M-1}{k}c_1^{\alpha+1}(1-\omega(\delta))^{M-k-1} \tag{3.53}$$

Furthermore, $(1-\omega(\delta))^{-k-1}$ is constant and $\binom{M-1}{k} = O((M-1)^k)$ as $M \to \infty$. Hence

$$I(\delta) \leq O((M-1)^k)(1-\omega(\delta))^M \tag{3.54}$$

Let $y = (1-\omega(\delta))^M$ so that $\log y = M\log(1-\omega(\delta))$. Then since $\omega(\delta) > 0$ is fixed there exists some constant $c > 0$ such that $\log(1-\omega(\delta)) \leq -c$. Hence $\log y \leq -cM$ from which it follows that $(1-\omega(\delta))^M \leq e^{-cM}$ for $c > 0$. From the power series expansion of e^x for $x > 0$ we see that $e^{-x} \leq n!/x^n$ for each $n \geq 0$. Given any $\beta > 0$ we may therefore choose some $n > \beta + k$ such that $(1-\omega(\delta))^M = O(1/M^{\beta+k})$, and the result follows by (3.54). □

Thus, without loss of generality we may choose to evaluate the integral in (3.30) either over the range $[0, c_1]$ as stated, or alternatively over the range $[0, \delta]$ for any fixed $\delta > 0$.

3.5.2 The interior region and the boundary region

Let $\delta > 0$ be fixed and let $B \subset C$ denote the boundary region of width δ defined in (3.4), so that B is the set of points in C that are within distance δ of its boundary. Let $A = C \setminus B$ denote the corresponding interior region.

By definition, for each $x \in A$ and $r \leq \delta$ the neighbourhood ball $B_x(r)$ is completely contained in C and is not therefore subject to boundary effects as described above. Furthermore, Lemma 3.8 shows that when evaluating (3.30), we may restrict our attention to those balls of radius $r < \delta$. Thus we need only consider boundary effects for the points $x \in B$ and we decompose (3.29) as

$$\mathcal{E}(d^{\alpha}_{M,k}) = \int_A \mathcal{E}(d^{\alpha}_{M,k}(x))\phi(x)\,dx + \int_B \mathcal{E}(d^{\alpha}_{M,k}(x))\phi(x)\,dx \tag{3.55}$$

so that all boundary effects are confined to the second integral in (3.55). As we shall see, since boundary effects are negligible for $x \in A$ we are able to find an exact asymptotic expression for the first integral in terms of the number of points M. Our task is therefore to construct the boundary region B in such a way that the second integral in (3.55) is of smaller order than the first in the limit as $M \to \infty$.

If $\delta > 0$ is small and the boundary of C is smooth it may be reasonable to suppose that the proportion of points that are within distance δ of the boundary will be small relative to the total number M. For example, suppose we select M points according to a uniform distribution from the unit square $C = [0,1]^2$. Since points are selected uniformly and since $[0,1]^2$ is of unit area, the probability that a point falls in some subset $A \subseteq C$ is equal to its area. For the unit square, the boundary region B of width δ is contained in the union of four rectangles, each having width δ and height 1, one of which is located adjacent to each edge of the square. Since the total area of these rectangles is 4δ, the probability that a point is selected from B is at most equal to 4δ and the number of points that are subject to boundary effects can therefore be controlled by choosing δ sufficiently small.

This illustrates the motivation behind condition **C.3** which requires the measure of the boundary region $C(\delta)$ to be bounded above by some constant multiple of its width δ.

3.5.3 Asymptotic shrinking of the boundary region

The problem with keeping $\delta > 0$ fixed is that the *proportion* of points falling in the boundary region B (and therefore subject to boundary effects) will remain constant as the number of points M increases. In order to ensure that the second integral in (3.55) is of smaller order than the first integral (relative to the number of points M), we in fact need that the proportion of points in B *decreases* as $M \to \infty$.

In other words we require that the *probability measure* of the boundary region B approaches zero as $M \to \infty$ and by condition **C.3**, it is sufficient that $\delta \to 0$ as $M \to \infty$

to ensure this. However, by defining δ in terms of M both the interior region A and the boundary region B now become dependent on M. In particular, A approaches C as $M \to \infty$ and consequently we cannot assert that every point of A is located at least at a *fixed* positive distance away from the boundary of C for all M.

In order that we may continue to ignore boundary effects in the first integral of (3.55) we therefore need a stronger result than that of Lemma 3.8. Specifically, we must define some $\delta \to 0$ as $M \to \infty$ such that every ball $B_x(r)$ of radius $r > \delta$ is asymptotically negligible regarding any polynomial expansion of (3.30) in terms of M. It turns out that for any $0 < \epsilon < 1/m$, taking $\delta = 1/M^\epsilon$ is sufficient to achieve this.

Lemma 3.9. *If* $0 < \epsilon < 1/m$ *and*

$$\delta = \frac{1}{M^\epsilon} \tag{3.56}$$

then for every $\beta > 0$*,*

$$I(\delta) = k\binom{M-1}{k}\int_\delta^{c_1} r^\alpha \omega_x(r)^{k-1}(1-\omega_x(r))^{M-k-1}\, d\omega_x(r) = O\left(\frac{1}{M^\beta}\right) \quad as \quad M \to \infty \tag{3.57}$$

Proof. Since $\omega_x(r)$ is an increasing function of r it follows that $1 - \omega_x(r) \leq 1 - \omega_x(\delta)$ for all $\delta \leq r \leq c_1$. Furthermore, by Lemma 3.2 we have that $\omega_x(\delta) \geq c_4\delta^m$ for some constant $c_4 > 0$. Hence $1 - \omega(r) \leq 1 - c_4\delta^m$ for all $\delta \leq r \leq c_1$ and since $r \leq c_1$ and $|\omega(r)| \leq 1$ we have that

$$I(\delta) \leq c_1^{\alpha+1} k\binom{M-1}{k}(1-c_4\delta^m)^{M-k-1} \tag{3.58}$$

Clearly, $\binom{M-1}{k} = O((M-1)^k)$ as $M \to \infty$ and since $\delta^m = o(1)$ it follows that $(1 - c_4\delta^m)^{-k-1} = O(1)$ as $M \to \infty$. Substituting for δ in (3.58) we thus obtain

$$I(\delta) \leq O((M-1)^k)\left(1 - \frac{c_4}{M^{m\epsilon}}\right)^M \quad \text{as} \quad M \to \infty \tag{3.59}$$

Hence, since $0 < m\epsilon < 1$ it follows by Lemma 3.4 that $(1 - c_4/M^{m\epsilon})^M$ converges to zero exponentially fast as $M \to \infty$ and the result follows. □

3.5.4 An integral over neighbourhood balls

Recall from (3.30) that

$$\mathcal{E}(d^\alpha_{M,k}(x)) = k\binom{M-1}{k}\int_0^{c_1} r^\alpha \omega_x(r)^{k-1}(1-\omega_x(r))^{M-k-1} d\omega_x(r) \tag{3.60}$$

In order to obtain the leading term in an asymptotic expansion for $\mathcal{E}(d^\alpha_{M,k})$ relative to the number of points M we will certainly have to evaluate the integral in (3.60), albeit

only for those points $x \in C \setminus C(\delta)$ and for r in the range $[0, \delta]$ (i.e. for those balls $B_x(r)$ that are completely contained in C).

To this end we change the variable of integration in (3.60) by defining $\omega = \omega_x(r)$. Since the probability measure of a ball of zero radius is zero we have that $\omega = 0$ when $r = 0$, and since the probability measure of a ball containing the whole of C is equal to one, it follows that $\omega = 1$ when $r = c_1$.

Let h_x denote the inverse function of ω_x so that $h_x(\omega)$ is the radius of the ball centred at x having probability measure ω. Then $r = h_x(\omega)$ and provided $h_x(\omega)$ exists over the range $0 \leq \omega \leq 1$, (3.60) becomes

$$\mathcal{E}(d^{\alpha}_{M,k}(x)) = k\binom{M-1}{k}\int_0^1 h_x(\omega)^{\alpha}\omega^{k-1}(1-\omega)^{M-k-1}d\omega \tag{3.61}$$

Clearly, $\omega_x(r)$ is an increasing function of r but in order to ensure that $h_x(\omega)$ is well defined we need that $\omega_x(r)$ is *strictly* increasing for all $0 \leq r \leq c_1$. In general however, $\omega_x(r)$ is not likely to be strictly increasing over the whole range $0 \leq r \leq c_1$, since the ball $B_x(r)$ may contain the whole of C before its radius r reaches c_1, the diameter of C – if $B_x(r_0)$ contains the whole of C then $\omega_x(r) = 1$ for all $r_0 \leq r \leq c_1$.

It emerges that we need only evaluate the integral in (3.60) explicitly for those points in the interior region A. By definition, if $x \in A$ then the balls $B_x(r)$ are completely contained in C for all $0 \leq r \leq \delta$. Thus, since $\phi > 0$ over C it follows that $\omega_x(r)$ is indeed strictly increasing for $0 \leq r \leq \delta$ and $x \in A$ and hence $h_x(\omega)$ is well defined for $0 \leq \omega \leq \omega_x(\delta)$.

For completeness we now show that the error incurred by neglecting balls of probability measure $\omega > \omega_x(\delta)$ (corresponding to balls of radius $r > \delta$) when evaluating the integral in (3.61) becomes exponentially small as $M \to \infty$. Note that if $\delta > 0$ then by Lemma 3.2 we have that $\omega_x(\delta) > 0$ for every $x \in C$.

Lemma 3.10. *If $0 < \epsilon < 1/m$ and $\delta = 1/M^{\epsilon}$ then for every $\beta > 0$,*

$$I(\delta) = k\binom{M-1}{k}\int_{\omega_x(\delta)}^1 h_x(\omega)^{\alpha}\omega^{k-1}(1-\omega)^{M-k-1}\,d\omega = O\left(\frac{1}{M^{\beta}}\right) \quad as \quad M \to \infty \tag{3.62}$$

where $h_x(\omega)$ is the radius of the ball centred at x having probability measure ω

Proof. Clearly, $1 - \omega \leq 1 - \omega_x(\delta)$ for all $\omega_x(\delta) \leq \omega \leq 1$. Furthermore, by Lemma 3.2 we know that $1 - \omega_x(\delta) \leq c_4\delta^m$ for some constant $c_4 > 0$. Since $|h_x(\omega)| \leq c_1$ and $|\omega| \leq 1$ it thus follows that

$$I(\delta) \leq c_1^{\alpha}k\binom{M-1}{k}(1 - c_4\delta^m)^{M-k-1} \tag{3.63}$$

It is easy to see that $\binom{M-1}{k} = O((M-1)^k)$ as $M \to \infty$ and since $\delta^m = o(1)$ it follows that $(1 - c_4\delta^m)^{-k-1} = O(1)$ as $M \to \infty$. Substituting for δ in (3.63) we get

$$I(\delta) \leq O((M-1)^k) \left(1 - \frac{c_4}{M^{m\epsilon}}\right)^M \quad \text{as} \quad M \to \infty \tag{3.64}$$

Hence, since $0 < m\epsilon < 1$ it follows by Lemma 3.4 that $(1 - c_4/M^{m\epsilon})^M$ converges to zero exponentially fast as $M \to \infty$ and the result follows. □

If we choose to evaluate $\mathcal{E}(d^\alpha_{M,k}(x))$ using (3.61) we may therefore restrict our attention to those balls $B_x(r)$ having probability measure in the range $[0, \omega_x(\delta)]$, where $\delta = 1/M^\epsilon$ for any $0 < \epsilon < 1/m$.

3.6 Asymptotic moments for a uniform sampling distribution

We shall now focus on deriving an asymptotic expression for $\mathcal{E}(d^\alpha_{M,k})$ in the case where the sample points are selected *uniformly* from some closed set $C \subset \mathbb{R}^m$ satisfying conditions **C.1** to **C.3**. It is a simple matter to show that the uniform distribution satisfies conditions **P.1** to **P.3** over any such set C. For convenience, we impose the condition that $\mu(C) = 1$ so that probability measure over C coincides with Lebesgue measure over C.

Theorem 3.1. *Let $C \subset \mathbb{R}^m$ be a closed set satisfying conditions **C.1** to **C.3** and suppose that $\mu(C) = 1$. Let $X = (X_1, \ldots, X_M)$ be a random sample of independent and identically distributed random variables where each X_i takes values uniformly at random from C. Let $d_{M,k}$ denote the distance between X_i and its kth nearest neighbour in X. Then for all $\rho > 0$,*

$$\mathcal{E}(d^\alpha_{M,k}) = \frac{c(m,\alpha,k,\phi)}{M^{\alpha/m}} + O\left(\frac{1}{M^{(\alpha+1)/m-\rho}}\right) \quad \textit{as} \quad M \to \infty \tag{3.65}$$

where

$$c(m,\alpha,k) = V_m^{-\alpha/m} \frac{\Gamma(k+\alpha/m)}{\Gamma(k)} \tag{3.66}$$

is a constant not depending on M.

Proof. Since $\mu(C) = 1$ and ϕ is uniform, (3.29) becomes

$$\mathcal{E}(d^\alpha_{M,k}) = \int_C \mathcal{E}(d^\alpha_{M,k}(x))\, dx \tag{3.67}$$

where

$$\mathcal{E}(d^\alpha_{M,k}(x)) = k\binom{M-1}{k} \int_0^1 h_x(\omega)^\alpha \omega^{k-1} (1-\omega)^{M-k-1}\, d\omega \tag{3.68}$$

and $h_x(\omega)$ is the radius of the ball centred at x having probability measure ω. By uniformity it also follows that the probability measure of the ball $B_x(r)$ is equal to the Lebesgue measure of its intersection with C, i.e.

$$\omega_x(r) = \mu(B_x(r) \cap C) \tag{3.69}$$

Let $0 < \epsilon < 1/m$ and define $\delta = 1/M^\epsilon$. Let B denote the boundary region of width δ as defined in (3.4), let $A = C \setminus B$ be its complement and write (3.67) as

$$\mathcal{E}(d^\alpha_{M,k}) = \int_A \mathcal{E}(d^\alpha_{M,k}(x))\,dx + \int_B \mathcal{E}(d^\alpha_{M,k}(x))\,dx \tag{3.70}$$

We treat the cases $x \in A$ and $x \in B$ separately.

Case (1): $x \in A$.

If $x \in A$ then x is *at least* a distance δ from the boundary of C so $B_x(r)$ is completely contained in C for all $0 \le r \le \delta$. Hence by (3.69) and since $\mu(C) = 1$ we obtain

$$\omega_x(r) = V_m r^m \quad \text{for all} \quad 0 \le r \le \delta \tag{3.71}$$

where V_m is the volume of the unit ball in $\mathbb{R}^m$. Inverting this, the radius $h_x(\omega)$ of the ball centred at x having probability measure ω satisfies

$$h_x(\omega) = V_m^{-1/m}\omega^{1/m} \tag{3.72}$$

provided the ball is completely contained in C. By Lemma 3.9 and (3.68) we know that for any $\beta > 0$,

$$\mathcal{E}(d^\alpha_{M,k}(x)) = k\binom{M-1}{k}\int_0^\delta h_x(\omega)^\alpha \omega^{k-1}(1-\omega)^{M-k-1}\,d\omega + O\left(\frac{1}{M^\beta}\right) \tag{3.73}$$

as $M \to \infty$. Since $x \in A$, the balls over which the integral in (3.73) is defined are *all* contained in C so

$$\mathcal{E}(d^\alpha_{M,k}(x)) = V_m^{-\alpha/m}k\binom{M-1}{k}\int_0^\delta \omega^{k+\alpha/m-1}(1-\omega)^{M-k-1}\,d\omega + O\left(\frac{1}{M^\beta}\right) \tag{3.74}$$

as $M \to \infty$. Changing the variable of integration and using Lemma 3.10 it thus follows that

$$\mathcal{E}(d^\alpha_{M,k}(x)) = V_m^{-\alpha/m}k\binom{M-1}{k}I_{M,k} + O\left(\frac{1}{M^\beta}\right) \quad \text{as} \quad M \to \infty \tag{3.75}$$

where

$$I_{M,k} = \int_0^1 \omega^{k+\alpha/m-1}(1-\omega)^{M-k-1}\,d\omega \tag{3.76}$$

We recognise the integral $I_{M,k}$ as the Beta function $B(a,b)$ with parameters $a = k+\alpha/m$ and $b = M-k$, defined by

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{3.77}$$

Hence

$$I_{M,k} = \frac{\Gamma(k+\alpha/m)\Gamma(M-k)}{\Gamma(M+\alpha/m)} \tag{3.78}$$

and writing

$$k\binom{M-1}{k} = \frac{\Gamma(M)}{\Gamma(M-k)\Gamma(k)} \tag{3.79}$$

in (3.75) we obtain

$$\mathcal{E}(d_{M,k}^{\alpha}(x)) = V_m^{-\alpha/m}\frac{\Gamma(k+\alpha/m)}{\Gamma(k)}\frac{\Gamma(M)}{\Gamma(M+\alpha/m)} + O\left(\frac{1}{M^{\beta}}\right) \quad \text{as} \quad M\to\infty \tag{3.80}$$

Applying Lemma 3.4 to (3.80) then leads to

$$\mathcal{E}(d_{M,k}^{\alpha}(x)) = V_m^{-\alpha/m}\frac{\Gamma(k+\alpha/m)}{\Gamma(k)}\frac{1}{M^{\alpha/m}}\left(1+O\left(\frac{1}{M}\right)\right) \quad \text{as} \quad M\to\infty \tag{3.81}$$

As a consequence of uniformity we see that this is independent of x. Furthermore, since $\mu(C) = 1$ the probability measure of A is equal to its Lebesgue measure $\mu(A)$. The first integral of (3.70) is thus given by

$$\int_A \mathcal{E}(d_{M,k}^{\alpha}(x))\,dx = \mu(A)\mathcal{E}(d_{M,k}^{\alpha}(x)) \tag{3.82}$$

Since $C = A\cup B$ is a disjoint union and $\mu(C) = 1$ we have $\mu(A) = 1-\mu(B)$ and by condition **C.3** we know that $\mu(B) = O(\delta)$ and hence $\mu(A) = 1-O(\delta)$ as $M\to\infty$. Furthermore, since $\delta = 1/M^{\epsilon}$ for $0<\epsilon<1/m$ it follows that $1/M = o(\delta)$ as $M\to\infty$ for all $m\geq 1$. Thus the $O(1/M)$ term in (3.81) is subsumed by the $O(\delta)$ term in this expression for $\mu(A)$ so

$$\int_A \mathcal{E}(d_{M,k}^{\alpha}(x))\,dx = V_m^{-\alpha/m}\frac{\Gamma(k+\alpha/m)}{\Gamma(k)}\frac{1}{M^{\alpha/m}} + O\left(\frac{\delta}{M^{\alpha/m}}\right) \quad \text{as} \quad M\to\infty \tag{3.83}$$

and hence

$$\int_A \mathcal{E}(d_{M,k}^{\alpha}(x))\,dx = \frac{c(m,\alpha,k)}{M^{\alpha/m}} + O\left(\frac{\delta}{M^{\alpha/m}}\right) \quad \text{as} \quad M\to\infty \tag{3.84}$$

where

$$c(m,\alpha,k) = V_m^{-\alpha/m}\frac{\Gamma(k+\alpha/m)}{\Gamma(k)} \tag{3.85}$$

Case (2): $x \in B$.

In this case, for any $0 \le r \le c_1$ the ball $B_x(r)$ is not necessarily contained in C so we cannot specify an exact expression for its probability measure $\omega_x(r) = \mu(B_x(r) \cap C)$. However, by condition **C.2** we know that there exists some constant $c_2 > 0$ such that

$$\omega_x(r) = \mu(B_x(r) \cap C) \ge c_2 r^m \quad \text{for all} \quad 0 \le r \le c_1 \tag{3.86}$$

Inverting this we see that the radius $h_x(\omega)$ of the ball centred at x of probability measure ω satisfies

$$h(\omega_x) \le c_2^{-1/m}\omega_x^{1/m} \tag{3.87}$$

regardless of whether or not the ball is completely contained in C. Substituting this into (3.68) we obtain

$$\mathcal{E}(d_{M,k}^{\alpha}(x)) \le c_2^{-\alpha/m} k \binom{M-1}{k} I_{M,k} + O\left(\frac{1}{M^{\beta}}\right) \quad \text{as} \quad M \to \infty \tag{3.88}$$

where $I_{M,k}$ is defined in (3.76). Proceeding as in Case (1) we get

$$\mathcal{E}(d_{M,k}^{\alpha}(x)) \le c_2^{-\alpha/m}\frac{\Gamma(k+\alpha/m)}{\Gamma(k)}\frac{\Gamma(M)}{\Gamma(M+\alpha/m)} \tag{3.89}$$

so by Lemma 3.4 it follows that

$$\mathcal{E}(d_{M,k}^{\alpha}(x)) = O\left(\frac{1}{M^{\alpha/m}}\right) \quad \text{as} \quad M \to \infty \tag{3.90}$$

Hence, the second integral of (3.70) satisfies

$$\int_B \mathcal{E}(d_{M,k}^{\alpha}(x))\,dx = \mu(B)O\left(\frac{1}{M^{\alpha/m}}\right) \quad \text{as} \quad M \to \infty \tag{3.91}$$

and since $\mu(B) = O(\delta)$ as $M \to \infty$ (condition **C.3**) we conclude that

$$\int_B \mathcal{E}(d_{M,k}^{\alpha}(x))\,dx = O\left(\frac{\delta}{M^{\alpha/m}}\right) \quad \text{as} \quad M \to \infty \tag{3.92}$$

Combining Case (1) and Case (2) via (3.70), (3.84) and (3.92) we obtain

$$\mathcal{E}(d_{M,k}^{\alpha}) = \frac{c(m,\alpha,k)}{M^{\alpha/m}} + O\left(\frac{\delta}{M^{\alpha/m}}\right) \quad \text{as} \quad M \to \infty \tag{3.93}$$

and substituting for $\delta = 1/M^{\epsilon}$ it follows that for any $0 < \epsilon < 1/m$,

$$\mathcal{E}(d^{\alpha}_{M,k}) = \frac{c(m,\alpha,k)}{M^{\alpha/m}} + O\left(\frac{1}{M^{\alpha/m+\epsilon}}\right) \quad \text{as} \quad M \to \infty \tag{3.94}$$

Finally, taking $\epsilon = 1/m - \rho$ in (3.94) we conclude that for all $\rho > 0$,

$$\mathcal{E}(d^{\alpha}_{M,k}) = \frac{c(m,\alpha,k)}{M^{\alpha/m}} + O\left(\frac{1}{M^{(\alpha+1)/m-\rho}}\right) \quad \text{as} \quad M \to \infty \tag{3.95}$$

as required. □

Remark: Equation (3.80) states that for each $x \in A$,

$$\mathcal{E}(d^{\alpha}_{M,k}(x)) = V_m^{-\alpha/m}\frac{\Gamma(k+\alpha/m)}{\Gamma(k)}\frac{\Gamma(M)}{\Gamma(M+\alpha/m)} + O\left(\frac{1}{M^{\beta}}\right) \tag{3.96}$$

as $M \to \infty$. This expression for $\mathcal{E}(d^{\alpha}_{M,k}(x))$ is *exact* to any polynomial order in M and displays a remarkable separation of the k–dependence and the M–dependence. We also note an intriguing symmetry between the near neighbour index k and the total number of points M.

3.7 Asymptotic moments for a non–uniform sampling distribution

We now extend Theorem 3.1 to the more general case where M points are selected from C according to *any* sampling distribution Φ satisfying **P.1** to **P.3**. Intuitively speaking, we should not expect the asymptotic behaviour of near neighbour distances determined by such sampling distributions to be significantly different to that observed in the uniform case. This is because, subject to conditions **P.1** to **P.3**, the sampling density at any point in a small neighbourhood of a point x will be approximately equal to the density at x. Hence the sampling density is approximately uniform in small neighbourhoods and as we have seen, it is precisely these small neighbourhoods that determine the asymptotic behaviour of the expected near neighbour distances as $M \to \infty$.

In order to deal with more general sampling distributions, we must make the further condition that the set $C \subset \mathbb{R}^m$ is *convex*. Compact convex bodies in $\mathbb{R}^m$ have been shown to satisfy conditions **C.1** to **C.3** in Proposition 3.1.

Theorem 3.2. *Let C be a compact convex body in $\mathbb{R}^m$. Let $X = (X_1, \ldots, X_M)$ be a random sample of independent and identically distributed random variables where each X_i takes values in C according to a probability distribution Φ satisfying conditions **P.1**,*

P.2 *and* ***P.3****. Let $d_{M,k}$ denote the distance between X_i and its kth nearest neighbour in the random sample X. Then for all $\rho > 0$,*

$$\mathcal{E}(d_{M,k}^{\alpha}) = \frac{c(m,\alpha,k,\phi)}{M^{\alpha/m}} + O\left(\frac{1}{M^{(\alpha+1)/m-\rho}}\right) \quad as \quad M \to \infty \tag{3.97}$$

where

$$c(m,\alpha,k,\phi) = V_m^{-\alpha/m}\frac{\Gamma(k+\alpha/m)}{\Gamma(k)}\int_C \phi(x)^{1-\alpha/m}\,dx \tag{3.98}$$

is a constant not depending on M.

Proof. We need to find an asymptotic expression for the integral

$$\mathcal{E}(d_{M,k}^{\alpha}) = \int_C \mathcal{E}(d_{M,k}^{\alpha}(x))\phi(x)\,dx \tag{3.99}$$

as the number of points M increases where

$$\mathcal{E}(d_{M,k}^{\alpha}(x)) = k\binom{M-1}{k}\int_0^1 h_x(\omega)^{\alpha}\omega^{k-1}(1-\omega)^{M-k-1}\,d\omega \tag{3.100}$$

and $h_x(\omega)$ is the radius of the ball centred at x having probability measure ω. Recall that

$$\omega_x(r) = \int_{B_x(r)\cap C}\phi(t)\,dt \tag{3.101}$$

In order to evaluate (3.100) we obtain an expression for $h_x(\omega)$ by exploiting the convexity of C. That C is convex implies that it is connected, and since ϕ is continuous on C (condition **P.2**) we can therefore apply the first mean value theorem of the *integral* calculus to ϕ . This asserts the existence of a point $\xi_1 \in B_x(r)\cap C$ such that

$$\omega_x(r) = \phi(\xi_1)\mu(B_x(r)\cap C) \tag{3.102}$$

Furthermore, since ϕ is differentiable at every point of C (condition **P.3**) we can also apply the first mean value theorem of the *differential* calculus to ϕ. Hence there exists a point ξ_2 on the line segment joining x and ξ_1 such that

$$\phi(\xi_1) = \phi(x) + (x-\xi_1)\phi'(\xi_2) \tag{3.103}$$

and since C is convex we have that $\xi_2 \in C$. Furthermore, since all partial derivatives of ϕ are assumed bounded on C we have

$$\phi(\xi_1) = \phi(x) + O(|x-\xi_1|) \tag{3.104}$$

Let $0 < \epsilon < 1/m$ and define $\delta = 1/M^{\epsilon}$. Since $\xi_1 \in B_x(r)\cap C$ we may assume that $|x-\xi_1| \le r$ and hence

$$\phi(\xi_1) = \phi(x) + O(\delta) \quad \text{as} \quad M \to \infty \tag{3.105}$$

for all $0 \leq r \leq \delta$. Substituting in (3.102) we see that for all $x \in C$

$$\omega_x(r) = (\phi(x) + O(\delta))\mu(B_x(r) \cap C) \quad \text{as} \quad M \to \infty \tag{3.106}$$

provided $0 \leq r \leq \delta$.

Let B denote the boundary region of width δ, let $A = C \setminus B$ be the corresponding interior region and write (3.99) as

$$\mathcal{E}(d^{\alpha}_{M,k}) = \int_A \mathcal{E}(d^{\alpha}_{M,k}(x))\phi(x)\,dx + \int_B \mathcal{E}(d^{\alpha}_{M,k}(x))\phi(x)\,dx \tag{3.107}$$

Case (1): $x \in A$.

If $x \in A$ then x is *at least* a distance δ from the boundary of C so $B_x(r)$ is completely contained in C for all $0 \leq r \leq \delta$. In this case we have that $\mu(B_x(r) \cap C) = \mu(B_x(r)) = V_m r^m$ where V_m is the volume of the unit ball in $\mathbb{R}^m$ and hence by (3.106) we obtain

$$\omega_x(r) = (\phi(x) + O(\delta))V_m r^m \quad \text{as} \quad M \to \infty \tag{3.108}$$

Rearranging this we get

$$r^m = \frac{\omega_x(r)}{V_m\phi(x)}\left(1 + \frac{O(\delta)}{\phi(x)}\right)^{-1} \tag{3.109}$$

Since the sampling distribution Φ satisfies conditions **P.1** to **P.3**, by Lemma 3.1 there exist constants a_1 and a_2 such that $0 < a_1 \leq \phi(x) \leq a_2 < \infty$ for all $x \in C$. The existence of a_1 ensures that (3.109) is well defined for all $x \in A$, and the existence of both a_1 and a_2 implies that

$$\left(1 + \frac{O(\delta)}{\phi(x)}\right)^{-1} = 1 + O(\delta) \quad \text{as} \quad M \to \infty \tag{3.110}$$

Furthermore, since $m \geq 1$ is fixed it follows that $(1 + O(\delta))^{1/m} = 1 + O(\delta)$ as $M \to \infty$ so

$$r = \left(\frac{\omega_x(r)}{V_m\phi(x)}\right)^{1/m}(1 + O(\delta)) \quad \text{as} \quad M \to \infty \tag{3.111}$$

Hence the inverse function $r = h_x(\omega)$ is given by

$$h_x(\omega) = \left(\frac{\omega}{V_m\phi(x)}\right)^{1/m}(1 + O(\delta)) \quad \text{as} \quad M \to \infty \tag{3.112}$$

and since $\alpha \geq 1$ is fixed it follows that $(1 + O(\delta))^{\alpha} = 1 + O(\delta)$ as $M \to \infty$ so

$$h_x(\omega)^{\alpha} = \left(\frac{\omega}{V_m\phi(x)}\right)^{\alpha/m}(1 + O(\delta)) \quad \text{as} \quad M \to \infty \tag{3.113}$$

provided $B_x(r)$ is completely contained in C. By Lemma 3.9 and (3.100) we have that for any $\beta > 0$,

$$\mathcal{E}(d_{M,k}^{\alpha}(x)) = k\binom{M-1}{k}\int_0^{\delta} h_x(\omega)^{\alpha}\omega^{k-1}(1-\omega)^{M-k-1}\,d\omega + O\left(\frac{1}{M^{\beta}}\right) \tag{3.114}$$

as $M \to \infty$. Hence, since $x \in A$ the balls over which the integral in (3.114) is defined are all contained in C so

$$\begin{aligned}\mathcal{E}(d_{M,k}^{\alpha}(x)) = (V_m\phi(x))^{-\alpha/m}k\binom{M-1}{k}(1+O(\delta))\int_0^{\delta}\omega^{k+\alpha/m-1}(1-\omega)^{M-k-1}\,d\omega \\ + O\left(\frac{1}{M^{\beta}}\right) \quad \text{as} \quad M \to \infty\end{aligned} \tag{3.115}$$

Changing the variable of integration and using Lemma 3.10 we thus obtain

$$\mathcal{E}(d_{M,k}^{\alpha}(x)) = (V_m\phi(x))^{-\alpha/m}k\binom{M-1}{k}(1+O(\delta))I_{M,k} + O\left(\frac{1}{M^{\beta}}\right) \tag{3.116}$$

as $M \to \infty$ where

$$I_{M,k} = \int_0^1 \omega^{k+\alpha/m-1}(1-\omega)^{M-k-1}\,d\omega \tag{3.117}$$

As in Theorem 3.1 we recognize $I_{M,k}$ as the Beta function $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ with parameters $a = k + \alpha/m$ and $b = M - k$. Hence

$$I_{M,k} = \frac{\Gamma(k+\alpha/m)\Gamma(M-k)}{\Gamma(M+\alpha/m)} \tag{3.118}$$

and writing

$$k\binom{M-1}{k} = \frac{\Gamma(M)}{\Gamma(M-k)\Gamma(k)} \tag{3.119}$$

we get

$$\mathcal{E}(d_{M,k}^{\alpha}(x)) = (V_m\phi(x))^{-\alpha/m}\frac{\Gamma(k+\alpha/m)}{\Gamma(k)}\frac{\Gamma(M)}{\Gamma(M+\alpha/m)}(1+O(\delta)) \tag{3.120}$$

as $M \to \infty$. Once again we note the separation and symmetry of the k–dependence and the M–dependence in this expression. By Lemma 3.4,

$$\frac{\Gamma(M)}{\Gamma(M+\alpha/m)} = \frac{1}{M^{\alpha/m}}\left(1 + O\left(\frac{1}{M}\right)\right) \quad \text{as} \quad M \to \infty \tag{3.121}$$

and since $1/M = o(\delta)$ as $M \to \infty$ the $O(1/M)$ term is subsumed by the $O(\delta)$ term and we have

$$\mathcal{E}(d_{M,k}^{\alpha}(x)) = (V_m\phi(x))^{-\alpha/m}\frac{\Gamma(k+\alpha/m)}{\Gamma(k)}\frac{1}{M^{\alpha/m}}(1+O(\delta)) \quad \text{as} \quad M \to \infty \tag{3.122}$$

Thus the first integral in (3.107) is equal to

$$\int_A \mathcal{E}(d_{M,k}^{\alpha}(x))\phi(x)\,dx = \frac{c'(m,\alpha,k,\phi)}{M^{\alpha/m}}(1+O(\delta)) \quad \text{as} \quad M\to\infty \tag{3.123}$$

where

$$c'(m,\alpha,k,\phi) = V_m^{-\alpha/m}\frac{\Gamma(k+\alpha/m)}{\Gamma(k)}\int_A \phi(x)^{1-\alpha/m}\,dx \tag{3.124}$$

To complete Case (1), we show that the error incurred by replacing the integral in (3.124) with the equivalent integral over C is at most of order $O(\delta)$ as $M\to\infty$. First note that since $0 < a_1 \le \phi(x) \le a_2 < \infty$ for each $x \in C$, it follows that $|\phi(x)^{1-\alpha/m}| \le a_3 < \infty$ where $a_3 = \max\{1/a_1^{1-\alpha/m}, a_2^{1-\alpha/m}\}$. Furthermore, by condition **C.3** there exists some constant c_3 such that $\mu(B) \le c_3\delta$. Hence

$$\int_B \phi(x)^{1-\alpha/m}\,dx \le a_3c_3\delta = O(\delta) \quad \text{as} \quad M\to\infty \tag{3.125}$$

and since $C = A\cup B$ is a disjoint union we have

$$\int_A \phi(x)^{1-\alpha/m}\,dx = \int_C \phi(x)^{1-\alpha/m}\,dx + O(\delta) \quad \text{as} \quad M\to\infty \tag{3.126}$$

Thus, by (3.123) and (3.124) we conclude that for every $x \in A$,

$$\int_A \mathcal{E}(d_{M,k}^{\alpha}(x))\phi(x)\,dx = \frac{c(m,\alpha,k,\phi)}{M^{\alpha/m}} + O\left(\frac{\delta}{M^{\alpha/m}}\right) \quad \text{as} \quad M\to\infty \tag{3.127}$$

where

$$c(m,\alpha,k,\phi) = V_m^{-\alpha/m}\frac{\Gamma(k+\alpha/m)}{\Gamma(k)}\int_C \phi(x)^{1-\alpha/m}\,dx \tag{3.128}$$

Case (2): $x \in B$.

In this case, for any $0 \le r \le c_1$ the ball $B_x(r)$ is not necessarily contained in C and we cannot specify an exact expression for its probability measure $\omega_x(r) = \mu(B_x(r)\cap C)$. However, by Lemma 3.2 there exists some constant $c_4 > 0$ such that for all $x \in C$ and $0 \le r \le c_1$ we have that

$$\omega_x(r) \ge c_4 r^m \tag{3.129}$$

Inverting this, we obtain an upper bound for $h_x(\omega)$, defined to be the radius of the sphere centred at x of probability measure ω, given by

$$h_x(\omega) \le c_4^{-1/m}\omega^{1/m} \tag{3.130}$$

and substituting this into (3.100) we obtain

$$\mathcal{E}(d_{M,k}^{\alpha}(x)) \leq c_4^{-\alpha/m} k \binom{M-1}{k} I_{M,k} \tag{3.131}$$

where $I_{M,k}$ is as defined in (3.117). Proceeding as in Case (1) we get

$$\mathcal{E}(d_{M,k}^{\alpha}(x)) \leq c_4^{-\alpha/m} \frac{\Gamma(k+\alpha/m)}{\Gamma(k)} \frac{\Gamma(M)}{\Gamma(M+\alpha/m)} \tag{3.132}$$

and by Lemma 3.4,

$$\mathcal{E}(d_{M,k}^{\alpha}(x)) = O\left(\frac{1}{M^{\alpha/m}}\right) \quad \text{as} \quad M \to \infty \tag{3.133}$$

Hence, the second integral of (3.107) satisfies

$$\int_B \mathcal{E}(d_{M,k}^{\alpha}(x))\phi(x)\,dx = O\left(\frac{1}{M^{\alpha/m}}\right) \int_B \phi(x)\,dx \quad \text{as} \quad M \to \infty \tag{3.134}$$

For each $x \in B$, by Lemma 3.1 there exists some constant a_2 such that $|\phi(x)| \leq a_2 < \infty$ so the integral of $\phi(x)$ over B is therefore bounded above by $a_2\mu(B)$. Furthermore, by condition **C.3** there exists some constant c_3 such that $\mu(B) \leq c_3\delta$ so

$$\int_B \phi(x)\,dx = O(\delta) \quad \text{as} \quad M \to \infty \tag{3.135}$$

and hence

$$\int_B \mathcal{E}(d_{M,k}^{\alpha}(x))\,dx = O\left(\frac{\delta}{M^{\alpha/m}}\right) \quad \text{as} \quad M \to \infty \tag{3.136}$$

Combining Case (1) and Case (2) via (3.107), (3.127) and (3.136) we obtain

$$\mathcal{E}(d_{M,k}^{\alpha}) = \frac{c(m,\alpha,k,\phi)}{M^{\alpha/m}} + O\left(\frac{\delta}{M^{\alpha/m}}\right) \quad \text{as} \quad M \to \infty \tag{3.137}$$

and substituting for $\delta = 1/M^{\epsilon}$ it follows that for any $0 < \epsilon < 1/m$,

$$\mathcal{E}(d_{M,k}^{\alpha}) = \frac{c(m,\alpha,k,\phi)}{M^{\alpha/m}} + O\left(\frac{1}{M^{\alpha/m+\epsilon}}\right) \quad \text{as} \quad M \to \infty \tag{3.138}$$

Finally, taking $\epsilon = 1/m - \rho$ in (3.138) we conclude that for all $\rho > 0$,

$$\mathcal{E}(d_{M,k}^{\alpha}) = \frac{c(m,\alpha,k,\phi)}{M^{\alpha/m}} + O\left(\frac{1}{M^{(\alpha+1)/m-\rho}}\right) \quad \text{as} \quad M \to \infty \tag{3.139}$$

where

$$c(m,\alpha,k,\phi) = V_m^{-\alpha/m} \frac{\Gamma(k+\alpha/m)}{\Gamma(k)} \int_C \phi(x)^{1-\alpha/m}\,dx \tag{3.140}$$

as required. □

3.8 Nearest neighbour distances for fractal sets

We now turn our attention to the distribution of kth nearest neighbour distances in a set of points that are confined to some *fractal* subset $C \subset \mathbb{R}^m$, defined to be a set having non–integral dimension $d < m$. This is motivated by the study of chaotic dynamical systems, whose trajectories in state space $X \subset \mathbb{R}^m$ are often confined to some fractal set (the attractor of the system).

In (3.30) the expected distance from an arbitrary point x to its kth nearest neighbour in a set of M points is defined in terms of the probability measure $\omega_x(r)$ of the neighbourhood balls $B_x(r)$ centred at x and having radius r. When C is of integral dimension, the probability measure $\omega_x(r)$ can be defined in terms of Lebesgue measure. However, if C is of non–integral dimension in $\mathbb{R}^m$, the Lebesgue measure of C is zero and in order to exploit (3.30) we must define a more subtle notion of its measure.

First we describe the *Hausdorff measure* on a fractal set C. This is defined purely in terms of the geometric properties of C and as a result it may be rather simplistic in the case where C is the attractor of a dynamical system as it does not take into account how frequently a typical trajectory of the system visits various parts of its attractor. In this sense the Hausdorff measure corresponds to *uniform density* over C. Following this we define a more natural measure on the attractor of a dynamical system which does take relative densities across the attractor into account.

3.8.1 Hausdorff measure

For any $\delta > 0$, a *δ–cover* of C is defined to be a countable collection of sets $\{E_i\}_1^\infty$, each having diameter $|E_i| < \delta$ such that C is contained their union. The *δ–approximating s–dimensional Hausdorff measure* of C is defined to be

$$H_\delta^s(C) = \inf\left\{\sum_{i=1}^{\infty} |E_i|^s : C \subset \cup_{i=1}^{\infty} E_i, |E_i| \leq \delta\right\} \tag{3.141}$$

where the infimum is taken over all δ–covers of C. Since the class of permissible covers decreases as δ decreases it follows that H_δ^d increases as $\delta \to 0$. The *s-dimensional Hausdorff measure* of C is then defined by

$$H^s(C) = \lim_{\delta \to 0} H_\delta^s(C) \tag{3.142}$$

Since $H_\delta^s(C)$ is a monotonic decreasing function of s, it can be shown that there exists a unique transition point d, called the *Hausdorff dimension* of C, such that

$$H^s(C) = \begin{cases} \infty & \text{for} \quad s < d \\ 0 & \text{for} \quad s > d \end{cases} \tag{3.143}$$

The d–dimensional Hausdorff measure of C will be called the *Hausdorff measure* of C. In general, $0 \leq H^d(C) \leq \infty$. However, in order to employ the Hausdorff measure

as a probability measure on C, we must restrict our attention to those sets C having $0 < H^d(C) < \infty$. Such sets are known as d–sets, see [Falconer 1990] and we assume without loss of generality that $H^d(C) = 1$ for such sets.

Having defined the Hausdorff dimension C, the Hausdorff measure of any subset $A \subset C$ is then defined by

$$H^d(A) = \lim_{\delta \to 0} \inf \left\{ \sum_{i=1}^{\infty} |E_i|^s : A \subset \cup_{i=1}^{\infty} E_i, |E_i| \leq \delta \right\} \tag{3.144}$$

where the infimum is taken over all δ–covers of A.

3.8.2 An integral representation of the moments

Let $x \in C$ and define the probability measure of its neighbourhood balls $B_x(r) \subset \mathbb{R}^m$ by

$$\omega_x(r) = H^d(B_x(r) \cap C) \tag{3.145}$$

so that $\omega_x(r)$ is the Hausdorff measure of that part of C contained in $B_x(r)$. Recall from (3.61) that

$$\mathcal{E}(d_{M,k}^{\alpha}(x)) = k \binom{M-1}{k} \int_0^1 h_x(\omega)^{\alpha} \omega^{k-1} (1-\omega)^{M-k-1} d\omega \tag{3.146}$$

where $h_x(\omega)$ is the radius of the ball centred at x having probability measure ω.

The inverse function $h_x(\omega)$

To obtain the leading term in an asymptotic expansion for (3.146) as $M \to \infty$ we need to find a first order expression for $h_x(\omega)$ in terms of ω. For some $\epsilon > 0$, using arguments similar to those employed in the proof of Lemma 3.9 we can show that neighbourhood balls of radius $r > 1/M^{\epsilon}$ are insignificant in determining the leading term in an asymptotic expansion for (3.146) as $M \to \infty$. Hence the following condition is sufficient to ensure the existence of a first order expression for $h_x(\omega)$ in terms of ω.

C.4 There exist constants $c, d > 0$ such that

$$\omega_x(r) = cr^d + o(r^d) \quad \text{as} \quad r \to 0 \tag{3.147}$$

where $c = c(x)$ and $d = d(x)$ may depend on x.

If **C.4** holds then a first order expression for $h_x(\omega)$ is given by

$$h_x(\omega) = (\omega/c)^{1/d}(1 + o(1)) \quad \text{as} \quad r \to 0 \tag{3.148}$$

Self-similarity

Taking logarithms of both sides of (3.147) we see that

$$d = \lim_{r \to 0} \left(\frac{\log(\omega_x(r))}{\log(r)} \right) \tag{3.149}$$

By analogy with (3.143) and other definitions of fractal dimension (see [Falconer 1990]), the exponent d of (3.147) can be interpreted as the (local) fractal dimension of C in the neighbourhood of x. Most elementary definitions of fractal dimension measure the 'average' dimension over the whole set. However many fractal sets (especially those that arise as attractors of chaotic systems) are not uniformly dense.

The variation in density over a fractal set is captured by its *generalised dimension*, defined as follows. We divide the imbedding space into $N(r)$ cells of size r, and let p_i be the probability that a point of the set lies in the ith cell. For each $q \neq 1$ the generalised (box-counting) dimension D_q is defined by

$$D_q = \lim_{r \to 0} \left(\frac{1}{q-1} \right) \frac{\log \sum_{i=1}^{N(r)} p_i^q}{\log r} \tag{3.150}$$

A set is then said to be *self-similar* if $D_q = D_0$ for all q. If C is self-similar then we might expect that the exponent d of (3.147) is independent of any particular $x \in C$.

3.8.3 A conjectured first order approximation

Let μ be a measure concentrated on a set C. If μ is a probability measure we say that a sequence of points $\{x_i\}$ is *distributed in C according to μ* if for every μ-measurable subset $A \subset C$,

$$\frac{1}{M} \sum_{i=1}^{M} I_A(x_i) \to \mu(A) \quad \text{as} \quad M \to \infty \tag{3.151}$$

where I_A denotes the indicator function for the set A.

We propose the following conjecture.

Conjecture 3.1. *Let $C \subset \mathbb{R}^m$, let d be its Hausdorff dimension, let H^d denote the d–dimensional Hausdorff measure on C and suppose that $H^d(C) = 1$. Let $\{x_1, \ldots, x_M\}$ be a sequence of points distributed in C according to H^d and let $k \geq 1$ be a fixed integer. Then if $x \in C$ satisfies condition **C.4**,*

$$\mathcal{E}(d_{M,k}^{\alpha}(x)) \sim \frac{c}{M^{\alpha/d}} \quad \text{as} \quad M \to \infty \tag{3.152}$$

where $c = c(\alpha, k, d)$ is a constant not depending on M.

Remark: If this conjecture holds, it may provide a useful method of estimating the local dimension of C at x.

3.8.4 Orders of magnitude

While finding an *exact* expression for the first order term in the asymptotic expansion of (3.146) in terms of the number of points M, less ambitious would be to determine its order of magnitude. To achieve this we require the following condition.

C.5 There exist constants $c, d > 0$ such that

$$\omega_x(r) \geq cr^d \quad \text{for all} \quad 0 \leq r \leq c_1 \tag{3.153}$$

where $c = c(x)$ and $d = d(x)$ may depend on x.

If **C.5** holds for all $x \in C$ then $\omega_x(r) \geq \tilde{c}r^{\tilde{d}}$ where $\tilde{c} = \max\{c(x) \,\big|\, x \in C\}$ and $\tilde{d} = \min\{d(x) \,\big|\, x \in C\}$, in which case

$$h_x(\omega) \leq (\omega/\tilde{c})^{1/\tilde{d}} \quad forall \quad x \in C \tag{3.154}$$

If C is self-similar then we might expect that $d = d(x)$ is independent of any particular $x \in C$ and furthermore that d is (related to) the Hausdorff dimension of C. In view of this we define the following (stronger) condition.

C.6 There exists a constant $c > 0$ such that for all $x \in C$

$$\omega_x(r) \geq cr^d \quad \text{for all} \quad 0 \leq r \leq c_1 \tag{3.155}$$

where d is the Hausdorff dimension of C.

Proposition 3.2. *Let $C \subset \mathbb{R}^m$, let d be its Hausdorff dimension, let H^d denote the d–dimensional Hausdorff measure on C and suppose that $H^d(C) = 1$. Let $\{x_1, \ldots, x_M\}$ be a sequence of points distributed in C according to H^d and let $k \geq 1$ be a fixed integer. Then if $x \in C$ satisfies condition **C.6**, the αth moment of the distance between x and its kth nearest neighbour in the set $\{x_1, \ldots, x_M\}$ satisfies*

$$\mathcal{E}(d^\alpha_{M,k}(x)) = O\left(\frac{1}{M^{\alpha/d}}\right) \quad as \quad M \to \infty \tag{3.156}$$

Proof. By condition **C.6** there exists some $c > 0$ such that $\omega_x(r) \geq cr^d$ for all $0 \leq r \leq c_1$. Hence, the radius $h_x(\omega)$ of the ball centred at x having probability measure ω satisfies $h_x(\omega) \leq (\omega/c)^{1/d}$. By (3.146) we thus have that

$$\mathcal{E}(d^\alpha_{M,k}(x)) \leq c_4^{-\alpha/d} k \binom{M-1}{k} \int_{\omega=0}^{1} \omega^{k+\alpha/d-1}(1-\omega)^{M-k-1} d\omega \tag{3.157}$$

This integral is the Beta function with parameters $k + \alpha/d$ and $M - k$ so

$$\mathcal{E}(d^\alpha_{M,k}(x)) \leq c_4^{-\alpha/d} \frac{\Gamma(k+\alpha/d)}{\Gamma(k)} \frac{\Gamma(M)}{\Gamma(M+\alpha/d)} \tag{3.158}$$

and the result follows by Lemma 3.7. □

We would like to show that condition **C.6** holds for all $x \in C$. Failing this, we might settle for showing that **C.6** holds for H^d–almost all $x \in C$, in the sense that for some $c > 0$, the set

$$C' = \left\{ x \in C \,|\, \omega_x(r) < cr^d \quad \text{for some} \quad 0 < r \leq c_1 \right\} \tag{3.159}$$

is of H^d–measure zero. While we have been unable to prove this, we have obtained a weaker result where the 'for some' condition is replaced by a 'for all' condition.

Lemma 3.11. *For any $\delta > 0$ and $0 < c < 1$, the set*

$$C'' = \left\{ x \in C \,|\, \omega_x(r) < cr^d \quad \textit{for all} \quad 0 < r \leq \delta \right\} \tag{3.160}$$

is of H^d–measure zero.

Proof. Let $\{U_i\}_1^\infty$ be any δ–cover of C''. Then

$$H^d(C'') = H^d(C'' \cap C) \leq H^d(\cup U_i \cap C) \leq H^d(\cup (U_i \cap C)) \leq \sum H^d(U_i \cap C) \tag{3.161}$$

For each U_i intersecting C'' choose some $x_i \in C'' \cap U_i$ and define $B_i = B_{x_i}(|U_i|)$ to be the ball centred at x_i having radius $|U_i|$. Since $x_i \in U_i$ we have $U_i \subset B_i$ and hence $(U_i \cap C) \subset (B_i \cap C)$. Furthermore, since $|U_i| < \delta$ it follows that $H^d(B_i \cap C) < c|U_i|^d$ by definition of $x_i \in C''$, so

$$H^d(C'') \leq \sum H^d(B_i \cap C) < c \sum_{i=1}^{M} |U_i|^d \tag{3.162}$$

Since $\{U_i\}$ is *any* δ–cover of C'', taking the infimum of the RHS over all such covers we get

$$H^d(C'') \leq c H^d_\delta(C') \leq c H^d(C'') \tag{3.163}$$

and since $c < 1$, this can only hold if $H^d(C'') = 0$ as required. □

3.9 Near neighbour distances for chaotic attractors

In the study of chaotic dynamical systems we often encounter sets of points that are confined to regions of $\mathbb{R}^m$ having non–integral dimension.

3.9.1 Dynamical systems

A (discrete–time) dynamical system is defined by some subspace $X \subset \mathbb{R}^m$, called the *state space* of the system along with a function $f : X \to X$. Given an initial state $x_0 \in X$, the time evolution of the system is then described by the iterative equation

$$x_i = f(x_{i-1}) = f^i(x_0) \qquad i \in \mathbb{N} \tag{3.164}$$

where x_i represents the state of the system at time i. For various initial points x_0 we are interested in the behaviour of the sequences $\{x_i\}_{i=0}^{\infty}$ as i increases. Such sequences are called *trajectories* or *orbits* of the dynamical system.

A subset $A \subset X$ is said to be *invariant* if $f(A) \subset A$. Invariant sets have the property that whenever $x_0 \in A$, the subsequent trajectory of the system remains confined to A for all $i \in \mathbb{N}$. A (minimal) closed invariant set $C \subset X$ is said to be an *attractor* of the dynamical system if there exists some open neighbourhood V of C such that

- For any $x_0 \in V$, the distance from x_i to C converges to zero as $i \to \infty$.
- For some $x_0 \in V$, the closure of $\{x_i\}_0^{\infty}$ contains C.

The set V is called the *basin of attraction* of C and for any initial point $x_0 \in V$, once transient effects have diminished the trajectory of the system settles near one of its attractors. If an attractor C has a fractal structure or exhibits sensitive dependence on initial conditions (where nearby points in C diverge under iterates of f) then the dynamics are said to be *chaotic*. Such attractors often have non–integral dimension and are commonly associated with *dissipative* dynamical systems, where volumes in state space are contracted by time evolution.

3.9.2 The Hénon Map

The *Hénon map* $f : \mathbb{R}^2 \to \mathbb{R}^2$ ([Hénon 1976]) with parameters $a, b \in \mathbb{R}$ is defined by

$$f(x, y) = (a + by - x^2,\ x) \tag{3.165}$$

and the time evolution of the associated dynamical system can be represented by

$$\begin{aligned} x_{i+1} &= a + by_i - x_i^2 \\ y_{i+1} &= x_i \end{aligned} \tag{3.166}$$

where (x_i, y_i) is the state of the system at time i. For $a = 1.4$ and $b = 0.3$ the dynamics of the Hénon map are known to be chaotic, and its attractor is a fractal set. This will serve as a test case for our investigations into the behaviour of certain statistics associated with the Gamma test.

3.9.3 Probability measures on attractors

Let $\{x_1, \ldots, x_M\}$ be a time series generated by a dynamical system $f : X \to X$ where $X \subset \mathbb{R}^m$ and suppose for simplicity that f has a single attractor, denoted by C. Then C provides a global picture of the long–term behaviour of the system and loosely corresponds to the 'sample space' from which the points x_i are selected. The *distribution* of the points x_i on C is determined by the dynamical system itself, along with the initial state x_0.

In order to estimate the expected value of the distance between nearest neighbours in the set $\{x_1, \ldots, x_M\}$ we need to define a *probability measure* on C. As we have seen, the Hausdorff measure on C is defined in terms of its geometric properties and thus corresponds to a uniform density over the set. In general however, a typical trajectory of a chaotic dynamical system is likely to visit some regions of its attractor more frequently than others, in which case the points of the time series will *not* be uniformly distributed over the attractor. It therefore becomes necessary to specify a measure on C which takes such variations into account.

Invariant measures

Firstly, any measure defined on C must be *invariant* under time evolution, i.e. for every subset A of C we have that $\mu\left(f^{-1}(A)\right) = \mu(A)$ where $f^{-1}(A) = \{x \in \mathbb{R}^m \mid f(x) \in A\}$. It is possible to construct many invariant measures on C, not all of which are particularly useful. For example, the measure concentrated on an unstable fixed point satisfies $\mu\left(f^{-1}(A)\right) = \mu(A)$, but tells us nothing about the general time evolution of the system.

To describe the distribution of iterates on an attractor C, we need an invariant probability measure which describes how frequently various parts of C are visited by a 'typical' trajectory. In view of this, the natural invariant measure on C, called the *residence time measure* (see [Falconer 1990, page 263]), is defined by the time average

$$\mu(A) = \lim_{M\to\infty} \frac{1}{M} \sum_{i=0}^{M} I_A(x_i) \tag{3.167}$$

where x_0 is a 'typical' initial state and I_A is the indicator function for the set A. Clearly, μ is invariant since $x_i = f^i(x_0) \in A$ if and only if $x_{i-1} = f^{i-1}(x_0) \in f^{-1}(A)$. Thus $\mu(A)$ represents the proportion of iterates which fall in A and μ is concentrated on the set of points to which the trajectory $f^i(x_0)$ comes arbitrarily close infinitely often (so μ is supported by the attractor of f). As it stands however, the definition of μ in (3.167) depends on the initial state x_0.

Remark: The "ergodic average" (3.167) need not converge for arbitrary measurable sets A – we usually need that the boundary of A has measure zero.

The ergodic theorem

An invariant probability measure μ may be decomposable into several different components, each of which are again invariant. If not then μ is said to be *ergodic*. The following theorem (see [Falconer 1990, page 191]) asserts the existence of an ergodic measure on the attractor of any dynamical system.

Theorem 3.3. *If the compact set C is invariant under the dynamical system f^n then there exists an invariant ergodic measure μ with support contained in C*

The fundamental property of ergodic measures is that a time average is equal to a space average weighted by the ergodic measure.

Theorem 3.4 (The Ergodic Theorem). *If μ is ergodic then for any continuous function g,*

$$\lim_{M\to\infty}\frac{1}{M}\sum_{i=0}^{M} g(x_i) = \int g(x)\,d\mu(x) \tag{3.168}$$

for almost all initial states x_0 with respect to μ.

In particular, taking $g = I_A$ we have that for μ–almost all initial states x_0,

$$\lim_{M\to\infty}\frac{1}{M}\sum_{i=0}^{M} I_A(x_i) = \int_A d\mu(x) \tag{3.169}$$

As with (3.167), to ensure that (3.169) converges we usually need that the boundary of A has measure zero. Further details regarding ergodic theory can be found in the original paper of [Birkhoff 1927] and also [Ott 1993].

3.9.4 Near neighbour distances on chaotic attractors

Theorems 3.3 and 3.4 ensure the existence of invariant measures defined by time averages for μ–almost all initial conditions x_0 where μ is defined in (3.167). In view of this we may define the probability measure of the ball $B_x(r)$ in $\mathbb{R}^m$ by

$$\omega_x(r) = \lim_{M\to\infty}\frac{1}{M}\sum_{i=0}^{M} I_{B_x(r)}(x_i) \tag{3.170}$$

By the Poincaré Recurrence Theorem [Falconer 1990], if x is a point on the attractor then the trajectory $f^i(x_0)$ comes arbitrarily close to x infinitely often. This suggests that perhaps $\omega_x(r) > 0$ for small $r > 0$, provided ω_x is defined carefully.

We propose the following analogue of Conjecture 3.1.

Conjecture 3.2. *Let $k \geq 1$ be a fixed integer and let $\{x_1, \ldots, x_M\}$ be a time series representing a trajectory of a dynamical system f having attractor $C \subset \mathbb{R}^m$. Then if $x \in C$ satisfies condition **C.4** with $\omega_x(r)$ as defined in (3.170),*

$$\mathcal{E}(d^{\alpha}_{M,k}(x)) \sim \frac{c}{M^{\alpha/d}} \quad as \quad M \to \infty \tag{3.171}$$

where $c = c(\alpha, k, d, f)$ and d can be interpreted as the fractal dimension of C.

Remark: If this conjecture holds, it may provide a useful method of estimating the fractal dimension of a chaotic attractor.

Table 3.1: Scaling behaviour of $\delta_M(k)$ as M increases.

	$\delta_M(k) \approx cM^{-\beta}$					
	Uniform distribution				Hénon Map	
	Experimental		Theoretical		Experimental	
k	c	β	c	β	c	β
1	0.32159	1.00087	0.318309	1.0	7.06533	1.59765
2	0.68335	1.00629	0.636619	1.0	24.6000	1.62701
3	1.03724	1.00711	0.954929	1.0	51.4725	1.64240
4	1.40184	1.00817	1.273239	1.0	85.8252	1.65035
5	1.77714	1.00935	1.591549	1.0	129.529	1.65781
6	2.14373	1.00970	1.909859	1.0	178.568	1.66184
7	2.52942	1.01069	2.228169	1.0	235.310	1.66521
8	2.91086	1.01125	2.546479	1.0	295.118	1.66679
9	3.30097	1.01193	2.864788	1.0	351.112	1.66556
10	3.68841	1.01239	3.183098	1.0	416.175	1.66573

3.10 Experimental results

We examine the rate at which the mean squared kth nearest neighbour distance $\delta_M(k)$ converges to zero as the number of points M increases.

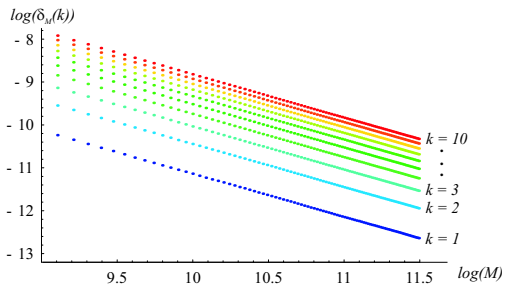

Figure 3.7: Graph of $\log(\delta_M(k))$ against $\log(M)$ for the uniform distribution.

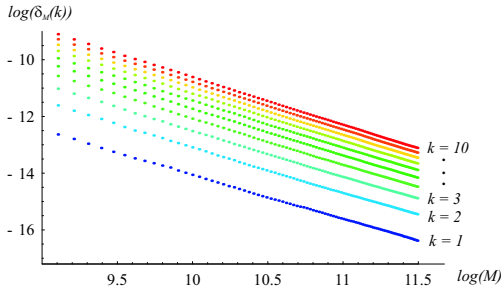

Figure 3.8: Graph of $\log(\delta_M(k))$ against $\log(M)$ for the Hénon map.

For each k in the range $1 \leq k \leq 10$ we compute the pairs $(\log(M), \log(\delta_M(k)))$ as M increases from 1000 to 100000 in steps of 1000. Figures 3.7 and 3.8 show plots of $\log(\delta_M(k))$ against $\log(M)$ for the uniform distribution and the Hénon map (3.165) with $a = 1.4$ and $b = 0.3$ respectively. Performing linear regression on these points we obtain the results shown in Table 3.1.

The theoretical values for the uniform case are provided by Theorem 3.2 and while the results are encouraging we should remark the errors in estimating the constant (which must be exponentiated from the least squares fit) are very high.

Note that given Conjecture 3.2, the scaling exponent of approximately 1.6 for the Hénon map in Table 3.1 would suggest a fractal dimension of approximately 1.25. This

is close to the published estimates which tend to be around 1.26 [Falconer 1990, page 180].

3.11 Summary

We have estimated the asymptotic behaviour of the moments of the kth nearest neighbour distance distribution under quite general hypotheses. Although a result of this type for the first moment of the distribution under uniform sampling has previously appeared, this result was subject to the assumption of periodic boundary conditions. Using the novel technique of boundary shrinking suggested by W. M. Schmidt we have removed the assumption of periodic boundary conditions and generalised the result to all moments and arbitrary smooth positive sampling densities.

This raises the possibility of using an inversion technique to construct an asymptotic form for the actual kth nearest neighbour distance distribution on M points. This involves technically difficult issues which we shall not pursue here.

We conjecture that such asymptotic results may be true under even more general conditions, where the sampling is driven by an ergodic process over a chaotic attractor of zero Lebesgue measure and positive Hausdorff dimension in $\mathbb{R}^m$. The further pursuit of these questions would raise some rather difficult issues and would somewhat divert us from the immediate goal of proving the Gamma test. Still, the fact remains that if such asymptotic results could be established they might be helpful in extending the provable range of applicability of the Gamma test and may also provide a route to an efficient $O(M \log M)$ algorithm for estimating Hausdorff dimension.

Chapter 4

Near neighbour geometry

4.1 Introduction

The primary objective of this chapter is to establish certain geometric results that assist us in dealing with the various sums of dependent random variables identified in Chapter 2. In pursuing this goal we find ourselves studying the directed graphs obtained by joining each point in a set of points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\} \subset \mathbb{R}^m$ to its kth nearest neighbour in the set – we call these *kth near neighbour graphs*. For our purposes we are mainly interested in *random* kth nearest neighbour graphs where the points are chosen from $\mathbb{R}^m$ according to some sampling distribution Φ. Whether random or not, these graphs are of considerable interest in their own right.

The result from this chapter required for the proof of the Gamma test is Theorem 4.1 which gives an explicit bound for the number of points that can share a common kth nearest neighbour. This result may also be of interest in coding theory.

Having spent some time in Chapter 3 developing techniques for dealing with boundary effects when computing the expected distance between kth nearest neighbours in a set of M points it seems appropriate to further illustrate the power of these methods by addressing the rather different but related question of determining the expected number of connected components in a random first nearest neighbour graph. This we do in section 4.5 where we are able to establish a precise asymptotic formula for the number of components in the case of a uniform sampling distribution. These results are also confirmed experimentally.

4.2 Nearest neighbour graphs

The *kth nearest neighbour graph* associated with a set of points $\{x_1, \ldots, x_M\} \subset \mathbb{R}^m$ is defined to be the directed graph $G = G(V, E)$ where the vertex set V is the set of points itself and the edge set E contains the directed edge (x_i, x_j) if and only if x_j is

the kth nearest neighbour of x_i. Note that if x_j is the kth nearest neighbour of x_i, this does not necessarily imply that x_i is the kth nearest neighbour of x_j.

In Figure 4.1 we plot the first nearest neighbour graph for a set of 500 points selected uniformly at random from the unit square $[0,1]^2$.

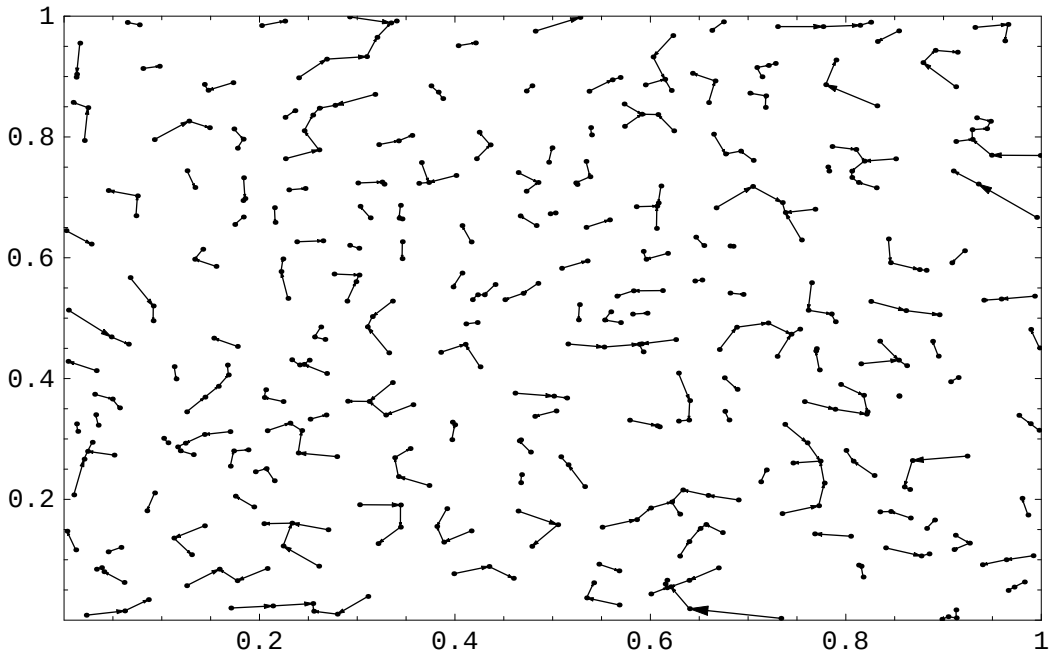

Figure 4.1: The first nearest neighbour graph for 500 uniformly distributed points in $[0,1]^2$ (*Courtesy of Dr A.P.M. Tsui*)

Figure 4.2 shows a single connected 'component' of a first nearest neighbour graph, where the word 'component' refers to the underlying undirected graph and $A \to B$ means that 'B is a nearest neighbour of A'.

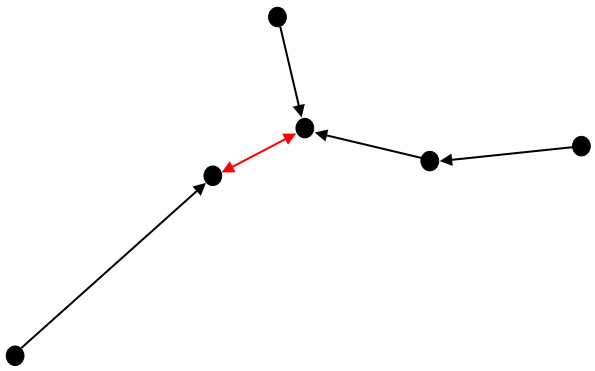

Figure 4.2: A connected component of a first nearest neighbour graph.

Using geometric methods we show that the maximum *in–degree* of any vertex in the kth nearest neighbour graph (subsequently called simply the maximum *degree* of a

vertex) is bounded above by some constant that is independent of the total number of points M. This result will be used in subsequent chapters to control the dependence between certain random variables associated with the Gamma test.

4.3 The maximum vertex in–degree

The following result [Bickel and Breiman 1983] shows that the maximum degree of any vertex in a first nearest neighbour graph remains absolutely bounded as the number of points increases to infinity.

Lemma 4.1. *For any set of distinct points $x_1, \ldots, x_M$ in $\mathbb{R}^m$, any point x_i can be the nearest neighbour of at most some finite number $N(m)$ other points, where $N(m)$ is independent of the number of points M.*

Proof. Let $S(x_i)$ denote the surface of the unit sphere in $\mathbb{R}^m$ centred at x_i. Since $S(x_i)$ is compact there exist disjoint sets $S_1, \ldots, S_N$ with

$$S(x_i) = \bigcup_{j=1}^{N(m)} S_j \tag{4.1}$$

such that for every $1 \leq j \leq N(m)$,

$$|a - b| < 1 \qquad \text{for all} \qquad a, b \in S_j \tag{4.2}$$

From this partition we define a set of disjoint cones

$$C_j = \{x_i + \lambda a \,|\, a \in S_j, \lambda > 0\} \qquad 1 \leq j \leq N(m) \tag{4.3}$$

having x_i as their common peak.

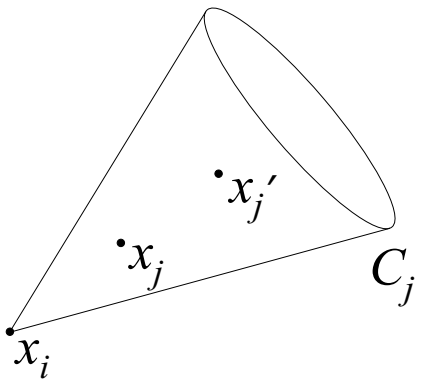

Figure 4.3: If x_j is the nearest point to x_i in the cone C_j, then any other point x_j' in C_j must be closer to x_j than it is to x_i.

To prove the lemma we show that *at most* one point from each C_j can have x_i as its nearest neighbour. Let x_j be a closest point to x_i in C_j and let x_j' be another point in

C_j. We claim that x'_j is closer to x_j than it is to x_i and cannot therefore have x_i as its nearest neighbour. To this end let

$$\begin{array}{rclcl} x_j - x_i & = & \lambda a & \text{for some} \quad a \in S_j & \text{and} \quad \lambda > 0 \\ x'_j - x_i & = & \mu b & \text{for some} \quad b \in S_j & \text{and} \quad \mu > 0 \end{array} \tag{4.4}$$

where $0 < \lambda \leq \mu$. Then $|x'_j - x_j| = |\mu b - \lambda a|$ which we write as

$$|x'_j - x_j| = \mu \left| \left(1 - \frac{\lambda}{\mu}\right) b - \frac{\lambda}{\mu}(a - b) \right| \tag{4.5}$$

from which it follows that

$$|x'_j - x_j| \leq \mu \left(\left(1 - \frac{\lambda}{\mu}\right) |b| + \frac{\lambda}{\mu}|a - b| \right) \tag{4.6}$$

Since $a, b \in S_j$ it follows that $|b| = 1$ and $|a - b| < 1$ so

$$|x'_j - x_j| < \mu \left(\left(1 - \frac{\lambda}{\mu}\right) + \frac{\lambda}{\mu} \right) = \mu = |x'_j - x_j| \tag{4.7}$$

Hence x'_j is closer to x_j than it is to x_i and cannot therefore have x_i as its nearest neighbour. Thus there can be at most one point in each cone C_j having x_i as its nearest neighbour, as required. □

While Lemma 4.1 is a step in the right direction, it does not specify how large the maximum vertex degree $N(m)$ is likely to be in any given dimension m. A quantification of this maximum vertex degree may obtained using the notion of *kissing numbers*.

4.4 Nearest neighbours and the maximum kissing number in $\mathbb{R}^m$

A *sphere packing* in any m–dimensional space is a collection of disjoint open spheres of unit radius, and the *kissing number* of any sphere S is the number of open unit spheres in the packing that share a common tangent with S. In a lattice packing, each sphere has the same kissing number. The maximum kissing number in $\mathbb{R}^m$, denoted by $K(m)$, is the largest kissing number that can be attained by *any* packing of m–dimensional spheres. $K(m)$ is known exactly in only a few dimensions: $K(1) = 2$, $K(2) = 6$, $K(3) = 12$, $K(8) = 240$ and $K(24) = 196560$.

The following lemma gives upper and lower bounds for the maximum kissing number in $\mathbb{R}^m$, found in [Kabatiansky and Levenshtein 1978] and [Wyner 1965] respectively.

Lemma 4.2. *The maximum kissing number $K(m)$ in $\mathbb{R}^m$ satisfies*

$$2^{0.2075m(1+o(1))} \leq K(m) \leq 2^{0.401m(1+o(1))} \tag{4.8}$$

The following lemma [Zeger and Gersho 1994] shows that the degree of a vertex in any *first* nearest neighbour graph in $\mathbb{R}^m$ is equal to at most the maximum kissing number in $\mathbb{R}^m$.

Lemma 4.3 (Zeger and Gersho). *The maximum number of points in $\mathbb{R}^m$ that can share a common nearest neighbour is equal to the maximum kissing number in $\mathbb{R}^m$.*

Proof. Suppose we have a set of $N > K(m)$ distinct points $\{x_1, \ldots, x_N\}$ having a common nearest neighbour c. Let $d_i = |x_i - c|$, $d = \min\{d_i : 1 \leq i \leq N\}$ and define

$$y_i = \left(\frac{d_i - d}{d_i}\right) c + \left(\frac{d}{d_i}\right) x_i \tag{4.9}$$

so that each point y_i lies on the sphere of radius d centred at c (see Figure 4.4). We

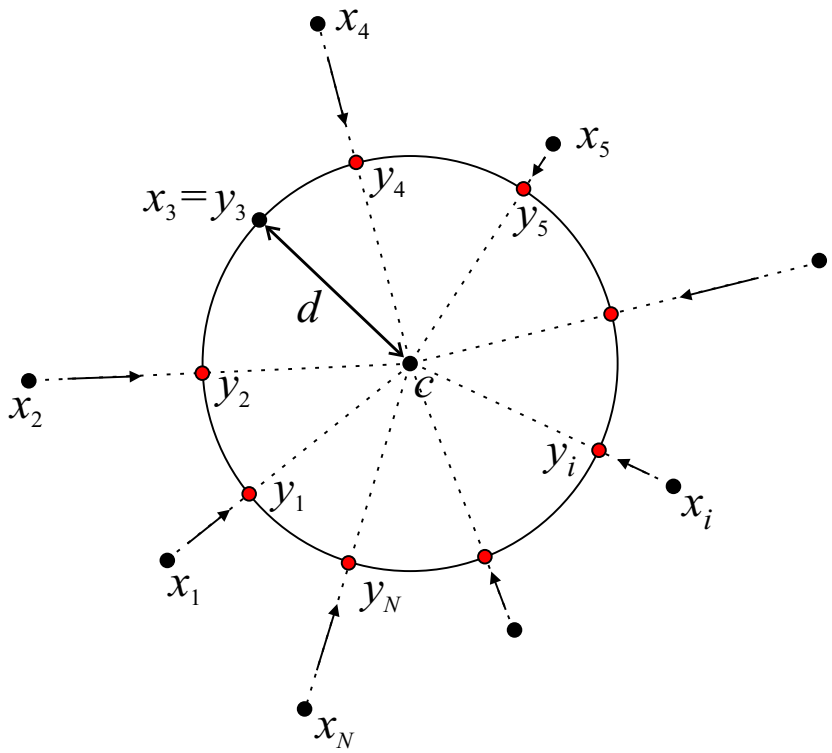

Figure 4.4: The vectors $x_i - c$ are scaled so that they all have the same length d.

claim that $|y_i - y_j| \geq d$ for each $i \neq j$. Let $i \neq j$ and suppose without loss of generality that

$$|x_i - c| \leq |x_j - c| \leq |x_i - x_j| \tag{4.10}$$

Using the identity

$$2(x_i - c) \cdot (x_j - c) = |x_i - c|^2 + |x_j - c|^2 - |x_i - x_j|^2 \tag{4.11}$$

we get

$$2(x_i - c) \cdot (x_j - c) \leq d_i^2 \leq d_i d_j \tag{4.12}$$

where the second inequality follows by (4.10). By definition of y_i and y_j,

$$2(y_i - c) \cdot (y_j - c) = \frac{2d^2}{d_i d_j}(x_i - c) \cdot (x_j - c) \tag{4.13}$$

so by (4.12) we have that $2(y_i - c) \cdot (y_j - c) \leq d^2$. Hence, using the identity

$$|y_i - y_j|^2 = |y_i - c|^2 + |y_j - c|^2 - 2(y_i - c) \cdot (y_j - c) \tag{4.14}$$

it follows that $|y_i - y_j|^2 \geq d^2$, which proves the claim.

The set $\{y_1, \ldots, y_N\}$ is therefore a set of $N > K(m)$ distinct points, all of which are located at a distance d from c and which are all at least a distance d apart. Thus we can place a set of N non–overlapping spheres of radius $d/2$ centred at each y_i and each of these will be tangent to the sphere of radius $d/2$ centred at c, contradicting the fact that the sphere of radius $d/2$ centred at c can have *at most* $K(m)$ such tangent spheres. Hence $N \leq K(m)$ as required. □

We now extend Lemma 4.3 to obtain an upper bound on the maximum number of points in $\mathbb{R}^m$ that can share a common kth nearest neighbour and in this way we get an upper bound on the maximum degree of a vertex in any kth nearest neighbour graph in $\mathbb{R}^m$.

Theorem 4.1. *The maximum number of points $F(m, k)$ in $\mathbb{R}^m$ that can share a common kth nearest neighbour satisfies*

$$F(m, k) \leq kK(m) \tag{4.15}$$

where $K(m)$ is the maximum kissing number in $\mathbb{R}^m$.

Proof. Suppose that $F(m, k) > kK(m)$ and let $S = \{x_1, \ldots, x_t\}$ be a set of $t > kK(m)$ points in $\mathbb{R}^m$ having a common kth nearest neighbour c.

First we choose α_1 to be a point of S which is furthest away from c,

$$|\alpha_1 - c| \geq |x_i - c| \quad \text{for all} \quad x_i \in S \tag{4.16}$$

and define $S_1 \subset S$ to be the set of points in S that are strictly closer to α_1 than c is to α_1, i.e.

$$S_1 = \{x_j \in S : |\alpha_1 - x_j| < |\alpha_1 - c|\} \tag{4.17}$$

Then S_1 contains at least one point (namely α_1 itself) and since c is the kth nearest neighbour of α_1 it follows by definition of kth nearest neighbours (Section 1.7) that S_1 contains at most k points. Hence, the cardinality of S_1 satisfies $1 \leq |S_1| \leq k$.

Next we eliminate the points of S_1 and choose $\alpha_2 \in T_1 = S \setminus S_1$ among the remaining points to be a furthest point away from c, as illustrated in Figure 4.5. We then define

$$S_2 = \{x_j \in S \setminus T_1 : |\alpha_2 - x_j| < |\alpha_2 - c|\} \tag{4.18}$$

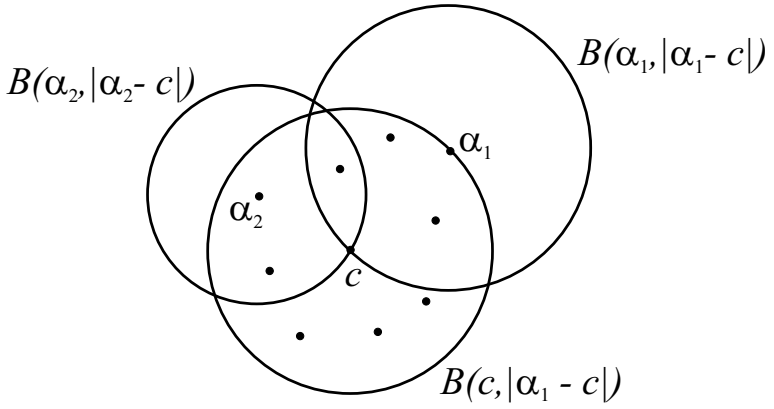

Figure 4.5: α_2 is the furthest point away from c that is not contained in the open ball $B(\alpha_1, |\alpha_1 - c|)$.

whose cardinality again satisfies $1 \leq |S_2| \leq k$. Since $\alpha_2 \in T_1$ it follows that $\alpha_1 \neq \alpha_2$ and $|\alpha_1 - c| \leq |\alpha_1 - \alpha_2|$. By construction we also have $|\alpha_2 - c| \leq |\alpha_1 - c|$ and hence

$$|\alpha_2 - c| \leq |\alpha_1 - c| \leq |\alpha_1 - \alpha_2| \tag{4.19}$$

We continue this process, after n steps obtaining the sequence of disjoint sets $S_1, \ldots, S_n$ and a corresponding set of distinct points $\{\alpha_1, \ldots, \alpha_n\}$. If $T_n = S \setminus \cup_{i=1}^{n} S_i$ is empty, the process terminates. Otherwise we choose $\alpha_{n+1} \in T_n$ to be a point which is furthest away from c and define

$$S_{n+1} = \{x_j \in S \setminus T_n : |\alpha_{n+1} - x_j| < |\alpha_{n+1} - c|\} \tag{4.20}$$

so that $1 \leq |S_{n+1}| \leq k$ as above. Since $\alpha_{n+1} \in T_n$ then α_{n+1} is distinct from each point in the set $\{\alpha_1, \ldots, \alpha_n\}$ and furthermore, $|\alpha_i - c| \leq |\alpha_i - \alpha_{n+1}|$ for each $1 \leq i \leq n$. By construction we also have that $|\alpha_{n+1} - c| \leq |\alpha_i - c|$ for each $1 \leq i \leq n$ and hence

$$|\alpha_{n+1} - c| \leq |\alpha_{n+1} - \alpha_i| \quad \text{for each} \quad 1 \leq i \leq n \tag{4.21}$$

The condition $|S_n| \geq 1$ means that at least one point is eliminated at each stage and hence the process must eventually terminate, say after N steps where $N \leq t$. Furthermore, since $|S_n| \leq k$ for each n, *at most* k points are eliminated at each stage so it also follows that $N \geq t/k$.

By hypothesis we have assumed that $t > kK(m)$ so it follows that $N > K(m)$. Thus we have a set of $N > K(m)$ distinct points $\{\alpha_1, \ldots, \alpha_n\}$ which by construction, have the property that $|\alpha_i - c| \leq |\alpha_i - \alpha_j|$ for each $1 \leq i \neq j \leq n$.

Hence the point c is the first nearest neighbour of the $N > K(m)$ points α_i. This contradicts Lemma 4.3 which asserts that *at most* $K(m)$ distinct points have this property and we conclude that $F(m, k) \leq kK(m)$ as required. □

Theorem 4.1 is all we need in this regard for the proof of the Gamma test in Chapter 7.

4.5 Expected number of components in a first nearest neighbour graph

We derive an expression for the expected number of connected components[1] in the first near neighbour graph of a set of points selected uniformly at random from a compact convex body in $\mathbb{R}^m$. Our result is based on the observation, found in [Eppstein *et al.* 1997], that every component of a first nearest neighbour graph contains *exactly one* pair of vertices that are nearest neighbours of each other. For completeness we include a proof of this observation.

Lemma 4.4. *Every connected component of a first nearest neighbour graph has exactly one pair of vertices that are nearest neighbours of each other.*

Proof. Let G be a (directed) first nearest neighbour graph having vertices $X_1, \ldots, X_M$. Let H be a (simply) connected component of G. We show that H contains exactly one cycle and that the length of this cycle is equal to 2.

Suppose that H has N vertices ($N \leq M$). By definition, each X_i has exactly one nearest neighbour so H contains exactly N arcs. Since H is (simply) connected, $N-1$ of these arcs must be such that they form a spanning tree for the underlying (undirected) graph of H. Thus follows that H contains exactly one directed cycle. The cycle is directed because each vertex of H has exactly one arc directed away from it (i.e. towards its unique nearest neighbour). Furthermore, since a vertex cannot be its own nearest neighbour the length of the cycle must be greater than 1.

Suppose that the length of the cycle is equal to $n \geq 3$. Letting σ denote an appropriate permutation of the index list $1, \ldots, M$ we represent the cycle by

$$X_{\sigma(1)} \,\text{NN}\, X_{\sigma(2)} \,\text{NN}\, \ldots \,\text{NN}\, X_{\sigma(n)} \,\text{NN}\, X_{\sigma(1)} \tag{4.22}$$

where the symbol 'NN' represents the 'is the nearest neighbour of' relation (not that this relation is not necessarily symmetric). If we define $\sigma(n+1) \equiv \sigma(1)$, this can be summarised by saying that $X_{\sigma(i)}$ is the nearest neighbour of $X_{\sigma(i+1)}$ for each $1 \leq i \leq n$. Clearly,

$$\begin{array}{rcll} |X_{\sigma(2)} - X_{\sigma(1)}| & \leq & |X_{\sigma(2)} - X_{\sigma(3)}| & \text{since} \quad X_{\sigma(1)} \,\text{NN}\, X_{\sigma(2)} \\ |X_{\sigma(3)} - X_{\sigma(2)}| & \leq & |X_{\sigma(3)} - X_{\sigma(4)}| & \text{since} \quad X_{\sigma(2)} \,\text{NN}\, X_{\sigma(3)} \\ \vdots & \vdots & \vdots & \vdots \\ |X_{\sigma(1)} - X_{\sigma(n)}| & \leq & |X_{\sigma(1)} - X_{\sigma(2)}| & \text{since} \quad X_{\sigma(n)} \,\text{NN}\, X_{\sigma(1)} \end{array}$$

and hence

$$|X_{\sigma(2)} - X_{\sigma(1)}| = |X_{\sigma(3)} - X_{\sigma(2)}| = \ldots = |X_{\sigma(n)} - X_{\sigma(n-1)}| = |X_{\sigma(1)} - X_{\sigma(n)}| \tag{4.23}$$

Thus each vertex $X_{\sigma(i)}$ is equidistant from its nearest neighbour $X_{\sigma(i-1)}$ and the point $X_{\sigma(i+1)}$ having $X_{\sigma(i)}$ as its nearest neighbour, i.e.

$$|X_{\sigma(i)} - X_{\sigma(i-1)}| = |X_{\sigma(i)} - X_{\sigma(i+1)}| \quad \text{for all} \quad 1 \leq i \leq n \tag{4.24}$$

[1] The components in question are those of the underlying undirected graph.

Let $\sigma(j) = \min\{\sigma(i) \,\big|\, 1 \leq i \leq n\}$ and consider the nearest neighbour $X_{\sigma(j-1)}$ of $X_{\sigma(j)}$. As we have seen, the distance between the nearest neighbour $X_{\sigma(j-2)}$ of $X_{\sigma(j-1)}$ is equal to the distance between $X_{\sigma(j)}$ and $X_{\sigma(j-1)}$. However, by definition the index $\sigma(j)$ is (strictly) less than the index $\sigma(j-2)$, contradicting the fact that $X_{\sigma(j-2)}$ is the nearest neighbour of $X_{\sigma(j-1)}$. Thus the cycle must be of length 2, as required □

Remark: That Lemma 4.4 does not hold for general kth nearest neighbour graphs (i.e. $k > 1$) is illustrated in Figure 4.6, where each arrow points to the second nearest neighbour of its starting point.

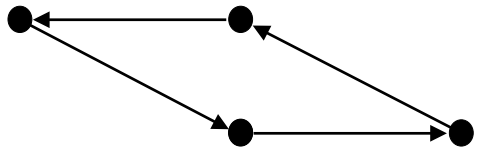

Figure 4.6: A connected component in a second nearest neighbour graph

4.5.1 Preliminaries

Let $X'_i = X_{N[i,1]}$ denote the (first) nearest neighbour of X_i in the random sample $X = (X_1, \ldots, X_M)$. By Lemma 4.4, we need to determine the conditional probability that X_i is the nearest neighbour of X'_i, given that X'_i is the nearest neighbour of X_i.

Remark: For a set of points selected from $\mathbb{R}^m$ according to any well behaved sampling distribution, an explicit expression for the probability that a point is the jth nearest neighbour of its own kth nearest neighbour appears in [Henze 1987]. Related results for point processes can be found in [Henze 1986], [Pickard 1982] and [Cox 1981]. The work of [Henze 1987] is considerably more general than that presented below – we consider only the case $j = k = 1$ for a uniform distribution. We do however give asymptotic results with error terms as $M \to \infty$.

Let $N = N(X)$ be the number of components in the first nearest neighbour graph of the random sample X and let $X_i \,\text{NN}\, X'_i$ represent the event that X_i is the nearest neighbour of X'_i. Then

$$\mathcal{E}(N) = \frac{1}{2} M \, \mathbf{P}(X_i \,\text{NN}\, X'_i) \tag{4.25}$$

For every $x \in C$ let

$$\sigma(x) = \mathbf{P}(X_i \,\text{NN}\, X'_i \,\big|\, X_i = x) \tag{4.26}$$

be the conditional probability that X_i is the nearest neighbour of X'_i given that X_i takes the value x, so that

$$\mathbf{P}(X_i \,\text{NN}\, X'_i) = \int_{x \in C} \sigma(x) \phi(x) \, dx \tag{4.27}$$

For every $x \in C$ and for all $0 \leq r \leq c_1$ let

$$u_x(r) = \mathbf{P}(X_i \,\text{NN}\, X_i' \,\big|\, X_i = x, |X_i' - X_i| = r) \tag{4.28}$$

be the conditional probability that X_i is the nearest neighbour of X_i' given that X_i takes the value x and that the distance from X_i to its nearest neighbour is equal to r.

As in (3.24) of Chapter 3 we define

$$q_x(r) = \mathbf{P}(|X_i' - X_i| \leq r \,\big|\, X_i = x) \tag{4.29}$$

to be the conditional distribution function of the first nearest neighbour distance of X_i given that X_i takes the value x, so that

$$\sigma(x) = \int_0^{c_1} u_x(r)\, dq_x(r) \tag{4.30}$$

By Lemma 3.3,

$$\sigma(x) = (M-1) \int_0^{c_1} u_x(r)(1 - \omega_x(r))^{M-2}\, d\omega_x(r) \tag{4.31}$$

We seek to express $u_x(r)$ in terms of $\omega_x(r)$ so that the integral may be evaluated by changing the variable of integration from r to $\omega = \omega_x(r)$.

4.5.2 Boundary effects

If X_i' takes a value near the boundary of C then the probability that X_i is the nearest neighbour of X_i is likely to be *greater* than it would be if X_i and X_i' were both located away from the boundary. Under certain circumstances it may be that this probability is equal to one (i.e. that X_i is certain to be the nearest neighbour of X_i') so the only thing we can say for sure regarding $\sigma(x)$ is that it satisfies $|\sigma(x)| \leq 1$.

Let $0 < \delta < c_1$ and let $B \subset C$ denote the boundary region of width 2δ, defined to be the set of points in C that are within distance 2δ of the boundary. Let $A = C \setminus B$ be the corresponding interior region and write

$$\mathbf{P}(X_i \,\text{NN}\, X_i') = \int_{x \in A} \sigma(x)\phi(x)\, dx + \int_{x \in B} \sigma(x)\phi(x)\, dx \tag{4.32}$$

By Lemma 3.1 there exists some constant $0 < a_2 < \infty$ such that $|\phi(x)| < a_2$ for all $x \in C$. Furthermore, by condition **C.3** there exists some constant $0 < c_3 < \infty$ such that $\mu(B) \leq c_3\delta$. Since $|\sigma(x)| < 1$ it thus follows that the second integral in (4.32) is bounded above by $a_2 c_2 \delta$ and hence

$$\mathbf{P}(X_i \,\text{NN}\, X_i') = \int_{x \in A} \sigma(x)\phi(x)\, dx + O(\delta) \quad \text{as} \quad M \to \infty \tag{4.33}$$

As in Lemma 3.9 we show that for every $x \in A$ and for suitable $\delta \to 0$ as $M \to \infty$, the error incurred by evaluating the integral in (4.30) over $[0, \delta]$ rather than $[0, c_1]$ becomes exponentially small as $M \to \infty$.

Lemma 4.5. *Let $0 < \epsilon < 1/m$ and $\delta = 1/M^{\epsilon}$. Then for all $x \in C$ and every $\beta > 0$,*

$$\sigma(x) = (M-1)\int_0^{\delta} u_x(r)(1-\omega_x(r))^{M-2}\, d\omega_x(r) + O\left(\frac{1}{M^{\beta}}\right) \quad as \quad M \to \infty \tag{4.34}$$

Proof. Let

$$I(\delta) = (M-1)\int_{\delta}^{c_1} u_x(r)(1-\omega_x(r))^{M-2}\, d\omega_x(r) \tag{4.35}$$

Since $\omega_x(r)$ is monotonic increasing in r and $\omega_x(r) \geq c_4 r^m$ for all $x \in C$ it follows that $1 - \omega_x(r) \leq 1 - c_4\delta^m$ for all $\delta \leq r \leq c_1$.

Since $0 \leq u_x(r) \leq 1$ and $\delta > 0$ we therefore have that

$$I(\delta) \leq c_1(M-1)(1-c_4\delta^m)^{M-2} \tag{4.36}$$

and substituting for δ we obtain

$$I_{\delta} \leq (M-1)\left(1 - \frac{c_2}{M^{m\epsilon}}\right)^{M-2} \tag{4.37}$$

Hence, since $0 < m\epsilon < 1$ we can apply Lemma 3.4 and the result follows. □

4.5.3 The probability $u_x(r)$

Let $x, y \in C$ and define

$$\xi(x,y) = \mathbf{P}(X_i \,\text{NN}\, X_i' \,\big|\, X_i = x, X_i' = y) \tag{4.38}$$

Let $X = (X_1, \ldots, X_M)$ be a random sample in which X_i takes the value x and X_i' takes the value y, and consider the ball $B_x(|x-y|)$ centred at x having y on its boundary. Since X_i' is the nearest neighbour of X_i it follows that $B_x(|x-y|)$ is empty, i.e. $B_x(|x-y|)$ contains no other point of the sample.

Now consider the ball $B_y(|x-y|)$ centred at y and having x on its boundary. As illustrated in Figure 4.7 the probability $\xi(x,y)$ that X_i is also the nearest neighbour of X_i' is equal to the conditional probability that $B_y(|x-y|)$ is empty given that the ball $B_x(|x-y|)$ is empty,

$$\xi(x,y) = \mathbf{P}\big(B_y(|x-y|) \text{ is empty} \,\big|\, B_x(|x-y|) \text{ is empty}\big) \tag{4.39}$$

Using the identity $\mathbf{P}(A \,|\, B) = \mathbf{P}(A \cap B)/\mathbf{P}(B)$ for conditional probabilities we obtain

$$\xi(x,y) = \frac{\mathbf{P}(B_x(|x-y|) \cup B_y(|x-y|) \text{ is empty})}{\mathbf{P}(B_x(|x-y|) \text{ is empty})} \tag{4.40}$$

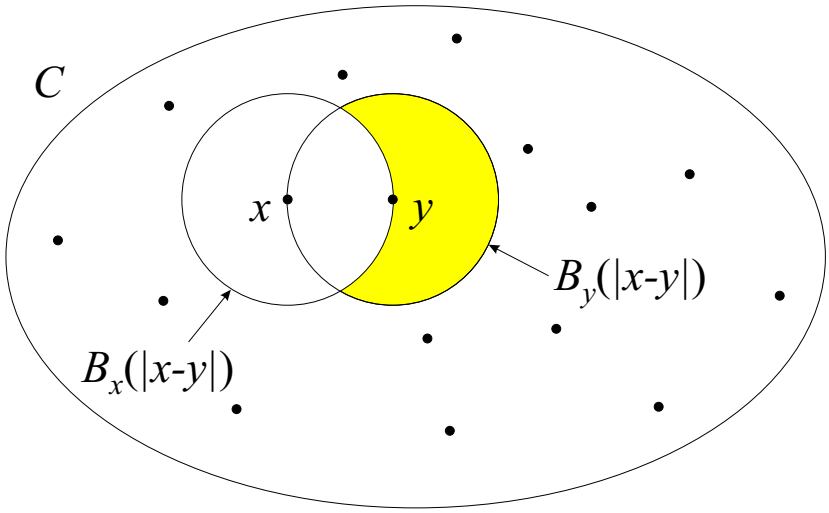

Figure 4.7: If y is the nearest neighbour of x, then x is the nearest neighbour of y if and only if the shaded region contains no other point of the sample.

The probability that $B_x(|x-y|)$ is empty is the probability that the remaining $M-2$ points of the random sample $X = (X_1, \ldots, X_M)$ fall outside $B_x(|x-y|)$.

Let $C_x(r)$ denote the boundary of the ball $B_x(r)$,

$$C_x(r) = \{y \in C \,\big|\, |x-y| = r\} \tag{4.41}$$

By (4.28) and (4.38) we see that $u_x(r)$ is the expected value of $\xi(x,y)$ over $C_x(r)$. Writing this as

$$u_x(r) = \mathcal{E}\big(\xi(x,y) \,\big|\, y \in C_x(r)\big) \tag{4.42}$$

we thus obtain

$$\sigma(x) = (M-1)\int_0^{c_1} \mathcal{E}\big(\xi(x,y) \,\big|\, y \in C_x(r)\big)(1-\omega_x(r))^{M-2}\, d\omega_x(r) \tag{4.43}$$

For every $y \in C_x(r)$ the probability that $B_x(|x-y|)$ is empty is the probability that $B_x(r)$ is empty. Hence the probability that X_j takes a value outside $B_x(r)$ is equal to $1-\omega_x(r)$ so

$$\mathbf{P}(B_x(|x-y|) \text{ is empty}) = (1-\omega_x(r))^{M-2} \quad \text{for all} \quad y \in C_x(r) \tag{4.44}$$

Let $\alpha(x,y)$ be the probability measure of $B_x(|x-y|) \cup B_y(|x-y|)$, given by

$$\alpha(x,y) = \int_{B_x(|x-y|)\cup B_y(|x-y|)} \phi(t)\, dt \tag{4.45}$$

so that

$$\mathbf{P}(B_x(|x-y|) \cup B_y(|x-y|) \text{ is empty}) = (1-\alpha(x,y))^{M-2} \tag{4.46}$$

By (4.39) and (4.44) we obtain

$$\xi(x,y) = \frac{(1-\alpha(x,y))^{M-2}}{(1-\omega_x(r))^{M-2}} \quad \text{for all} \quad y \in C_x(r) \tag{4.47}$$

Since $\omega_x(r)$ is constant over $C_x(r)$, by (4.42) we have that

$$u_x(r) = \frac{\mathcal{E}\big((1-\alpha(x,y))^{M-2} \,\big|\, y \in C_x(r)\big)}{(1-\omega_x(r))^{M-2}} \tag{4.48}$$

Hence by (4.43) it follows that

$$\sigma(x) = (M-1)\int_0^{c_1} v_x(r)\, d\omega_x(r) \tag{4.49}$$

where

$$v_x(r) = \mathcal{E}\big((1-\alpha(x,y))^{M-2} \,\big|\, y \in C_x(r)\big) \tag{4.50}$$

and thus we seek to express $v_x(r)$ in terms of $\omega_x(r)$.

4.5.4 The Lebesgue measure of a circle pair

The *circle pair* associated with a pair of points $x, y \in \mathbb{R}^m$ is defined to be the set $B_x(|x-y|) \cup B_y(|x-y|)$ – this is the union of the ball centred at x having y on its boundary and the ball centred at y having x on its boundary.

Let

$$\eta = \eta(m) = \frac{\mu\big(B_x(|x-y|) \cup B_y(|x-y|)\big)}{\mu(B_x(|x-y|))} \tag{4.51}$$

denote the Lebesgue measure of the circle pair $B_x(|x-y|) \cup B_y(|x-y|)$ expressed as a proportion of the Lebesgue measure of $B_x(|x-y|)$. Note that $1 < \eta < 2$ provided $x \neq y$. Simple geometric arguments lead to the following.

Lemma 4.6. *For every $x, y \in \mathbb{R}^m$,*

$$\mu(B_x(r) \cap B_y(r)) = 2V_{m-1}r^m \int_0^{\pi/3} \sin^m\theta\, d\theta \tag{4.52}$$

where V_{m-1} is the volume of the unit ball in $\mathbb{R}^{m-1}$.

The value of $\eta(m)$ is given by the following.

Lemma 4.7.

$$\eta = 2\left(1 - \frac{\Gamma(1+m/2)}{\sqrt{\pi}\Gamma(1/2+m/2)}\int_0^{\pi/3}\sin^m\theta\, d\theta\right) \tag{4.53}$$

Proof. Let $x, y \in \mathbb{R}^m$ and define $r = |x - y|$ so that

$$\eta = \frac{\mu(B_x(r) \cup B_y(r))}{\mu(B_x(r))} \tag{4.54}$$

Let V_m be the volume of the unit ball in $\mathbb{R}^m$ and consider

$$\mu(B_x(r) \cup B_y(r)) = \mu(B_x(r)) + \mu(B_y(r)) - \mu(B_x(r) \cap B_y(r)) \tag{4.55}$$

Clearly, $\mu(B_x(r)) = \mu(B_y(r)) = V_m r^m$ and by Lemma 4.6 we have that

$$\mu(B_x(r) \cap B_y(r)) = 2V_{m-1} r^m \int_0^{\pi/3} \sin^m \theta \, d\theta \tag{4.56}$$

Hence by (4.55) we obtain

$$\mu(B_x(r) \cup B_y(r)) = 2V_m r^m \left(1 - \frac{V_{m-1}}{V_m} \int_0^{\pi/3} \sin^m \theta \, d\theta\right) \tag{4.57}$$

and since $\mu(B_x(r)) = V_m r^m$, the result follows by (4.54) and the formula $V_m = \pi^{m/2}/\Gamma(1 + m/2)$. □

We omit the proof of the following elementary result which facilitates direct computation of η from Lemma 4.7.

Lemma 4.8. *If*

$$I_m = \int_0^{\pi/3} \sin^m \theta \, d\theta \tag{4.58}$$

then

$$I_m = \left(1 - \frac{1}{m}\right) I_{m-2} - \frac{1}{2m} \left(\frac{\sqrt{3}}{2}\right)^{m-1} \tag{4.59}$$

with $I_1 = 1/2$ *and* $I_2 = \pi/6 - \sqrt{3}/8$.

4.5.5 Theorem for uniform distributions

By Lemma 4.5 and (4.33) we can restrict our attention to the case where $x \in A$ and $0 \leq r \leq \delta$ in which case for every point $y \in C_x(r)$ the region $B_x(r) \cup B_y(r)$ is completely contained in C. If the points are selected from C according to a uniform distribution it thus follows that

$$\alpha(x, y) = \eta \omega_x(r) = \eta V_m r^m \tag{4.60}$$

Theorem 4.2. *Let $X = (X_1, \ldots, X_M)$ be a random sample of independent and uniformly distributed random variables X_i taking values in the set $C \subset \mathbb{R}^m$ where $\mu(C) = 1$ and C satisfies conditions **C.1**, **C.2** and **C.3** of Chapter 3. Let N be the number of connected components in the first nearest neighbour graph of X. Then for every $\rho > 0$,*

$$\mathcal{E}(N) = \frac{1}{2}M\left(\frac{1}{\eta} + O\left(\frac{1}{M^{1/m-\rho}}\right)\right) \quad as \quad M \to \infty \tag{4.61}$$

Proof. We need to show that

$$\mathbf{P}(X_i \,\text{NN}\, X_i') = \frac{1}{\eta} + O(\delta) \quad \text{as} \quad M \to \infty \tag{4.62}$$

where $\delta = 1/M^{1/m-\rho}$.

Let $A = C \setminus C(2\delta)$ be the set of points in C whose distance from the boundary is at least 2δ. By (4.33),

$$\mathbf{P}(X_i \,\text{NN}\, X_i') = \int_{x\in A} \sigma(x)\phi(x)\,dx + O(\delta) \quad \text{as} \quad M \to \infty \tag{4.63}$$

where

$$\sigma(x) = (M-1)\int_0^{c_1} v_x(r)\,d\omega_x(r) \tag{4.64}$$

and

$$v_x(r) = \mathcal{E}\big((1-\alpha(x,y))^{M-2} \,\big|\, y \in C_x(r)\big) \tag{4.65}$$

Let $x \in A$. Then for every $0 \le r \le \delta$ the circle pair $B_x(r) \cup B_y(r)$ is completely contained in C so by (4.60) it follows that

$$\alpha(x,y) = \eta\omega_x(r) \text{ for all } y \in C_x(r) \tag{4.66}$$

and since this is independent of any particular $y \in C_x(r)$ we have that

$$v_x(r) = (1-\eta\omega_x(r))^{M-2} \text{ for all } 0 \le r \le \delta \tag{4.67}$$

Substituting this into (4.30) we get

$$\sigma(x) = (M-1)\int_0^{c_1} (1-\eta\omega_x(r))^{M-2}\,d\omega_x(r) \tag{4.68}$$

and changing the variable of integration from $\omega_x(r)$ to ω we obtain

$$\sigma(x) = (M-1)\int_0^1 (1-\eta\omega)^{M-2}\,d\omega \tag{4.69}$$

which we write as

$$\sigma(x) = -\frac{1}{\eta}\int_0^1 (M-1)(1-\eta\omega)^{M-2}(-\eta)\,d\omega \tag{4.70}$$

Recognising the integrand as the derivative of $(1-\eta\omega)^{M-1}$ with respect to ω we evaluate the integral to obtain

$$\sigma(x) = \frac{1}{\eta}\big(1-(1-\eta)^{M-1}\big) \tag{4.71}$$

Since $1 < \eta < 2$ it follows that $|1-\eta| < 1$ so the $(1-\eta)^{M-1}$ term converges to zero exponentially fast as $M \to \infty$, i.e. for every $\beta > 0$,

$$\sigma(x) = \frac{1}{\eta} + O\left(\frac{1}{M^\beta}\right) \quad \text{as} \quad M \to \infty \tag{4.72}$$

Since this is independent of x then by uniformity we have by (4.63) that

$$\mathbf{P}(X_i \,\text{NN}\, X_i') = \frac{1}{\eta}\mu(A)O(\delta) \quad \text{as} \quad M \to \infty \tag{4.73}$$

Finally, since $A = C \setminus B$, $\mu(C) = 1$ and $\mu(B) = O(\delta)$ (condition **C.3**) we have that $\mu(A) = 1 - O(\delta)$ and hence

$$\mathbf{P}(X_i \,\text{NN}\, X_i') = \frac{1}{\eta} + O(\delta) \quad \text{as} \quad M \to \infty \tag{4.74}$$

as required. □

Using Lemma 4.7 and Lemma 4.8 we compute the first few values of η and η^{-1}, shown in the following table. These values have also been confirmed experimentally for each $1 \le m \le 10$ by performing five experiments, each time generating $M = 1000$ points uniformly at random in the unit hypercube in $[0,1]^m$ and counting the number of components in the associated first nearest neighbour graphs, then averaging over all experiments. The results are shown in Table 4.1.

Table 4.1: Theoretical and experimental values of $\mathbf{P}(X_i \,\text{NN}\, X_i')$.

m	1	2	3	4	5	6	7	8	9	10
η	1.5	1.609	1.688	1.747	1.793	1.830	1.859	1.883	1.902	1.918
η^{-1}	0.667	0.623	0.593	0.572	0.558	0.547	0.538	0.531	0.526	0.521
Expt	0.663	0.624	0.594	0.573	0.534	0.539	0.526	0.506	0.495	0.492

4.5.6 Difficulties with the non–uniform case

In [Henze 1987] we find an analogue of Theorem 4.2 for point sets selected according to *non–uniform* sampling distributions. However, the methods we have developed in this section cannot be easily extended to encompass non–uniform distributions. Recall that for the uniform case we have the exact expression

$$v_x(r) = (1 - \eta\omega_x(r))^{M-2} \text{ for all } 0 \leq r \leq \delta \tag{4.75}$$

For the non–uniform case, using the mean value theorems of the integral and differential calculus it can be shown that

$$\alpha(x, y) = \eta\omega_x(r)(1 + O(r)) \quad \text{as} \quad r \to 0 \tag{4.76}$$

for every $y \in C_x(r)$ and since this is independent of any particular $y \in C_x(r)$ it follows that

$$v_x(r) = (1 - \eta\omega_x(r) + O(r))^{M-2} \quad \text{as} \quad r \to 0 \tag{4.77}$$

Furthermore, since we need only consider r in the range $0 \leq r \leq \delta = 1/M^\epsilon$, in contrast to (4.67) we now have

$$v_x(r) = \left(1 - \eta\omega_x(r) + O\left(\frac{1}{M^\epsilon}\right)\right)^{M-2} \quad \text{as} \quad M \to \infty \tag{4.78}$$

From here we see that the error associated with this approximate expression for $v_x(r)$ in terms of $\omega_x(r)$ grows *exponentially* as the number of points $M \to \infty$. This error will therefore dominate the leading term in any (polynomial) asymptotic expansion of $\sigma(x)$, and consequently of $\mathbf{P}(X_i \,\text{NN}\, X_i')$, in terms of the number of points as $M \to \infty$.

4.6 Summary

In this chapter we have developed an analysis of near neighbour graphs of which only Theorem 4.1 is explicitly required for the proof of the Gamma test in Chapter 7. We have also used the technique of 'boundary shrinking' developed in Chapter 3 to establish an asymptotic formula for the expected number of components in a random first nearest neighbour graph for which the points are sampled uniformly from a set $C \subset \mathbb{R}^m$ satisfying conditions **C.1**,**C.2** and **C.3**. This result is also confirmed experimentally.

In the next two chapters we turn our attention to establishing the required asymptotic upper bounds on the variance of terms $A_M(k)$, $B_M(k)$ and $C_M(k)$ that arise from the decomposition of $\gamma_M(k)$ specified in Chapter 2.

Chapter 5

L–dependent random variables

5.1 Introduction

In this chapter we establish some quite elegant statistical results that allow us to obtain the required upper bounds on the variance of terms $A_M(k)$ and $B_M(k)$.

By hypothesis, each component variable R_i of the random noise sample $R = (R_1, \ldots, R_M)$ is independent of every other R_j and also of the random point sample $X = (X_1, \ldots, X_M)$. The terms $(R_{N[i,k]} - R_i)^2$ of the sum $A_M(k)$ are therefore statistically independent of one another unless they share a common subscript. By Theorem 4.1 this means that any one term in the sum $A_M(k)$ can be statistically dependent on *at most* some fixed number of other terms that is independent of the sample size M. Consequently, it is relatively straightforward to obtain the required upper bound on the variance of $A_M(k)$ and as we shall see in Chapter 7, similar notions are sufficient to establish an adequate upper bound on the variance of $B_M(k)$.

5.2 A weak law of large numbers for independent random variables

The weak law of large numbers for independent random variables is a result of classical probability theory. Let $Y = (Y_1, \ldots, Y_M)$ be a random sample of independent and identically distributed random variables Y_i, each defined on the probability space C and having common mean and variance μ and σ^2 respectively.

The *sample mean* of the random sample $Y \in C^M$ is defined by

$$\bar{Y}_M = \frac{1}{M} \sum_{i=1}^{M} Y_i \tag{5.1}$$

and is itself a random variable on the sample space C^M. Since the component variables Y_i are identically distributed, the *expected value* of $\bar{Y}_M$ over all random samples $Y \in C^M$ is given by

$$\mathcal{E}(\bar{Y}_M) = \frac{1}{M}\sum_{i=1}^{M}\mathcal{E}(Y_i) = \mu \tag{5.2}$$

The following lemma gives the *variance* of $\bar{Y}_M$.

Lemma 5.1. *Let $Y = (Y_1, \ldots, Y_M)$ be a random sample of independent and identically distributed random variables having common mean and variance μ and σ^2 respectively. Then*

$$\mathrm{Var}(\bar{Y}_M) = \frac{\sigma^2}{M} \tag{5.3}$$

Proof. Without loss of generality suppose that $\mu = 0$ so that $\mathrm{Var}(\bar{Y}_M) = \mathcal{E}(\bar{Y}_M^2)$, and write this as

$$\mathrm{Var}(\bar{Y}_M) = \frac{1}{M^2}\left(\sum_{i=1}^{M}\mathcal{E}(Y_i^2) + \sum_{\substack{i,j=1\\ i\neq j}}^{M}\mathcal{E}(Y_iY_j)\right) \tag{5.4}$$

Since the Y_i are identically distributed, $\mathcal{E}(Y_i^2) = \sigma^2$ for all $1 \leq i \leq M$. Furthermore, since the Y_i are independent with $\mathcal{E}(Y_i) = 0$ it follows that $\mathcal{E}(Y_iY_j) = \mathcal{E}(Y_i)\mathcal{E}(Y_j) = 0$ for all $1 \leq i \neq j \leq M$. Hence $\mathrm{Var}(\bar{Y}_M) = \sigma^2/M$ as required. □

Using Chebyshev's inequality (Lemma 2.1) we obtain the following result of classical probability theory, known as the *the weak law of large numbers* for independent and identically distributed random variables.

Corollary 5.1. *Let $Y_1, \ldots, Y_M$ be a sample of independent and identically distributed random variables having common mean and variance μ and σ^2 respectively. Then for every $\epsilon > 0$,*

$$\mathbf{P}(|\bar{Y}_M - \mu| > \epsilon) \leq \frac{\sigma^2}{M\epsilon^2} \tag{5.5}$$

and if $\sigma^2 < \infty$, the sample mean $\bar{Y}_M$ converges in probability to its expected value μ as $M \to \infty$.

Proof. Apply Chebyshev's inequality to the random variable $\bar{Y}_M$ and apply the definition (2.21) of convergence in probability. □

The following result quantifies the rate at which the sample mean $\bar{Y}_M$ converges in probability to its expected value μ as $M \to \infty$.

Corollary 5.2. *For every $\kappa > 0$,*

$$\bar{Y}_M = \mu + O\left(\frac{1}{M^{1/2-\kappa}}\right) \tag{5.6}$$

with probability greater than $1 - O(1/M^{2\kappa})$ as $M \to \infty$.

Proof. Taking $\epsilon = 1/M^{1/2-\kappa}$ in Corollary 5.1,

$$\mathbf{P}\left(|\bar{Y}_M - \mu| > \frac{1}{M^{1/2-\kappa}}\right) \leq \frac{\sigma^2}{M^{2\kappa}} \tag{5.7}$$

and hence

$$\mathbf{P}\left(|\bar{Y}_M - \mu| \leq \frac{1}{M^{1/2-\kappa}}\right) > 1 - O\left(\frac{1}{M^{2\kappa}}\right) \tag{5.8}$$

as required. □

5.3 A weak law of large numbers for L–dependent random variables

We now generalise the results of Section 5.2 to a class of dependent random samples where for every component variable Y_i, there exist some fixed number of other component variables Y_j such that Y_i is independent of any subset of $Y_1, \ldots, Y_M$ not containing one or more of these Y_j. More precisely, let $L \geq 1$ be a fixed integer (independent of M). A random sample $Y = (Y_1, \ldots, Y_M)$ of identically distributed random variables is said to be *L–dependent* if for every $1 \leq i \leq M$, there exists a subset $V_i \subset \{1, \ldots, M\}$ of cardinality $|V_i| \leq L + 1$ and which contains i, such that for every $U \subset \{1, \ldots, M\}$ with $V_i \cap U = \emptyset$ we have that Y_i is independent of $\{Y_j : j \in U\}$. Note that the case $L = 0$ corresponds to a random sample of independent random variables.

Lemma 5.2. *Let $Y = (Y_1, \ldots, Y_M)$ be a random sample of identically distributed L–dependent random variables having common mean and variance μ and σ^2 respectively and define*

$$\bar{Y}_M = \frac{1}{M}\sum_{i=1}^{M} Y_i \tag{5.9}$$

Then

$$\mathrm{Var}(\bar{Y}_M) \leq \frac{(L+1)\sigma^2}{M} \tag{5.10}$$

Proof. Without loss of generality suppose that $\mu = 0$ and consider

$$\mathrm{Var}(\bar{Y}_M) = \mathcal{E}(\bar{Y}_M^2) = \frac{1}{M^2}\left(\sum_{i=1}^{M} \mathcal{E}(Y_i^2) + \sum_{\substack{i,j=1 \\ i \neq j}}^{M} \mathcal{E}(Y_i Y_j)\right) \tag{5.11}$$

By definition of L–dependence, for any particular Y_i there are at most L other Y_j (with $i \neq j$) such that $\mathcal{E}(Y_iY_j) \neq \mathcal{E}(Y_i)\mathcal{E}(Y_j)$. Hence there are at most ML pairs Y_i and Y_j $(i \neq j)$ such that $\mathcal{E}(Y_iY_j) - \mu^2 \neq 0$ and thus we see that there are at most ML non–zero terms in the second sum of (5.11). Moreover, since Y_i and Y_j are identically distributed and since $ab \leq \frac{1}{2}(a^2 + b^2)$ for any pair of real numbers a and b we have that

$$\mathcal{E}(Y_iY_j) \leq \frac{1}{2}(\mathcal{E}(Y_i^2) + \mathcal{E}(Y_j^2)) = \mathcal{E}(Y_i^2) \tag{5.12}$$

and hence

$$\mathcal{E}(Y_iY_j) - \mu^2 \leq \mathcal{E}(Y_i^2) - \mu^2 = \sigma^2 \tag{5.13}$$

Thus the second sum in (5.11) is bounded above by $ML\sigma^2$, and since the first sum in (5.11) is equal to $M\sigma^2$ it follows that

$$\mathrm{Var}(\bar{Y}_M) \leq \frac{(L+1)\sigma^2}{M} \tag{5.14}$$

as required. □

Using Chebyshev's inequality we obtain the following corollary of Lemma 5.2. This constitutes a *weak law of large numbers for L–dependent random variables.*

Corollary 5.3. *Let $Y = (Y_1, \ldots, Y_M)$ be a random sample of identically distributed L–dependent random variables having common mean and variance μ and σ^2 respectively. Then for every $\epsilon > 0$,*

$$\mathbf{P}(|\bar{Y}_M - \mu| > \epsilon) \leq \frac{(L+1)\sigma^2}{M\epsilon^2} \tag{5.15}$$

and if $\sigma^2 < \infty$, the sample mean $\bar{Y}_M$ converges in probability to its expected value μ as $M \to \infty$.

5.4 Statistical dependence in the noise sample R

In Chapter 2, corresponding to any function $g : \mathbb{R}^2 \to \mathbb{R}$ and any random point sample $X \in C^M$ we defined a set of random variables $(g_1(R), \ldots, g_M(R))$ on the space of random noise samples $\mathbb{R}^M$ by

$$g_i(R) = g(R_i, R_{N[i,k]}) \qquad 1 \leq i \leq M, \quad R \in \mathbb{R}^M \tag{5.16}$$

where $N[i,k]$ is the index of the kth nearest neighbour of X_i in X. The indexing structure $\mathcal{N}(X)$ inherited by a noise sample R from the associated point sample X thus imposes a *dependence structure* on the random variables $(g_1(R), \ldots, g_M(R))$, in the sense that the value taken by a particular $g_i(R) = g(R_i, R_{N[i,k]})$ may depend on the value taken by some other $g_j(R) = g(R_j, R_{N[j,k]})$ where $i \neq j$. Since each R_i is independent and identically distributed, this can happen only if one of the following occurs.

- $j = N[i,k]$: X_j is the kth nearest neighbour of X_i
  - by definition, this occurs for exactly one index j.
- $N[j,k] = i$: X_i is the kth nearest neighbour of X_j
  - by Theorem 4.1 this can occur for at most $kK(m)$ indices j.
- $N[j,k] = N[i,k]$: X_i and X_j have a common kth nearest neighbour
  - by Theorem 4.1 this can occur for at most $kK(m)$ indices j.

By Theorem 4.1, for *any* dependence structure imposed on the set $(g_1(R), \ldots, g_M(R))$ by a point sample X the value taken by any particular $g_i(R)$ can therefore depend on the value taken by *at most* $2kK(m) + 1$ of the other $g_j(R)$, where $K(m)$ is the maximum kissing number in $\mathbb{R}^m$.

Thus we have shown that $(g_1(R), \ldots, g_M(R))$ is a set of L–dependent random variables with $L = 2kK(m) + 1$ and applying Lemma 5.2 to $(g_1(R), \ldots, g_M(R))$ we obtain the following.

Theorem 5.1. *Let $X = (X_1, \ldots, X_M) \in C^M$ be a random sample of independent and identically distributed random variables, let $g : \mathbb{R}^2 \to \mathbb{R}$ be any function and define a set of identically distributed random variables $g_i : \mathbb{R}^m \to \mathbb{R}$ by*

$$g_i(R) = g(R_i, R_{N[i,k]}) \tag{5.17}$$

where $N[i,k]$ is the index of the kth nearest neighbour of X_i in X. Let $G_M = G_M(X,R)$ denote their sample mean,

$$G_M = \frac{1}{M} \sum_{i=1}^{M} g_i(R) \tag{5.18}$$

Then

$$\mathrm{Var}(G_M) \leq \frac{2(kK(m)+1)\,\mathrm{Var}(g_i(R))}{M} \tag{5.19}$$

where $K(m)$ is the maximum kissing number in $\mathbb{R}^m$, and this bound is independent of any particular $X \in C^M$.

Proof. By the definition of conditional variance (see [Feller 1971]) it is easily shown that

$$\mathrm{Var}(G_M) = \mathrm{Var}(\mathcal{E}(G_M \,|\, X)) + \mathcal{E}(\mathrm{Var}(G_M|X)) \tag{5.20}$$

Since the R_i are identically distributed, the expected value of G_M is independent of any particular $X \in C^M$ and hence the first term of (5.20) is zero. Furthermore, for any $X \in C^M$ the random sample $(g_1(R), \ldots, g_M(R))$ is a sequence of L–dependent random variables with $L = 2kK(m) + 1$. Hence by Lemma 5.2 the second term of (5.20) is bounded by $(L+1)\sigma^2/M$ where $\sigma^2 = \mathrm{Var}(g_i(R))$, and the result follows. □

For the proof of the Gamma test, Theorem 5.1 is all we need regarding L–dependent random variables. We now digress to prove a Central Limit Theorem for this class of dependent random variables.

5.5 A Central Limit Theorem for L–dependent random variables

A sequence of distribution functions F_M is said to *converge* to some distribution function F as $M \to \infty$ if $F_M(x) \to F(x)$ as $M \to \infty$ at each point of continuity of F.

Let Y_M and Y be random variables having distribution functions F_M and F respectively. Then $F_M \to F$ as $M \to \infty$ if

$$\mathbf{P}(Y_M \leq x) \to \mathbf{P}(Y \leq x) \quad \text{as} \quad M \to \infty \tag{5.21}$$

for every x such that $\mathbf{P}(Y = x) = 0$, in which case we say that Y_M *converges in distribution* to Y as $M \to \infty$.

Let $X_1, X_2, \ldots$ be a sequence of random variables and consider the partial sums $S_M = \sum_{i=1}^{M} X_i$. Letting $\mathcal{E}(S_M) = \mu_M$ and $\mathrm{Var}(S_M) = \sigma_M^2$ we define the normalised partial sums $S_M^* = (S_M - \mu_M)/\sigma_M$ so that $\mathcal{E}(S_M^*) = 0$ and $\mathrm{Var}(S_M^*) = 1$. The sequence $X_1, X_2, \ldots$ is said to satisfy the *Central Limit Theorem* if the distribution of their normalised partial sums S_M^* converges to the standard normal distribution $\Phi(x)$ as $M \to \infty$, i.e. for all $x \in \mathbb{R}$,

$$\mathbf{P}\left(\frac{S_M - \mu_M}{\sigma_M} \leq x\right) \to \Phi(x) \quad \text{as} \quad M \to \infty \tag{5.22}$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt \tag{5.23}$$

5.5.1 A result of Baldi and Rinott

The following appears in [Baldi and Rinott 1989] (we change their terminology and notation).

Theorem 5.2. *Let $X_1, X_2, \ldots$ be a sequence of identically distributed random variables with $|X_i| \leq B$ such that $X = (X_1, \ldots, X_M)$ is a random sample of L–dependent random variable for each M. Let $S_M = \sum_{i=1}^{M} X_i$ and define $\mu_M = \mathcal{E}(S_M)$ and $\sigma_M^2 = \mathrm{Var}(S_M)$. Then*

$$\left|\mathbf{P}\left(\frac{S_M - \mu_M}{\sigma_M} \leq x\right) - \Phi(x)\right| \leq \frac{32(1+\sqrt{6})(L+1)B^{3/2}M^{1/2}}{\sigma_M^{3/2}} \tag{5.24}$$

In [Baldi and Rinott 1989] the proof of Theorem 5.2 proceeds from a rather cumbersome inequality of [Stein 1986], and the role played by L–dependence is somewhat obscure. In Theorem 5.5 we present a more general result than that of Theorem 5.2 – instead of requiring that the X_i are uniformly bounded, we need only that their absolute moments are uniformly bounded. Our proof, in which the role played by L–dependence is clearly illustrated, is based on an argument of [Noether 1970] and uses techniques that are very close to those found in [Petrovskaya and Leontovich 1982].

5.5.2 The method of moments

The next two results are well known theorems of classical probability theory. Their proofs may be found in [Kendall and Stuart 1963] or [Billingsley 1979].

Theorem 5.3. *Let X be a random variable with $\mathcal{E}(X^r) < \infty$ for all $r = 1, 2, \ldots$ and suppose that*

$$\sum_{r=1}^{\infty} \frac{\mathcal{E}(X^r)}{r!} x^r < \infty \quad \text{for some} \quad x > 0 \tag{5.25}$$

(i.e. the power series has a positive radius of convergence). Then X is the only random variable with moments $\mathcal{E}(X), \mathcal{E}(X^2), \ldots$.

The distribution function of any random variable satisfying the conclusion of Lemma 5.3 is said to be *uniquely determined by its moments.*

Theorem 5.4 (The method of moments). *Let X_M be a sequence of random variables and suppose that each X_M has moments of all orders. Let X be a random variable whose distribution is uniquely determined by its moments and suppose that $\mathcal{E}(X_M^r) \to \mathcal{E}(X^r)$ as $M \to \infty$ for all $r = 1, 2, \ldots$. Then the distribution of X_M converges to the distribution of X as $M \to \infty$.*

5.5.3 The standard normal distribution

The *standard normal distribution* is defined by

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt \tag{5.26}$$

Lemma 5.3. *The moments of the standard normal distribution $\Phi(x)$ are given by*

$$n_r = \begin{cases} 0 & \text{for } r \text{ odd} \\ (r-1)(r-3)\ldots 3 & \text{for } r \text{ even} \end{cases} \tag{5.27}$$

Proof. By definition,

$$n_r = \int_{-\infty}^{\infty} x^r e^{-x^2/2}\, dx = -\int_{-\infty}^{\infty} x^{r-1} \frac{d}{dx}\left(e^{-x^2/2}\right) dx \tag{5.28}$$

Integrating by parts,

$$n_r = (r-1) \int_{-\infty}^{\infty} x^{r-2} e^{-x^2/2}\, dx = (r-1) n_{r-2} \tag{5.29}$$

and since $n_1 = 0$ and $n_2 = 1$, the result follows by induction on r. □

Lemma 5.4. *The standard normal distribution $\Phi(x)$ is uniquely determined by its moments.*

Proof. By Lemma 5.3, $|n_r| \leq r!$ so

$$\sum_{r=1}^{\infty} \frac{n_r}{r!} x^r \leq \sum_{r=1}^{\infty} x^r \tag{5.30}$$

which is bounded for all $0 < x < 1$. Hence by Theorem 5.3, $\Phi(x)$ is uniquely determined by its moments. □

Thus, in order to show that the distribution of some sequence S_M converges to the standard normal distribution $\Phi(x)$ as $n \to \infty$, by Lemma 5.4 and Theorem 5.4 it is sufficient to show that the rth moment of S_M converges to the rth moment of $\Phi(x)$ as $M \to \infty$ for each $r = 1, 2, \ldots$.

5.5.4 A Central Limit Theorem for triangular arrays of L–dependent random variables

We extend Theorem 5.2 to encompass triangular arrays of random variables having finite moments of all orders.

Theorem 5.5. *For each $M \in \mathbb{N}$ let $(X_{M,1}, \ldots, X_{M,M})$ be a random sample of identically distributed L–dependent random variables and suppose that there exist finite constants $c_1, c_2, \ldots$ such that for every $X_{M,i}$, $\mathcal{E}(|X_{M,i}|^r) \leq c_r$ for all $r \in \mathbb{N}$. Let $S_M = \sum_{i=1}^{M} X_{M,i}$ and define $\mu_M = \mathcal{E}(S_M)$ and $\sigma_M^2 = \mathrm{Var}(S_M)$. Then*

$$\mathbf{P}\left(\frac{S_M - \mu_M}{\sigma_M} \leq x \right) \to \Phi(x) \quad \textit{as} \quad M \to \infty \tag{5.31}$$

provided there exists some $\epsilon > 0$ and some constant $C > 0$ such that

$$\sigma_M^2 \geq C M^{2/3(1+\epsilon)} \tag{5.32}$$

Remark: Note that condition (5.32) can be expressed as $M^{1/2}\sigma_M^{-3/2} = o(1)$ as $M \to \infty$. This is precisely the condition required to ensure that the right hand side of (5.24) converges to zero as $M \to \infty$

Proof. For each $r \in \mathbb{N}$ let

$$m_r = \mathcal{E}\left(\left(\frac{S_M - \mu_M}{\sigma_M} \right)^r \right) \tag{5.33}$$

By Lemma 5.4 and Theorem 5.4 it is sufficient to prove that $m_r \to n_r$ as $M \to \infty$ for each $r \in \mathbb{N}$. Let $Y_i = X_{M,i} - \mu_M$ so that $\mathcal{E}(Y_i) = 0$ (for brevity, we suppress the

dependence of Y_i on M). By hypothesis there exist finite constants $c_1, c_2, \ldots$ such that for every $X_{M,i}$, $\mathcal{E}(|X_{M,i}|^r) \leq c_r$ for all $r \in \mathbb{N}$. Hence for every Y_i it follows that

$$\mathcal{E}(|Y_i|^r) \leq c'_r = \sum_{h=0} r \binom{r}{h} c_j c_1^{r-j} < \infty \tag{5.34}$$

for all $r \in \mathbb{N}$. Let $T_M = S_M - \mu_M = \sum_{i=1}^{M} Y_i$ so that $\mathcal{E}(T_M) = 0$ and $\mathcal{E}(T_M^2) = \sigma_M^2$, and define $\mu_r = \mathcal{E}(T_M^r)$. We need to show that

$$\mu_r / \sigma_M^r \to n_r \quad \text{as} \quad M \to \infty \tag{5.35}$$

for each $r \in \mathbb{N}$. Since $\mu_1 = 0$ and $\mu_2 = \sigma_M^2$, by (5.27) we see that (5.35) is satisfied for $r = 1$ and $r = 2$.

Since $X_{M,1}, \ldots X_{M,M}$ are L–dependent, the $Y_1, \ldots, Y_M$ are also L–dependent. For each $1 \leq i \leq M$ let $V_i \subset \{1, \ldots, M\}$ denote the subset (whose cardinality satisfies $V_i| \leq L+1$ and which contains i) such that for every $U \subset \{1, \ldots, M\}$ with $V_i \cap U = \emptyset$, Y_i is independent of $\{Y_j : j \in U\}$. We consider the cases where r is odd and r is even separately and to illustrate the method we look at the cases $r = 3$ and $r = 4$ in detail.

Special case 1: $r = 3$.

Consider

$$\mu_3 = \mathcal{E}(T_M^3) = \sum_{i,j,k} \mathcal{E}(Y_i Y_j Y_k) \tag{5.36}$$

If Y_k is independent of $\{Y_i, Y_j\}$ then $\mathcal{E}(Y_i Y_j Y_k) = \mathcal{E}(Y_i Y_j)\mathcal{E}(Y_k)$, and since $\mathcal{E}(Y_k) = 0$ and $\mathcal{E}(Y_i Y_j) \leq \mathcal{E}(|Y_i Y_j|) \leq \mathcal{E}(Y_i^2) \leq c'_2 < \infty$ it follows that $\mathcal{E}(Y_i Y_j Y_k) = 0$ whenever Y_k is independent of both $\{Y_i, Y_j\}$. For brevity of notation let $Y_i \rightleftharpoons Y_j$ indicate that Y_i is dependent on Y_j. By symmetry on the indices we see that $\mathcal{E}(Y_i Y_j Y_k)$ is non-zero only if one of the following holds.

(1) $Y_i \rightleftharpoons Y_j$ and $Y_i \rightleftharpoons Y_k$

(2) $Y_i \rightleftharpoons Y_j$ and $Y_j \rightleftharpoons Y_k$

(3) $Y_i \rightleftharpoons Y_k$ and $Y_j \rightleftharpoons Y_k$

Since $Y_1, \ldots, Y_M$ are L–dependent it follows that at most $M(L+1)^2$ of the terms $Y_i Y_j Y_k$ satisfy (1). This is because for each fixed i we need that

- $j \in V_i$ — there are at most $L + 1$ such j
- $k \in V_i$ — there are at most $L + 1$ such k

Thus it follows that there can be at most $3M(L+1)^2$ non–zero terms in (5.36). Since the Y_i are identically distributed it follows by the AM–GM inequality that $\mathcal{E}(|Y_iY_jY_k|) \leq \mathcal{E}(|Y_i|^3)$ and hence $|\mathcal{E}(Y_iY_jY_k)| \leq c_3' < \infty$. Thus we see that

$$\mu_3 \leq 3M(L+1)c_3' < \infty \tag{5.37}$$

and by condition (5.32) it follows that

$$\frac{\mu_3}{\sigma_M^3} \leq \frac{3(L+1)c_3'}{C^{3/2}M^{\epsilon}} \to 0 \quad \text{as} \quad M \to \infty \tag{5.38}$$

as required.

Special case 2: $r = 4$

Consider

$$\mu_4 = \mathcal{E}(T_M^4) = \sum_{i,j,k,l} \mathcal{E}(Y_iY_jY_kY_l) \tag{5.39}$$

where the sum is taken over all $1 \leq i, j, k, l \leq M$. If Y_l is independent of $\{Y_i, Y_j, Y_k\}$ then $\mathcal{E}(Y_iY_jY_kY_l) = \mathcal{E}(Y_iY_jY_k)\mathcal{E}(Y_l)$, and since $\mathcal{E}(Y_l) = 0$ and $\mathcal{E}(Y_iY_jY_k) \leq \mathcal{E}(Y_i^3) \leq c_3' < \infty$ it follows that $\mathcal{E}(Y_iY_jY_kY_l) = 0$ whenever Y_l is independent of $\{Y_i, Y_j, Y_k\}$. By symmetry on the indices, if $\mathcal{E}(Y_iY_jY_kY_l) \neq 0$ then one of the following must hold.

(1) $Y_i \rightleftharpoons Y_j\,,\ Y_k \rightleftharpoons Y_l$

(2) $Y_i \rightleftharpoons Y_k\,,\ Y_j \rightleftharpoons Y_l$

(3) $Y_i \rightleftharpoons Y_l\,,\ Y_j \rightleftharpoons Y_k$

We show that the terms $Y_iY_jY_kY_l$ for which $\mathcal{E}(Y_iY_jY_kY_l) \neq 0$ are principally provided by those that can be decomposed into two independent pairs of the form Y_iY_j and Y_kY_l. Let $\{Y_i, Y_j\} \not\rightleftharpoons \{Y_k, Y_l\}$ indicate that $\{Y_i, Y_j\}$ is independent of $\{Y_k, Y_l\}$. Any term $Y_iY_jY_kY_l$ for which (1) holds must satisfy one of the following.

(a) $Y_i \rightleftharpoons Y_j\,,\ Y_k \rightleftharpoons Y_l\,,\ \{Y_i, Y_j\} \not\rightleftharpoons \{Y_k, Y_l\}$

(b) $Y_i \rightleftharpoons Y_j\,,\ Y_k \rightleftharpoons Y_l\,,\ Y_i \rightleftharpoons Y_k$

(c) $Y_i \rightleftharpoons Y_j\,,\ Y_k \rightleftharpoons Y_l\,,\ Y_i \rightleftharpoons Y_l$

(d) $Y_i \rightleftharpoons Y_j\,,\ Y_k \rightleftharpoons Y_l\,,\ Y_j \rightleftharpoons Y_k$

(e) $Y_i \rightleftharpoons Y_j\,,\ Y_k \rightleftharpoons Y_l\,,\ Y_j \rightleftharpoons Y_l$

Among the M^4 terms $Y_iY_jY_kY_l$, at most $M^2(L+1)^2$ of them can satisfy (a) because for each i we require that

- $j \in V_i$ — there are at most $L+1$ such j
- $k \in \{1, \ldots, M\} \setminus (V_i \cup V_j)$ — there are at most M such k
- $l \in V_k \setminus (V_i \cup V_j)$ — there are at most $L+1$ such l

On the other hand, at most $M(L+1)^3$ of the $Y_iY_jY_kY_l$ satisfy (b) because for each i we require that

- $j \in V_i$ — there are at most $L+1$ such j
- $k \in V_i$ — there are at most $L+1$ such k
- $l \in V_k$ — there are at most $L+1$ such l

Thus at most $4M(L+1)^3$ of the terms $Y_iY_jY_kY_l$ satisfying (1) do not arise as a result of two independent pairs of the form (Y_i, Y_j) and (Y_k, Y_l). These terms can therefore be ignored when looking at the large M behaviour of (5.39), and since the above argument also applies to the pairs $\{(Y_i, Y_k), (Y_j, Y_l)\}$ and $\{(Y_i, Y_l), (Y_j, Y_k)\}$ it follows that

$$\mu_4 \sim 3 \sum_{i,j} \mathcal{E}(Y_iY_j) \sum_{k,l} \mathcal{E}(Y_kY_l) \quad \text{as} \quad M \to \infty \tag{5.40}$$

Thus, since $\sigma_M^2 = \sum_{i,j} \mathcal{E}(Y_iY_j)$ we see that $\mu_4 \to 3\sigma_M^4$ and hence $\mu_4/\sigma_M^4 \to 3$ as $M \to \infty$, as required.

General case 1: r odd.

For r odd, the sum

$$\mu_r = \sum_{i_1, \ldots i_r} \mathcal{E}(Y_{i_1} \ldots Y_{i_r}) \tag{5.41}$$

is dominated by $(r-3)/2$ dependent pairs Y_iY_j and one dependent triple $Y_iY_jY_k$. This is because the number of such arrangements is asymptotically of greater order (in terms of the number of points M) than the total number of other arrangements for which the summand is non–zero (e.g. those having $(r-5)/2$ dependent pairs and one dependent quintuple). If $C < \infty$ is the number of ways of choosing $(r-3)/2$ distinct pairs and one triple from a set of r elements where r is odd then

$$\mu_r \sim C \left(\sum_{i,j} \mathcal{E}(Y_iY_j) \right)^{(r-3)/2} \sum_{i,j,k} \mathcal{E}(Y_iY_jY_k) \quad \text{as} \quad M \to \infty \tag{5.42}$$

and since $\sigma_M^2 = \sum_{i,j} \mathcal{E}(Y_iY_j)$ and $\mu_3 = \sum_{i,j,k} \mathcal{E}(Y_iY_jY_k)$ we see that

$$\mu_r \sim C\sigma_M^{r-3}\mu_3 \quad \text{as} \quad M \to \infty \tag{5.43}$$

Thus $\mu_r/\sigma_M^r \sim C\mu_3/\sigma_M^3$ as $M \to \infty$, and we have previously seen that $\mu_3/\sigma_M^3 \to 0$ as $M \to \infty$. Hence $\mu_r/\sigma_M^r \to n_r$ as $M \to \infty$ for r odd, as required.

General case 2: r even.

For r even, the sum

$$\mu_r = \sum_{i_1,\ldots i_r} \mathcal{E}(Y_{i_1} \ldots Y_{i_r}) \tag{5.44}$$

is dominated by those arrangements consisting of $r/2$ dependent pairs Y_iY_j. Again, this is because the number of such arrangements is asymptotically of greater order (in terms of the number of points M) than the total number of other arrangements for which the summand is non–zero (e.g. those involving $r/2-2$ dependent pairs and one dependent quadruple). If r is even then there are $(r-1)(r-3)\ldots 3$ ways of selecting $r/2$ distinct pairs from r elements and we have that

$$\mu_r \sim (r-1)(r-3)\ldots 3 \left(\sum_{i,j} \mathcal{E}(Y_iY_j) \right)^{r/2} \quad \text{as} \quad M \to \infty \tag{5.45}$$

Thus, since $\sigma_M^2 = \sum_{i,j} \mathcal{E}(Y_iY_j)$ we obtain

$$\mu_r \sim (r-1)(r-3)\ldots 3\sigma_M^r \quad \text{as} \quad M \to \infty \tag{5.46}$$

an hence $\mu_r/\sigma_M^r \sim (r-1)(r-3)\ldots 3$ as $M \to \infty$ for r even, as required. □

5.6 Summary

In this chapter we have established the theory of sums of what we have called L-dependent random variables. While similar results have appeared scattered in the literature, we have presented a reasonably coherent theory for this class of random variables, of which only Theorem 5.1 is required for the proof of the Gamma test presented in Chapter 7. We hope that our simple proof of the Central Limit Theorem for L-dependent random variables serves to clarify the overall situation.

In the next chapter we address the more difficult issue of establishing an upper bound on the variance of term $C_M(k)$, which cannot be handled by the notion of L-dependence.

Chapter 6

Bounded functions of a point and its *k*th nearest neighbour

6.1 Introduction

As we shall see in Chapter 7, the notion of L–dependence is sufficient to establish the required upper bounds on the variance of terms $A_M(k)$ and $B_M(k)$. Recall that

$$C_M(k) = \frac{1}{M}\sum_{i=1}^{M} h_i(X) \tag{6.1}$$

where

$$h_i(X) = \frac{1}{2}((X_{N[i,k]} - X_i)\cdot\nabla f(X_i))^2 - A(M,k)|X_{N[i,k]} - X_i|^2 \tag{6.2}$$

To obtain an upper bound on the variance of $C_M(k)$ we are *not* able to apply the results of Chapter 5 because $(h_1(X),\ldots,h_M(X))$ is not a set of L–dependent random variables. To see this, suppose that M points are selected at random from the unit square $[0,1]^2 \subset \mathbb{R}^2$, suppose that the point X_i is located near the upper right corner of $[0,1]^2$ and suppose that the distance from X_i to its first nearest neighbour $X_{N[i,1]}$ is close to $\sqrt{2}$. Then *all* points other than X_i must be located in the lower left corner of $[0,1]^2$. Thus $|X_{N[j,1]} - X_j|$ is dependent on $|X_{N[i,1]} - X_i|$ for each $j \neq i$ and since statistical dependence is symmetric, it is clear that any one nearest neighbour distance may be statistically dependent on *all* the others.

6.2 The point sample X

We need an upper bound on the variance of

$$H_M = \frac{1}{M}\sum_{i=1}^{M} h_i(X) \tag{6.3}$$

Define $h_i^*(X) = h_i(X) - \mathcal{E}(h_i(X))$ and for brevity of notation write $h_i = h_i(X)$ and $h_i^* = h_i^*(X)$. By definition

$$\mathrm{Var}(H_M) = \frac{1}{M^2}\sum_{i=1}^{M}\mathcal{E}\left(h_i^{*2}\right) + \frac{1}{M^2}\sum_{i\neq j}\mathcal{E}\left(h_i^* h_j^*\right) \tag{6.4}$$

and since the h_i^* are identically distributed over C^M it follows that

$$\mathrm{Var}(H_M) \leq \frac{1}{M}|\mathcal{E}(h_1^{*2})| + |\mathcal{E}(h_1^* h_2^*)| \tag{6.5}$$

Lemma 6.1.

$$|\mathcal{E}(h_1^{*2})| \leq 4||h||\mathcal{E}(|h_1|) \tag{6.6}$$

where $||h|| = \sup\{|h(x,y)| \,:\, x,y \in C\}$.

Proof. Clearly,

$$|\mathcal{E}(h_1^{*2})| \leq \mathcal{E}(|h_1^*||h_1^*|) \tag{6.7}$$

and since $h_1^* = h_1 - \mathcal{E}(h_1)$ it follows that

$$|h_1^*| \leq |h_1| + \mathcal{E}(|h_1|) \leq 2||h|| \tag{6.8}$$

Hence $|\mathcal{E}(h_1^{*2})| \leq 2||h||\mathcal{E}(|h_1^*|)$ and since $\mathcal{E}(|h_1^*|) \leq 2\mathcal{E}(|h_1|)$ we obtain

$$|\mathcal{E}(h_1^{*2})| \leq 4||h||\mathcal{E}(|h_1|) \tag{6.9}$$

as required. □

By Lemma 6.1 and (6.5),

$$\mathrm{Var}(H_M) \leq \frac{4||h||}{M}\mathcal{E}(|h_1|) + |\mathcal{E}(h_1^* h_2^*)| \tag{6.10}$$

In section 6.4 we obtain an asymptotic upper bound on $|\mathcal{E}(h_1^* h_2^*)|$ of order $O(1/M)$ as $M \to \infty$. Our approach is based on methods developed in [Bickel and Breiman 1983] for any bounded function of a point X_i and its first nearest neighbour distance $|X_{N[i,1]} - X_i|$. We extend their treatment to encompass any bounded function $h(X_i, X_{N[i,k]})$ of a point X_i and its kth nearest neighbour $X_{N[i,k]}$.

6.3 The kth nearest neighbour ball

Let $\phi(x)$ denote the common density function of the X_i over C and denote the probability measure of any subset $A \subseteq C$ by

$$\mu(A) = \int_A \phi(x)\,dx \tag{6.11}$$

For any random point sample $X = (X_1, \ldots, X_M)$ in C^M, let d_1 be the distance from X_1 to its kth nearest neighbour $X_{N[1,k]}$ in X. We define the *kth nearest neighbour ball* of X_1 to be the ball $B_1 = B(X_1, d_1)$ centred at X_1 and having the kth nearest neighbour $X_{N[1,k]}$ of X_1 on its boundary.

For every random sample X there is a corresponding kth nearest neighbour ball $B_1 = B_x(|x - y|)$ where X_i takes the value $x \in C$ and $X_{N[1,k]}$ takes the value $y \in C$. The probability measure $\mu(B_1)$ of the kth nearest neighbour ball of X_1 is therefore a random variable over the sample space C^M and we seek to determine its distribution function and compute its moments.

Suppose X_1 is fixed at some value $x \in C$ and consider the set of samples

$$\{X \in C^M : X_1 = x\,, d_1 \leq r\} \tag{6.12}$$

for which the distance from $X_1 = x$ to its kth nearest neighbour $X_{N[1,k]}$ is at most equal to some $r > 0$. For any such sample X the ball $B_x(r)$ must contain *at least* k points distinct from x so the conditional probability that $d_1 \leq r$ given that $X_1 = x$ is equal to

$$\begin{aligned}
&\mathbf{P}(d_1 \leq r \,\big|\, X_1 = x) \\
&\quad = \mathbf{P}\big(B_x(r) \text{ contains at least } k \text{ points distinct from } x\big) \\
&\quad = 1 - \sum_{j=0}^{k-1} \mathbf{P}\big(B_x(r) \text{ contains exactly } j \text{ points distinct from } x\big)
\end{aligned}$$

For any X_j with $j \neq 1$, the probability that X_j takes a value in the ball $B_x(r)$ is equal to the probability measure $\mu(B_x(r))$. Similarly, the probability that X_j takes a value in the complement of $B_x(r)$ is equal to $1 - \mu(B_x(r))$. Noting that there are precisely $\binom{M-1}{j}$ different ways of selecting a set of j points from the remaining $M - 1$ points $X_2, \ldots, X_M$ we obtain

$$\mathbf{P}(d_1 \leq r \,\big|\, X_1 = x) = 1 - \sum_{j=0}^{k-1} \binom{M-1}{j} \mu(B_x(r))^j (1 - \mu(B_x(r)))^{M-j-1} \tag{6.13}$$

Clearly, if $d_1 \leq r$ then $B_x(d_1) \subseteq B_x(r)$ and hence $\mu(B_x(d_1)) \leq \mu(B_x(r))$. Conversely, if $\mu(B_x(d_1)) \leq \mu(B_x(r))$ then since the balls $B_x(d_1)$ and $B_x(r)$ are concentric it follows that $B_x(d_1) \subseteq B_x(r)$ and hence that $d_1 \leq r$. Thus

$$\mathbf{P}\big(\mu(B_x(d_1)) \leq \mu(B_x(r)) \,\big|\, X_1 = x\big) = 1 - \sum_{j=0}^{k-1} \binom{M-1}{j} \mu(B_x(r))^j (1 - \mu(B_x(r)))^{M-j-1} \tag{6.14}$$

Letting $z = \mu(B_x(r))$ this becomes

$$\mathbf{P}(\mu(B_x(d_1)) \leq z \,\big|\, X_1 = x) = 1 - \sum_{j=0}^{k-1} \binom{M-1}{j} z^j (1 - z)^{M-j-1} \tag{6.15}$$

and since this holds for all $x \in C$ we obtain

$$\mathbf{P}(\mu(B_1) \leq z) = 1 - \sum_{j=0}^{k-1} \binom{M-1}{j} z^j (1-z)^{M-j-1} \tag{6.16}$$

Thus the distribution function of the probability measure $\mu(B_1)$ of the kth nearest neighbour ball B_1 over the sample space C^M is given by

$$F(z) = \mathbf{P}(\mu(B_1) \leq z) = 1 - \sum_{j=0}^{k-1} \binom{M-1}{j} z^j (1-z)^{M-j-1} \tag{6.17}$$

Although we shall need only the first two moments of $\mu(B_1)$, it is convenient to compute an expression for its general moment $\mathcal{E}(\mu(B_1)^\alpha)$.

Lemma 6.2. *For every integer $\alpha \geq 1$,*

$$\mathcal{E}(\mu(B_1)^\alpha) = \frac{(k+\alpha-1)\ldots k}{(M+\alpha-1)\ldots M} \tag{6.18}$$

Proof. Since μ is a probability measure we have that $0 \leq \mu(B_1) \leq 1$ so integrating by parts we get

$$\mathcal{E}(\mu(B_1)^\alpha) = \int_0^1 z^\alpha F'(z)\, dz = 1 - \alpha \int_0^1 z^{\alpha-1} F(z)\, dz \tag{6.19}$$

where $F(z)$ is the distribution function of $\mu(B_1)$. By (6.17),

$$\mathcal{E}(\mu(B_1)^\alpha) = \sum_{j=0}^{k-1} \binom{M-1}{j} \int_0^1 z^{j+\alpha-1} (1-z)^{M-j-1}\, dz \tag{6.20}$$

The integral in (6.20) is the Beta function

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\, dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{6.21}$$

with parameters $a = j + \alpha$ and $b = M - j$. Hence

$$\mathcal{E}(\mu(B_1)^\alpha) = \alpha \sum_{j=0}^{k-1} \binom{M-1}{j} \frac{\Gamma(j+\alpha)\Gamma(M-j)}{\Gamma(M+\alpha)} \tag{6.22}$$

and writing

$$\binom{M-1}{j} = \frac{\Gamma(M)}{\Gamma(M-j)\Gamma(j+1)} \tag{6.23}$$

we see that

$$\mathcal{E}(\mu(B_1)^\alpha) = \alpha \sum_{j=0}^{k-1} \frac{\Gamma(M)\Gamma(j+\alpha)}{\Gamma(M+\alpha)\Gamma(j+1)} = \frac{\Gamma(M)\Gamma(k+\alpha)}{\Gamma(M+\alpha)\Gamma(k)} \tag{6.24}$$

and the result follows. □

6.4 An asymptotic upper bound on $\mathcal{E}(h_1^* h_2^*)$

We aim to show that $|\mathcal{E}(h_1^* h_2^*)| \leq c/M$ where c is a finite constant that is independent of the number of points M.

Let P be the probability measure on the space of random point samples C^M having the property that each component variable X_i of the random sample $X = (X_1, \ldots, X_M)$ is independent and identically distributed in C with probability density ϕ. Let $(a_1, a_2) \in C \times C$ be a pair of fixed points and define $S(a_1, a_2) \subset C$ to be the closed ball $B(a_1, |a_2 - a_1|)$ centred at a_1 and having radius $|a_2 - a_1|$ (so that a_2 is on the boundary of $S(a_1, a_2)$).

Let ϕ_S and $\phi_{C \setminus S}$ be the *conditional densities* of the X_i on hypothesis that $X_i \in S$ and $X_i \in C \setminus S$ respectively,

$$\phi_S(x) = \begin{cases} \phi(x)/\mu(S) & \text{if } x \in S \\ 0 & \text{if } x \notin S \end{cases} \tag{6.25}$$

$$\phi_{C \setminus S}(x) = \begin{cases} \phi(x)/\mu(C \setminus S) & \text{if } x \in C \setminus S \\ 0 & \text{if } x \notin C \setminus S \end{cases} \tag{6.26}$$

where $\mu(S)$ and $\mu(C \setminus S)$ are the probability measures of S and $C \setminus S$ respectively as defined in (6.11). Let $Q_{k+1}(\,\cdot\,\big|\,(a_1, a_2))$ be the probability measure on the sample space C^M such that

- $X_1, \ldots, X_{M-k-1}$ are i.i.d. in $C \setminus S$ according to $\phi_{C \setminus S}$
- $X_{M-k}, \ldots, X_{M-2}$ are i.i.d. in S according to ϕ_S
- $X_{M-1} = a_1$ and $X_M = a_2$

Let $Y = (Y_1, \ldots, Y_M)$ be a random sample where each Y_i takes values in $C \setminus S$ according to the conditional density $\phi_{C \setminus S}$ and similarly let $Z = (Z_1, \ldots, Z_M)$ be a random sample where each Z_i takes values in S according to the conditional density ϕ_S

For each random sample $X = (X_1, \ldots, X_M)$ in C^M consider the associated sample $\widetilde{X} = (\widetilde{X}_1, \ldots, \widetilde{X}_M)$ given by

$$\widetilde{X}_i = \begin{cases} X_i & \text{if} \quad X_i \in C \setminus S \quad \text{and} \quad 1 \leq i \leq M-k-1 \\ Y_i & \text{if} \quad X_i \in S \quad \text{and} \quad 1 \leq i \leq M-k-1 \\ Z_i & \text{for} \quad M-k \leq i \leq M-2 \\ a_1 & \text{for} \quad i = M-1 \\ a_2 & \text{for} \quad i = M \end{cases} \tag{6.27}$$

When constructing the restricted sample $\widetilde{X}$ associated with any particular $X \in C^M$, if any point X_i among the first $M-k-1$ components variables of X falls inside $S(a_1, a_2)$, it is replaced by the corresponding Y_i which is guaranteed to be in the complement $C \setminus S$.

Each of the next $k-1$ component variables are then replaced by the corresponding points Z_i (which are each guaranteed to be in S) and finally, the last two component variables X_{M-1} and X_M are replaced by the fixed points a_1 and a_2 respectively.

By construction, for every sample $\widetilde{X}$ defined in this way the ball $S(a_1, a_2)$ coincides with the kth nearest neighbour ball of its component variable of $\widetilde{X}_{M-1}$ (which is fixed at a_1).

Recall that $h_i(X) = h(X_i, X_{N[i,k]})$ where $X_{N[i,k]}$ is the kth nearest neighbour of X_i in the random sample $X = (X_1, \ldots, X_M)$. Let $\widetilde{X}_{\widetilde{N}[i,k]}$ denote the kth nearest neighbour of $\widetilde{X}_i$ in the associated sample $\widetilde{X} = (\widetilde{X}_1, \ldots, \widetilde{X}_M)$, and write $h_i(\widetilde{X}) = h(\widetilde{X}_i, \widetilde{X}_{\widetilde{N}[i,k]})$. Clearly, $h_i(X) \neq h_i(\widetilde{X})$ only if either $\widetilde{X}_i \neq X_i$ or $\widetilde{X}_{\widetilde{N}[i,k]} \neq X_{N[i,k]}$.

Let $r = k+1$ and let N be the number of 'changed' points among the first $M-r$ so that $N+r$ points have changed in total. Letting I denote the indicator function we write this as

$$N = \sum_{i=1}^{M-r} I(\widetilde{X}_i \neq X_i) \tag{6.28}$$

This number N is precisely the number of points X_i among the first $M-k-1$ points that fall in the set S and which are therefore replaced by the corresponding points Y_i in the complement of S. Its expected value is therefore given by $\mathcal{E}(N) = (M-r)\mu(S)$ where $\mu(S) = \mu(S(a_1, a_2))$ is the probability measure of S as defined in (6.11).

Suppose that $\widetilde{X}_i = X_i$. Then $\widetilde{X}_{\widetilde{N}[i,k]} \neq X_{N[i,k]}$ only if one of the following occurs,

- One of the first k nearest neighbours of X_i has been removed.
- One of the new points becomes one of the first k nearest neighbours of X_i

Let $X_j \neq \widetilde{X}_j$ be one of the changed points. For each $1 \leq u \leq k$, by Theorem 4.1 we know that X_j can be the uth nearest neighbour of *at most* $uK(m)$ other points X_i in $X = (X_1, \ldots, X_M)$, where $K(m)$ is the maximum kissing number in $\mathbb{R}^m$. Thus it follows that X_j could have been one of the kth nearest neighbours of *at most* $\beta(m,k)$ of the other points X_i where

$$\beta(m,k) = \sum_{u=1}^{k} uK(m) = \frac{1}{2}k(k+1)K(m) \tag{6.29}$$

Similarly, the new point $\widetilde{X}_j$ can be the nearest neighbour of at most $\beta(m,k)$ other points $\widetilde{X}_i$ in the modified sample $\widetilde{X} = (\widetilde{X}_1, \ldots, \widetilde{X}_M)$. Thus each of the $N+r$ changed points can affect at most $2\beta(m,k) = k(k+1)K(m)$ of the kth nearest neighbour relations so

$$\sum_{i=1}^{M-r} I(\widetilde{X}_i = X_i, \widetilde{X}_{\widetilde{N}[i,k]} \neq X_{N[i,k]}) \leq k(k+1)K(m)(N+r) \tag{6.30}$$

where I denotes the indicator function. For the sceptical reader we rigourise this argument in the following.

Lemma 6.3.

$$\sum_{i=1}^{M-r} I(\widetilde{X}_i = X_i, \widetilde{X}_{\widetilde{N}[i,k]} \neq X_{N[i,k]}) \leq k(k+1)K(m)(N+r) \tag{6.31}$$

where $K(m)$ is the maximum kissing number in $\mathbb{R}^m$.

Proof. For any $1 \leq i \leq M-r$,

$$\begin{aligned}
&I(\widetilde{X}_i = X_i, \widetilde{X}_{\widetilde{N}[i,k]} \neq X_{N[i,k]}) \\
&\quad \leq \sum_{u=1}^{k} \sum_{j,l=1}^{M} I\left(N[i,u] = j, \widetilde{N}[i,u] = l, X_j \neq \widetilde{X}_j \text{ or } X_l \neq \widetilde{X}_l\right) \\
&\quad \leq \sum_{u=1}^{k} \sum_{j,l=1}^{M} I(N[i,u] = j) I(\widetilde{N}[i,u] = l) I(X_j \neq \widetilde{X}_j) \\
&\qquad + \sum_{u=1}^{k} \sum_{j,l=1}^{M} I(N[i,u] = j) I(\widetilde{N}[i,u] = l) I(X_l \neq \widetilde{X}_l)
\end{aligned}$$

Hence

$$\begin{aligned}
&\sum_{i=1}^{M-r} I(\widetilde{X}_i = X_i, \widetilde{X}_{\widetilde{N}[i,k]} \neq X_{N[i,k]}) \\
&\quad \leq \sum_{u=1}^{k} \sum_{i=1}^{M} \sum_{j,l=1}^{M} I(N[i,u] = j) I(\widetilde{N}[i,u] = l) I(X_j \neq \widetilde{X}_j) \\
&\qquad + \sum_{u=1}^{k} \sum_{i=1}^{M} \sum_{j,l=1}^{M} I(N[i,u] = j) I(\widetilde{N}[i,u] = l) I(X_l \neq \widetilde{X}_l)
\end{aligned} \tag{6.32}$$

Let A denote the first sum in (6.32). Rearranging the order of summation, this is equal to

$$A = \sum_{j=1}^{M} I(X_j \neq \widetilde{X}_j) \sum_{u=1}^{k} \sum_{i=1}^{M} I(N[i,u] = j) \sum_{l=1}^{M} I(\widetilde{N}[i,u] = l) \tag{6.33}$$

For fixed u and i, $\widetilde{X}_i$ has precisely one uth nearest neighbour $\widetilde{X}_{\widetilde{N}[i,u]}$ in the modified sample $\widetilde{X}$ so there exists exactly one index l in the range $1 \leq l \leq M$ with $\widetilde{N}[i,u] = l$. Hence

$$A = \sum_{j=1}^{M} I(X_j \neq \widetilde{X}_j) \sum_{u=1}^{k} \sum_{i=1}^{M} I(N[i,u] = j) \tag{6.34}$$

Furthermore, for fixed j and u, by Theorem 4.1 it follows that X_j can be the uth nearest neighbour of at most $uK(m)$ points X_i, i.e. there are at most $uK(m)$ indices i in the range $1 \leq i \leq M$ such that $N[i,u] = j$. Hence

$$A \leq \sum_{j=1}^{M} I(X_j \neq \widetilde{X}_j) \sum_{u=1}^{k} uK(m) \tag{6.35}$$

so that

$$A \leq \frac{1}{2}k(k+1)K(m)\sum_{j=1}^{M} I(X_j \neq \widetilde{X}_j) \tag{6.36}$$

By definition of N we know that $X_j \neq \widetilde{X}_j$ for exactly $N+r$ indices j in the range $1 \leq j \leq M$ so

$$A \leq \frac{1}{2}k(k+1)K(m)(N+r) \tag{6.37}$$

Applying an identical argument to the second sum in (6.32) it thus follows that

$$\sum_{i=1}^{M-r} I(\widetilde{X}_i = X_i, \widetilde{X}_{\widetilde{N}[i,k]} \neq X_{N[i,k]}) \leq k(k+1)K(m)(N+r) \tag{6.38}$$

as required □

By construction, the samples $\{\widetilde{X} \,|\, X \in C^M\}$ are distributed in C^M according to the probability measure $Q_{k+1}(\cdot) = Q_{k+1}(\,\cdot\,\big|\,(a_1, a_2))$. Let $\mathcal{E}_{Q_{k+1}}$ denote an expectation taken with respect to the probability measure $Q_{k+1}(\cdot)$. The following lemma quantifies the difference between the expected value $\mathcal{E}_{Q_{k+1}}(h_1)$ taken with respect to Q_{k+1} and the expected value $\mathcal{E}(h_1)$ taken with respect to the (unrestricted) probability measure P. As we might expect, this difference depends both on k and on the set $S = S(a_1, a_2)$.

Lemma 6.4. *For any fixed $1 \leq k \leq M-1$,*

$$|\mathcal{E}_{Q_{k+1}}(h_1) - \mathcal{E}(h_1)| \leq C_0||h||\left(\frac{k+1}{M} + \mu(S)\right) \tag{6.39}$$

where $Q_{k+1}(\,\cdot\,) = Q_{k+1}(\,\cdot\,|\,(a_1, a_2))$, $S = S(a_1, a_2)$ is the ball centred at a_1 of radius $|a_2 - a_1|$ and

$$C_0 = 4(1 + k(k+1)K(m)) \tag{6.40}$$

is a constant depending only on m and k.

Proof. Let $r = k+1$. Since the left hand side of (6.39) is bounded above by $2||h||$, the bound holds trivially for $r \geq M/2$ since $C_0 \geq 4$. Suppose therefore that $r < M/2$.

For each $1 \leq i \leq M-r$ we have that $\mathcal{E}_{Q_r}(h_i(X)) = \mathcal{E}(h_i(\widetilde{X}))$ so

$$|\mathcal{E}_{Q_r}(h_i(X)) - \mathcal{E}(h_i(X))| = |\mathcal{E}(h_i(\widetilde{X})) - \mathcal{E}(h_i(X))| \tag{6.41}$$

Furthermore, since each X_i is identically distributed it follows that $\mathcal{E}(h_1(X)) = \mathcal{E}(h_i(X))$ and $\mathcal{E}(h_1(\widetilde{X})) = \mathcal{E}(h_i(\widetilde{X}))$ for each $1 \leq i \leq M-r$. Hence

$$\begin{aligned} |\mathcal{E}_{Q_r}(h_1(X)) - \mathcal{E}(h_1(X))| &= \frac{1}{M-r}\sum_{i=1}^{M-r} |\mathcal{E}(h_i(\widetilde{X})) - \mathcal{E}(h_i(X))| \\ &\leq \frac{1}{M-r}\,\mathcal{E}\left(\sum_{i=1}^{M-r} |h_i(\widetilde{X}) - h_i(X)|\right) \end{aligned} \tag{6.42}$$

Now, $h_i(X) = h(X_i, X_{N[i,k]})$ and $h_i(\widetilde{X}) = h(\widetilde{X}_i, \widetilde{X}_{\widetilde{N}[i,k]})$ can be different only if one of the following occurs,

- $\widetilde{X}_i \neq X_i$ (X_i has changed).
- $\widetilde{X}_i = X_i$ and $\widetilde{X}_{\widetilde{N}[i,k]} \neq X_{N[i,k]}$ (the kth nearest neighbour of X_i has changed).

and in each case, this difference cannot be greater than $2||h||$. Hence, by (6.42) it follows that

$$\begin{aligned} &|\mathcal{E}_{Q_r}(h_1(X)) - \mathcal{E}(h_1(X))| \\ &\quad \leq \frac{2||h||}{M-r}\mathcal{E}\left(\sum_{i=1}^{M-r}\Big(I(\widetilde{X}_i \neq X_i) + I(\widetilde{X}_i = X_i, \widetilde{X}_{\widetilde{N}[i,k]} \neq X_{N[i,k]})\Big)\right) \end{aligned} \tag{6.43}$$

where I denotes the indicator function. Let N be the number of changed points among the first $M-r$ points, i.e.

$$N = \sum_{i=1}^{M-r} I(\widetilde{X}_i \neq X_i) \tag{6.44}$$

By Lemma 6.3,

$$\sum_{i=1}^{M-r} I(\widetilde{X}_i = X_i, \widetilde{X}_{\widetilde{N}[i,k]} \neq X_{N[i,k]}) \leq k(k+1)K(m)(N+r) \tag{6.45}$$

so by (6.43) we obtain

$$\begin{aligned} |\mathcal{E}_{Q_r}(h_1(X)) - \mathcal{E}(h_1(X))| &\leq \frac{2||h||}{M-r}\,\mathcal{E}\big((1+k(k+1)K(m))(N+r)\big) \\ &= \frac{2||h||(1+k(k+1)K(m))}{M-r}(\mathcal{E}(N)+r) \end{aligned} \tag{6.46}$$

Writing $\mathcal{E}(N) = (M-r)\mu(S)$ where $\mu(S)$ is the probability measure of the ball centred at a_1 having radius $|a_2 - a_1|$ we obtain

$$|\mathcal{E}_{Q_r}(h_1(X)) - \mathcal{E}(h_1(X))| \leq 2||h||(1+k(k+1)K(m))\left(\frac{r}{M-r} + \mu(S)\right) \tag{6.47}$$

Finally, since $r < M/2$ we have that $r/(M-r) \leq 2r/M$ so

$$|\mathcal{E}_{Q_r}(h_1) - \mathcal{E}(h_1)| \leq 4||h||(1+k(k+1)K(m))\left(\frac{r}{M} + \mu(S)\right) \tag{6.48}$$

and the result follows on taking $C_0 = 4(1+k(k+1)K(m))$. □

Using Lemma 6.4 we can now compute an upper bound on the expected value of $h_1^* h_2^*$ as follows.

Lemma 6.5 (The critical lemma).

$$|\mathcal{E}(h_1^* h_2^*)| \leq C_1 ||h|| \left(\frac{1}{M} \mathcal{E}(|h_1|) + \mathcal{E}(|h_1 \mu(B_1)|) \right) \tag{6.49}$$

where $B_1 = B_1(X)$ is the kth nearest neighbour ball of X_1 and

$$C_1 = (3k+2)C_0 = 4(3k+2)(1 + k(k+1)K(m)) \tag{6.50}$$

is a constant depending only on m and k.

Proof. Consider

$$\mathcal{E}(h_1^* h_2^*) = \int_{C^M} h_1^* h_2^* \, dP \tag{6.51}$$

Writing $h_2^* = h_2 - \mathcal{E}(h_2)$ this becomes

$$\begin{aligned} \mathcal{E}(h_1^* h_2^*) &= \int_{C^M} h_1^* (h_2 - \mathcal{E}(h_2)) \, dP \\ &= \int h_1^* h_2 \, dP - \mathcal{E}(h_2) \int h_1^* \, dP \end{aligned} \tag{6.52}$$

For every pair $(a_1, a_2) \in C^2$ define the subset $C[a_1, a_2]$ of C^M by

$$C[a_1, a_2] = \left\{ X \in C^M \, : \, X_1 = a_1, X_{N[1,k]} = a_2 \right\} \tag{6.53}$$

so that $C[a_1, a_2]$ contains all those samples $X \in C^M$ where X_1 takes the value a_1 and the kth nearest neighbour of X_1 takes the value a_2. This induces a partition of C^M given by

$$C^M = \bigcup_{(a_1,a_2) \in C^2} C[a_1, a_2] \tag{6.54}$$

By Fubini's theorem we may evaluate the first integral on the right hand side of (6.52) by first integrating over each subset $C[a_1, a_2]$ separately, then integrating over all pairs $(a_1, a_2) \in C^2$. Thus we have that

$$\int_{C^M} h_1^* h_2 \, dP = \int_{(a_1,a_2) \in C^2} \left(\int_{C[a_1,a_2]} h_1^* h_2 \, dP_1 \right) dP_2 \tag{6.55}$$

where $P_1 = P_1(a_1, a_2)$ and P_2 are the projections of the probability measure P onto the subspaces $C[a_1, a_2]$ and C^2 respectively.

By construction, for each $X \in C[a_1, a_2]$ both X_1 and its kth nearest neighbour $X_{N[1,k]}$ are fixed at a_1 and a_2 respectively so by definition, $h_1^* = h_1^*(X) = h^*(X_1, X_{N[1,k]})$ is equal to $h^*(a_1, a_2)$ for all $X \in C[a_1, a_2]$. By (6.55) we thus have that

$$\int_{C^M} h_1^* h_2 \, dP = \int_{(a_1,a_2) \in C^2} h^*(a_1, a_2) \left(\int_{C[a_1,a_2]} h_2 \, dP_1 \right) dP_2 \tag{6.56}$$

Since P_1 is the projection of P onto $C[a_1, a_2]$, the inner integral in (6.56) corresponds to the conditional expectation of $h_2 = h_2(X)$ on hypothesis that X_1 is fixed at a_1 and and that its kth nearest neighbour $X_{N[1,k]}$ is fixed at a_2. Hence

$$\int_{C^M} h_1^* h_2 \, dP = \int_{(a_1,a_2)\in C^2} h^*(a_1, a_2) \mathcal{E}(h_2(X) \,\big|\, X_1 = a_1, X_{N[1,k]} = a_2) \, dP_2 \tag{6.57}$$

Furthermore, since

$$\int_{C[a_1,a_2]} dP_1 = 1 \tag{6.58}$$

and since $h^*(a_1, a_2)\mathcal{E}(h_2 \,\big|\, X_1 = a_1, X_{N[1,k]} = a_2)$ depends only on a_1 and a_2 it follows that

$$\begin{aligned}
&\int_{C^M} h_1^* h_2 \, dP \\
&= \int_{(a_1,a_2)\in C^2} h^*(a_1, a_2) \mathcal{E}(h_2(X) \,\big|\, X_1 = a_1, X_{N[1,k]} = a_2) \int_{C[a_1,a_2]} dP_1 \, dP_2 \\
&= \int_{(a_1,a_2)\in C^2} \int_{C[a_1,a_2]} h^*(a_1, a_2) \mathcal{E}(h_2(X) \,\big|\, X_1 = a_1, X_{N[1,k]} = a_2) \, dP_1 \, dP_2 \\
&= \int_{C^M} h_1^*(X) \mathcal{E}(h_2 \,\big|\, X_1, X_{N[1,k]}) \, dP
\end{aligned}$$

Substituting this into (6.52) we obtain

$$\mathcal{E}(h_1^* h_2^*) = \int_{C^M} h_1^*(X) \big(\mathcal{E}(h_2(X) \,\big|\, X_1, X_{N[1,k]}) - \mathcal{E}(h_2) \big) \, dP \tag{6.59}$$

and hence

$$|\mathcal{E}(h_1^* h_2^*)| \leq \int_{C^M} |h_1^*| \, |\mathcal{E}(h_2(X) \,\big|\, X_1, X_{N[1,k]}) - \mathcal{E}(h_2)| \, dP \tag{6.60}$$

The point samples $X \in C^M$ that are subject to the restrictions $X_1 = a_1$ and $X_{N[1,k]} = a_2$ (i.e. those contained in $C[a_1, a_2]$) are distributed according to the probability measure $Q_{k+1}(\,\cdot \,|\, (a_1, a_2))$[1], i.e.

$$\mathcal{E}(h_2(X) \,\big|\, X_1, X_{N[1,k]}) = \mathcal{E}_{Q_{k+1}}(h_2) \tag{6.61}$$

so by (6.60) it follows that

$$|\mathcal{E}(h_1^* h_2^*)| \leq \int_{C^M} |h_1^*| \, |\mathcal{E}_{Q_{k+1}}(h_2) - \mathcal{E}(h_2)| \, dP \tag{6.62}$$

By Lemma 6.4 we have that

$$|\mathcal{E}_{Q_{k+1}}(h_2) - \mathcal{E}(h_2)| \leq C_0 ||h|| \left(\frac{k+1}{M} + \mu(B_1) \right) \tag{6.63}$$

[1]Note that Q_{k+1} precisely corresponds to the projected measure P_1.

where B_1 is the kth nearest neighbour ball of X_1, so by (6.62) we obtain

$$|\mathcal{E}(h_1^* h_2^*)| \leq C_0||h|| \left(\frac{k+1}{M} \int_{C^M} |h_1^*| \, dP + \int_{C^M} |h_1^*| \mu(B_1) \, dP \right) \tag{6.64}$$

Now, since $|h_1^*| \leq |h_1| + \mathcal{E}(|h_1|)$ it follows that

$$\int_{C^M} |h_1^*| \, dP \leq \int_{C^M} |h_1| + \mathcal{E}(|h_1|) \, dP = 2\mathcal{E}(|h_1|) \tag{6.65}$$

and using the fact that $0 \leq \mu(B_1) \leq 1$, this also implies that

$$\int |h_1^*| \mu(B_1) \, dP \leq \int |h_1 \mu(B_1)| \, dP + \mathcal{E}(|h_1|) \int \mu(B_1) dP \tag{6.66}$$

i.e.

$$\int |h_1^*| \mu(B_1) \, dP \leq \mathcal{E}(|h_1 \mu(B_1)|) + \mathcal{E}(|h_1|) \mathcal{E}(\mu(B_1)) \tag{6.67}$$

Furthermore, taking $\alpha = 1$ in Lemma 6.2 we know that $\mathcal{E}(\mu(B_1)) = k/M$ so (6.67) becomes

$$\int_B |h_1^*| \mu(B_1) \, dP \leq \mathcal{E}(|h_1 \mu(B_1)|) + \frac{k}{M} \mathcal{E}(|h_1|) \tag{6.68}$$

Finally, substituting (6.65) and (6.68) into (6.64) we obtain

$$|\mathcal{E}(h_1^* h_2^*)| \leq C_0||h|| \left(\frac{(3k+2)}{M} \mathcal{E}(|h_1|) + \mathcal{E}(|h_1 \mu(B_1)|) \right) \tag{6.69}$$

Hence

$$|\mathcal{E}(h_1^* h_2^*)| \leq (3k+2) C_0||h|| \left(\frac{1}{M} \mathcal{E}(|h_1|) + \mathcal{E}(|h_1 \mu(B_1)|) \right) \tag{6.70}$$

and the result follows on taking $C_1 = (3k+2)C_0$. □

6.5 An asymptotic upper bound on $\mathrm{Var}(H_M)$

Lemma 6.6. *For all $M \geq 4$,*

$$|\mathcal{E}(h_1^* h_2^*)| \leq \frac{C_2||h||}{M} \mathcal{E}(|h_1|^2)^{1/2} \tag{6.71}$$

where

$$C_2 = (k+2)C_1 = 4(k+2)(3k+2)(1 + k(k+1)K(m)) \tag{6.72}$$

Proof. By the Cauchy–Schwarz inequality,

$$\mathcal{E}(|h_1\mu(B_1)|) \leq \mathcal{E}(|h_1|^2)^{1/2}\mathcal{E}(\mu(B_1)^2)^{1/2} \tag{6.73}$$

Taking $\alpha = 2$ in Lemma 6.2 we get

$$\mathcal{E}(\mu(B_1)^2) = \frac{k(k+1)}{M(M+1)} \leq \left(\frac{k+1}{M}\right)^2 \tag{6.74}$$

so by (6.73) it follows that

$$\mathcal{E}(|h_1\mu(B_1)|) \leq \frac{k+1}{M}\mathcal{E}(|h_1|^2)^{1/2} \tag{6.75}$$

Furthermore, $\mathrm{Var}(|h_1|) = \mathcal{E}(|h_1|^2) - \mathcal{E}(|h_1|)^2 \geq 0$ implies that $\mathcal{E}(|h_1|) \leq \mathcal{E}(|h_1|^2)^{1/2}$ so by Lemma 6.5,

$$\begin{aligned} |\mathcal{E}(h_1^* h_2^*)| &\leq C_1||h||\left(\frac{1}{M}\mathcal{E}(|h_1|) + \mathcal{E}(|h_1\mu(B_1)|)\right) \\ &\leq C_1||h||\left(\frac{1}{M}\mathcal{E}(|h_1|^2)^{1/2} + \frac{k+1}{M}\mathcal{E}(|h_1|^2)^{1/2}\right) \\ &\leq \frac{C_1(k+2)||h||}{M}\mathcal{E}(|h_1|^2)^{1/2} \end{aligned} \tag{6.76}$$

and the result follows on taking $C_2 = (k+2)C_1$. □

Thus we obtain the main result of the present chapter as follows.

Theorem 6.1.

$$\mathrm{Var}(H_M) \leq \frac{C_3||h||}{M}\mathcal{E}(|h_1|^2)^{1/2} \tag{6.77}$$

where

$$C_3 = 4 + C_2 = 4(1 + (k+2)(3k+2)(1 + k(k+1)K(m))) \tag{6.78}$$

Proof. Recall from (6.5) that

$$\mathrm{Var}(H_M) \leq \frac{1}{M}|\mathcal{E}(h_1^{*2})| + |\mathcal{E}(h_1^* h_2^*)| \tag{6.79}$$

By Lemma 6.1,

$$|\mathcal{E}(h_1^{*2})| \leq 4||h||\mathcal{E}(|h_1|) \leq 4||h||\mathcal{E}(|h_1|^2)^{1/2} \tag{6.80}$$

and by Lemma 6.6,

$$|\mathcal{E}(h_1^* h_2^*)| \leq \frac{C_2||h||}{M}\mathcal{E}(|h_1|^2)^{1/2} \tag{6.81}$$

Hence

$$\mathrm{Var}(H_M) \leq \frac{(C_2+4)||h||}{M}\mathcal{E}(|h_1|^2)^{1/2} \tag{6.82}$$

and the result follows on taking $C_3 = 4 + C_2$. □

6.6 A law of large numbers for $\delta_M(k)$

Using Chebyshev's inequality we obtain the following corollary of Theorem 6.1.

Corollary 6.1. *For every* $\epsilon > 0$

$$\mathbf{P}(|H_M - \mathcal{E}(H_M)| > \epsilon) \leq \frac{C_3||h||}{M\epsilon^2}\mathcal{E}(|h_1|^2)^{1/2} \tag{6.83}$$

If $||h|| < \infty$ then by Corollary 6.1 we see that the sample mean H_M converges in probability to its expected value as $M \to \infty$. Thus we have shown that bounded functions of a point and its kth nearest neighbour satisfy a *weak law of large numbers*. In particular, $\delta_M(k)$ is the sample mean of the random variables $h_i(X) = |X_{N[i,k]} - X_i|^2$ so by Corollary 6.1 it follows that δ_M satisfies the weak law of large numbers, i.e.

$$\frac{\delta_M(k)}{\mathcal{E}(|X_{N[i,k]} - X_i|^2)} \to 1 \qquad \text{in probability as} \qquad M \to \infty \tag{6.84}$$

To quantify the rate of the probabilistic convergence in Corollary 6.1 let $\kappa > 0$ and define

$$\epsilon = \frac{\mathcal{E}(|h_1|^2)^{1/4}}{M^{1/2-\kappa}} \tag{6.85}$$

By Corollary 6.1 we get

$$\mathbf{P}\left(|H_M - \mathcal{E}(H_M)| > \frac{\mathcal{E}(|h_1|^2)^{1/4}}{M^{1/2-\kappa}}\right) \leq \frac{C_3||h||}{M^{2\kappa}} \tag{6.86}$$

and since C_3 and $||h||$ are bounded independently of M it follows that

$$H_M = \mathcal{E}(H_M) + O\left(\frac{\mathcal{E}(|h_1|^2)^{1/4}}{M^{1/2-\kappa}}\right) \tag{6.87}$$

with probability greater than $1 - O(1/M^{2\kappa})$ as $M \to \infty$. Since $\mathcal{E}(|h_1|^2)^{1/4} \leq ||h||^{1/2}$ this implies that

$$H_M = \mathcal{E}(H_M) + O\left(\frac{1}{M^{1/2-\kappa}}\right) \tag{6.88}$$

with probability greater than $1 - O(1/M^{2\kappa})$ as $M \to \infty$.

However, if the function h is such that the expected value of $|h_i| = |h(X_i, X_{N[i,k]})|$ converges to zero as $M \to \infty$, then (6.87) incorporates this to give a faster probabilistic rate of convergence of the sample mean H_M to its expected value $\mathcal{E}(H_M)$ as $M \to \infty$. In particular, if the random sample $X = (X_1, \ldots, X_M)$ satisfies the conditions of Theorem 3.2 then

$$\delta_M(k) = \mathcal{E}(|X_{N[i,k]} - X_i|^2) + O\left(\frac{1}{M^{1/2+1/m-\kappa}}\right) \tag{6.89}$$

with probability greater than $1 - O(1/M^{2\kappa})$ as $M \to \infty$.

6.7 An asymptotic lower bound for the Travelling Salesman Problem

The kth nearest neighbour graph of a point set $X_1, \ldots, X_M$ is defined to be the graph obtained by including an edge between each point X_i and its kth nearest neighbour $X_{N[i,k]}$. Taking $\alpha = 1$ in Theorem 3.2 provides an interesting asymptotic expression for the expected length $L_M(\mathcal{N}_k)$ of the kth nearest neighbour graph of M points selected from $C \subset \mathbb{R}^m$ according to some sampling distribution function whose density ϕ is smooth and strictly positive over C.

Corollary 6.2. *Subject to the conditions of Theorem 3.2 we have*

$$\frac{\mathcal{E}(L_M(\mathcal{N}_k))}{M^{1-1/m}} \to c(m,k,\phi) \tag{6.90}$$

in probability as $M \to \infty$, where

$$c(m,k,\phi) = V_m^{-1/m} \frac{\Gamma(k+1/m)}{\Gamma(k)} \int_C \phi(x)^{1-1/m}\, dx \tag{6.91}$$

is a constant not depending on M.

By Corollary 6.1 it follows that $L_M(\mathcal{N}_k)/\mathcal{E}(L_M(\mathcal{N}_k)) \to 1$ in probability as $M \to \infty$ and hence

$$\frac{L_M(\mathcal{N}_k)}{M^{1-1/m}} \to c(m,k,\phi) \tag{6.92}$$

in probability as $M \to \infty$. For a uniform distribution with $m = 2$ and $k = 1$ this gives

$$\frac{L_M(\mathcal{N}_1)}{\sqrt{M}} \sim 0.5 \tag{6.93}$$

in probability as $M \to \infty$. This provides an interesting asymptotic lower bound for the minimal tour length of the random geometric Travelling Salesman Problem (TSP) on the unit square. If ϕ is uniform then [Beardwood *et al.* 1959] prove that the optimal tour length $L_M(\mathcal{T})$ of a set of independently distributed points in $[0,1]^m$ satisfies

$$\frac{L_M(\mathcal{T})}{M^{1-1/m}} \sim \beta \quad \text{as} \quad M \to \infty \tag{6.94}$$

for some constant $\beta > 0$, where the convergence is with probability one. For $m = 2$ the early estimate by [Stein 1977] of $\beta = 0.765$, derived empirically for relatively small TSP problems, was found to be too large (see [Valenzuela and Jones 1997]), and [Johnson *et al.* 1996] estimate $\beta = 0.7124$.

Since $L_M(\mathcal{T}) \geq L_M(\mathcal{N})$ then (6.93) shows that $\beta \geq 0.5$. While this is not particularly good as an estimate for the minimal tour length of the TSP on the unit square, it is interesting to observe that the method of proof is quite different from that presented in

[Beardwood *et al.* 1959]. Furthermore, it indicates that the optimal tour in a random geometric TSP is likely to contain a high proportion of edges that are not first nearest neighbour links.

Experimental evidence for the random geometric TSP suggests that a near–optimal tour can be constructed with with edges chosen from the associated kth nearest neighbour graphs for $1 \leq k \leq 20$. These results raise interesting questions regarding the distribution of the maximum value of such k, and may suggest new heuristics for the random geometric TSP which use only sets of kM edges ($1 \leq k \leq 20$), rather than the complete set of M^2 edges.

6.8 The asymptotic length of the k-nearest neighbours graph

In [Yukich 1998] the k-nearest neighbours graph of a point set $X_1, \ldots, X_M$ is defined to be the graph obtained by including an edge between each point X_i and its first k nearest neighbours $X_{N[i,1]}, \ldots, X_{N[i,k]}$. Theorem 8.3 of [Yukich 1998] states that if $X_1, \ldots, X_M$ are independent and identically distributed random variables with values in $[0,1]^m$ for $m \geq 2$, and if $N(k; X_1, \ldots, X_M)$ is the length of the k-nearest neighbours graph of $X_1, \ldots, X_M$, then

$$\lim_{M\to\infty} N(k; X_1, \ldots, X_M)/M^{(m-1)/m} = c(m,k) \int_{[0,1]^m} \phi(x)^{(m-1)/m}\, dx \tag{6.95}$$

where $c(k,m)$ is a constant not depending on M the convergence is complete(see [Yukich 1998]). The method of proof used in [Yukich 1998] is based on techniques first used in [Beardwood *et al.* 1959] and later extended by [Steele 1981]. We now show that (6.95) can also be obtained using the methods developed in this thesis.

By definition,

$$N(k; X_1, \ldots, X_M) = \sum_{j=1}^{k} \sum_{i=1}^{M} |X_{N[i,j]} - X_i| \tag{6.96}$$

Taking $h_i(X) = |X_{N[i,j]} - X_i|$ in (6.87),

$$\frac{1}{M} \sum_{i=1}^{M} |X_{N[i,j]} - X_i| = \mathcal{E}(|X_{N[i,j]} - X_i|) + O\left(\frac{\mathcal{E}(|X_{N[i,j]} - X_i|^2)^{1/4}}{M^{1/2-\kappa}} \right) \tag{6.97}$$

Taking $\alpha = 1$ and $\alpha = 2$ in Theorem 3.2, for every $\rho > 0$ we have that

$$\mathcal{E}(|X_{N[i,j]} - X_i|) = \frac{c'(m,j)}{M^{1/m}} + O\left(\frac{1}{M^{2/m-\rho}} \right) \quad \text{as} \quad M \to \infty \tag{6.98}$$

$$\mathcal{E}(|X_{N[i,j]} - X_i|^2) = O\left(\frac{1}{M^{2/m}} \right) \quad \text{as} \quad M \to \infty \tag{6.99}$$

where

$$c'(m,j) = V_m^{-1/m} \frac{\Gamma(j+1/m)}{\Gamma(j)} \int_{[0,1]^m} \phi(x)^{(m-1)/m}\, dx \tag{6.100}$$

By (6.97) it thus follows that for all $\kappa > 0$,

$$\frac{1}{M}\sum_{i=1}^{M} |X_{N[i,j]} - X_i| = \frac{c'(m,j)}{M^{1/m}} + O\left(\frac{1}{M^{1/2+1/2m-\kappa}}\right) \quad \text{as} \quad M \to \infty \tag{6.101}$$

and multiplying both sides of (6.101) by $M^{1/m}$ we obtain

$$\frac{1}{M^{(m-1)/m}}\sum_{i=1}^{M} |X_{N[i,j]} - X_i| = c'(m,j) + O\left(\frac{1}{M^{1/2-1/2m-\kappa}}\right) \quad \text{as} \quad M \to \infty \tag{6.102}$$

Hence by (6.96) and (6.100) we obtain

$$\begin{aligned} N(k; X_1, \ldots, X_M)/M^{(m-1)/m} = c(m,k) &\int_{[0,1]^m} \phi(x)^{(m-1)/m}\, dx \\ &+ O\left(\frac{1}{M^{1/2-1/2m-\kappa}}\right) \quad \text{as} \quad M \to \infty \end{aligned} \tag{6.103}$$

where

$$c(m,k) = V_m^{-1/m} \sum_{j=1}^{k} \Gamma(j+1/m)/\Gamma(j) \tag{6.104}$$

For all $m \geq 2$ we have therefore shown that

$$\lim_{M\to\infty} N(k; X_1, \ldots, X_M)/M^{(m-1)/m} = c(m,k) \int_{[0,1]^m} \phi(x)^{(m-1)/m}\, dx \tag{6.105}$$

where the convergence is in probability. Although convergence in probability is weaker than the complete convergence of Theorem 8.3 of [Yukich 1998], our result has the advantage of providing an explicit value for the constant $c(m,k)$. Furthermore, our result provides the asymptotic order of magnitude of the error term and hence the rate at which $N(k; X_1, \ldots, X_M)/M^{(m-1)/m}$ converges as $M \to \infty$.

6.9 Summary

In this chapter we have extended the ideas of [Bickel and Breiman 1983] to deal with sums involving bounded functions $h_i(X) = h(X_{N[i,k]}, X_i)$ of a point and its kth nearest neighbour. In Theorem 6.1 we have established an upper bound on the variance of such sums – this will be required for the proof of the Gamma test in Chapter 7. We

have then used this result to show that the mean squared distance $\delta_M(k)$ between kth nearest neighbours in a set of M points satisfies the (weak) law of large numbers as $M \to \infty$.

With considerable further effort it may be possible to develop a Central Limit theorem for sums of such functions, again following the ideas of [Bickel and Breiman 1983]. Furthermore, one could no doubt develop a similar theory for sums of bounded functions of the form $h_i(X) = h(X_{N[i,k]}, X_{N[i,k-1]}, \ldots, X_i)$.

All the required tools for a proof of the Gamma test have now been assembled, and in the next chapter we shall put these results together.

Chapter 7

Proof of the Gamma test

7.1 Introduction

We now proceed to the proofs of Theorems 1.1, 1.2 and 1.3. Recall (2.20) which states that

$$\gamma_M(k) = \mathrm{Var}(r) + A(M,k)\delta_M(k) + o(\delta_M(k)) + A_M(k) + B_M(k) + C_M(k) \tag{7.1}$$

To prove Theorem 1.1 we must show that

$$\gamma_M(k) = \mathrm{Var}(r) + A(M,k)\delta_M(k) + o(\delta_M(k)) + O\left(\frac{1}{M^{1/2-\kappa}}\right) \tag{7.2}$$

in probability as $M \to \infty$.

In Lemma 2.4 we have shown that the expected value of each of the terms $A_M(k)$, $B_M(k)$ and $C_M(k)$ is zero. To make use of Chebyshev's inequality we now apply Theorem 5.1 and Theorem 6.1 to obtain (probabilistic) upper bounds on $A_M(k)$, $B_M(k)$ and $C_M(k)$ in terms of the number of points M.

7.2 Upper bounds on variance

Lemma 7.1 (The noise term). *The variance of term $A_M(k)$ satisfies*

$$\mathrm{Var}(A_M(k)) \leq \frac{(kK(m)+1)(\mathcal{E}(r^4) + \mathrm{Var}(r)^2)}{M} \tag{7.3}$$

where $K(m)$ is the maximum kissing number in $\mathbb{R}^m$.

Proof. Define the random variables

$$g_i(R) = \frac{1}{2}(R_{N[i,k]} - R_i)^2 - \mathrm{Var}(r) \tag{7.4}$$

on the space of random noise samples $\mathbb{R}^M$ so that

$$A_M(k) = \frac{1}{M}\sum_{i=1}^{M} g_i(R) \tag{7.5}$$

By Theorem 4.1, $(g_1(R), \ldots, g_M(R))$ is a random sample of identically distributed L–dependent random variables with $L = 2kK(m) + 1$. By Theorem 5.1 it thus follows that

$$\mathrm{Var}(A_M(k)) \leq \frac{2(kK(m)+1)\,\mathrm{Var}(g_i)}{M} \tag{7.6}$$

By hypothesis, R_i and $R_{N[i,k]}$ are independent and identically distributed having expected value zero so

$$\mathcal{E}((R_{N[i,k]} - R_i)^2) = \mathcal{E}(R_{N[i,k]}^2) - 2\mathcal{E}(R_{N[i,k]})\mathcal{E}(R_i) + \mathcal{E}(R_i^2) = 2\mathcal{E}(R_i^2) = 2\,\mathrm{Var}(r) \tag{7.7}$$

and hence $\mathcal{E}(g_i) = 0$. Thus it follows that $\mathrm{Var}(g_i) = \mathcal{E}(g_i^2)$ which we write as

$$\begin{aligned}
\mathrm{Var}(g_i) &= \mathcal{E}\left(\frac{1}{4}(R_{N[i,k]} - R_i)^4 - (R_{N[i,k]} - R_i)^2\,\mathrm{Var}(r) + \mathrm{Var}(r)^2\right) \\
&= \mathcal{E}\left(\frac{1}{4}(R_{N[i,k]} - R_i)^4\right) - \mathrm{Var}(r)^2 \\
&= \frac{1}{4}\mathcal{E}\left(R_{N[i,k]}^4 - 4R_{N[i,k]}^3 R_i + 6R_{N[i,k]}^2 R_i^2 - 4R_{N[i,k]}R_i^3 + R_i^4\right) - \mathrm{Var}(r)^2
\end{aligned} \tag{7.8}$$

By hypothesis, $\mathcal{E}(R_i) = 0$, $\mathcal{E}(R_i^3) < \infty$ and R_i and $R_{N[i,k]}$ are independent so

$$\begin{array}{rcccl}
\mathcal{E}(R_{N[i,k]}^3 R_i) & = & \mathcal{E}(R_{N[i,k]}^3)\mathcal{E}(R_i) & = & 0 \\
\mathcal{E}(R_{N[i,k]} R_i^3) & = & \mathcal{E}(R_{N[i,k]})\mathcal{E}(R_i^3) & = & 0
\end{array} \tag{7.9}$$

Furthermore, since R_i and $R_{N[i,k]}$ are independent and identically distributed we have that

$$\mathcal{E}(R_{N[i,k]}^2 R_i^2) = \mathcal{E}(R_{N[i,k]}^2)\mathcal{E}(R_i^2) = \mathrm{Var}(r)^2 \tag{7.10}$$

Thus, since $\mathcal{E}(R_{N[i,k]}^4) = \mathcal{E}(R_i^4) = \mathcal{E}(r^4)$ we obtain

$$\mathrm{Var}(g_i) = \frac{1}{2}(\mathcal{E}(r^4) + \mathrm{Var}(r)^2) < \infty \tag{7.11}$$

and by (7.6) it follows that

$$\mathrm{Var}(A_M(k)) \leq \frac{(kK(m)+1)(\mathcal{E}(r^4) + \mathrm{Var}(r)^2)}{M} \tag{7.12}$$

as required. □

Lemma 7.2 (The mixed term). *The variance of term $B_M(k)$ satisfies*

$$\mathrm{Var}(B_M(k)) \leq \frac{4(kK(m)+1)\,\mathrm{Var}(r)c_1^2(b_1+c_1b_2)^2}{M} \tag{7.13}$$

where $K(m)$ is the maximum kissing number in $\mathbb{R}^m$.

Proof. Define the random variables

$$\begin{aligned} g_i(R) &= R_{N[i,k]} - R_i & (7.14)\\ h_i(X) &= (X_{N[i,k]} - X_i)\cdot\nabla f(X_i) + T_f(X_i, X_{N[i,k]}) & (7.15) \end{aligned}$$

on $\mathbb{R}^M$ and C^M respectively so that

$$B_M(k) = \frac{1}{M}\sum_{i=1}^{M} g_i(R)h_i(X) \tag{7.16}$$

By Lemma 2.4 we have that $\mathcal{E}(B_M(k)) = 0$ so $\mathrm{Var}(B_M(k)) = \mathcal{E}(B_M(k)^2)$. Taking the expectation of $B_M(k)^2$ over all pairs $(X,R) \in C^M \times \mathbb{R}^M$ we obtain

$$\mathcal{E}(B_M(k)^2) = \frac{1}{M^2}\sum_{i,j=1}^{M} \mathcal{E}(h_i(X)h_j(X)g_i(R)g_j(R)) \tag{7.17}$$

By hypothesis both $h_i(X)$ and $h_j(X)$ are completely independent of R, while $g_i(R)$ and $g_j(R)$ may depend on X. Thus we write

$$\mathcal{E}(B_M(k)^2) = \frac{1}{M^2}\sum_{i,j=1}^{M} \mathcal{E}_\phi\big(h_i(X)h_j(X)\mathcal{E}_\psi(g_i(R)g_j(R)\big|X)\big) \tag{7.18}$$

By Theorem 4.1 and the discussion contained in Chapter 2, for any $X \in C^M$ the random variables $g_1(R),\ldots,g_M(R)$ are L–dependent with $L = 2(kK(m)+1)$ where $K(m)$ is the maximum kissing number in $\mathbb{R}^m$. Hence, for any fixed i it follows that $\mathcal{E}_\psi(g_i(R)g_j(R)\big|X) \neq 0$ for at most L indices j. Furthermore, since $|ab| \leq \frac{1}{2}(a^2+b^2)$ for any pair of real numbers a and b and since $g_i(R)$ and $g_j(R)$ are identically distributed we have that

$$|\mathcal{E}_\psi(g_i(R)g_j(R)\big|X)| \leq \mathcal{E}_\psi(|g_i(R)g_j(R)|\big|X) \leq \mathcal{E}_\psi(g_i^2(R)\,\big|\,X) = 2\,\mathrm{Var}(r) < \infty \tag{7.19}$$

and since $|h_i(X)| \leq c_1b_1 + c_1^2b_2 < \infty$ for each $X \in C^M$ it thus follows that

$$\mathcal{E}(B_M(k)^2) \leq \frac{4(kK(m)+1)c_1^2(b_1+c_1b_2)^2\,\mathrm{Var}(r)}{M} \tag{7.20}$$

as required. □

Lemma 7.3 (The distance term). *The variance of term $C_M(k)$ satisfies*

$$\mathrm{Var}(C_M(k)) \leq \frac{4(1+(k+2)(3k+2)(1+k(k+1)K(m)))c_1^4 b_1^4}{M} \tag{7.21}$$

where $K(m)$ is the maximum kissing number in $\mathbb{R}^m$.

Proof. Define the random variable

$$h_i(X) = \frac{1}{2}\left((X_{N[i,k}-X_i)\cdot\nabla f(X_i)\right)^2 - A(M,k)|X_{N[i,k]}-X_i|^2 \tag{7.22}$$

on the space of random point samples C^M where

$$A(M,k) = \frac{\mathcal{E}_\phi\left(\left((X_{N[i,k]}-X_i)\cdot\nabla f(X_i)\right)^2\right)}{2\mathcal{E}_\phi(|X_{N[i,k]}-X_i|^2)} \tag{7.23}$$

so that

$$C_M(k) = \frac{1}{M}\sum_{i=1}^{M} h_i(X) \tag{7.24}$$

Since

$$|(X_{N[i,k]}-X_i)\cdot\nabla f(X_i)|^2 \leq |X_{N[i,k]}-X_i|^2|\nabla f(X_i)|^2 \leq |X_{N[i,k]}-X_i|^2 b_1^2 \tag{7.25}$$

it follows by (7.23) that $0 \leq A(M,k) \leq \frac{1}{2}b_1^2$ so

$$|h_i(X)| \leq \frac{1}{2}c_1^2 b_1^2 + \frac{1}{2}c_1^2 b_1^2 = c_1^2 b_1^2 \qquad \text{for all} \quad X \in C^M \tag{7.26}$$

Hence

$$||h|| = \sup\{|h_i(X)| \,:\, X \in C^M\} \leq c_1^2 b_1^2 < \infty \tag{7.27}$$

and we can thus apply Theorem 6.1 to $C_M(k)$ so that

$$\mathrm{Var}(C_M(k)) \leq \frac{4(1+(k+2)(3k+2)(1+k(k+1)K(m)))c_1^2 b_1^2}{M}(\mathcal{E}(|h_1|^2))^{1/2} \tag{7.28}$$

where $K(m)$ is the maximum kissing number in $\mathbb{R}^m$. Finally, since $|h_1|^2 \leq c_1^4 b_1^4$ it follows that $\mathcal{E}(|h_1|^2) \leq c_1^4 b_1^4$ and hence $(\mathcal{E}(|h_1|^2))^{1/2} \leq c_1^2 b_1^2$ so we conclude that

$$\mathrm{Var}(C_M(k)) \leq \frac{4(1+(k+2)(3k+2)(1+k(k+1)K(m)))c_1^4 b_1^4}{M} \tag{7.29}$$

as required. □

7.3 Probabilistic upper bounds on $A_M(k)$, $B_M(k)$ and $C_M(k)$

Since the expected value of $A_M(k)$, $B_M(k)$ and $C_M(k)$ is zero then by Lemma 7.1, Lemma 7.2, Lemma 7.3 and Chebyshev's inequality we obtain the following.

Corollary 7.1. *For every $\epsilon > 0$,*

$$\begin{aligned}
\mathbf{P}(|A_M(k)| > \epsilon) &\leq \frac{\lambda_A}{M\epsilon^2} \\
\mathbf{P}(|B_M(k)| > \epsilon) &\leq \frac{\lambda_B}{M\epsilon^2} \\
\mathbf{P}(|C_M(k)| > \epsilon) &\leq \frac{\lambda_C}{M\epsilon^2}
\end{aligned}$$

where

$$\begin{aligned}
\lambda_A &= (kK(m)+1)(\mathcal{E}(r^4) + \mathrm{Var}(r)^2) \\
\lambda_B &= 4(kK(m)+1)\,\mathrm{Var}(r)c_1^2(b_1+c_1b_2)^2 \\
\lambda_C &= 4(1+(k+2)(3k+2)(1+k(k+1)K(m)))c_1^4b_1^4
\end{aligned}$$

are finite constants not depending on M.

7.4 Proof of Theorem 1.1

The following result enables us to assemble the results of Corollary 7.1 into a proof of Theorem 1.1.

Lemma 7.4 (The transitive lemma). *Let X and Y be two random variables and suppose for every $\epsilon_1, \epsilon_2 > 0$ there exist $\eta_1, \eta_2 > 0$ such that*

$$\mathbf{P}(|X| > \epsilon_1) < \eta_1 \quad \textit{and} \quad \mathbf{P}(|Y| > \epsilon_2) < \eta_2 \tag{7.30}$$

where $\eta_1, \eta_2 \to 0$ as $\epsilon_1, \epsilon_2 \to 0$. Then

$$\mathbf{P}(|X \pm Y| > \epsilon_1 + \epsilon_2) < \eta_1 + \eta_2 \tag{7.31}$$

Proof. Denote the events $|X| > \epsilon_1$ and $|Y| > \epsilon_2$ by A and B respectively so that $\mathbf{P}(A) < \eta_1$ and $\mathbf{P}(B) < \eta_2$. Note that we do not assume the events A and B to be independent. Let $C(A)$ and $C(B)$ denote the complement of A and B respectively and consider the mutually exclusive and exhaustive set of events $A \cap B$, $C(A) \cap B$, $A \cap C(B)$ and $C(A) \cap C(B)$ illustrated in Figure 7.1.

First of all,

$$\mathbf{P}(A \cap B) + \mathbf{P}(C(A) \cap B) + \mathbf{P}(A \cap C(B)) + \mathbf{P}(C(A) \cap C(B)) = 1 \tag{7.32}$$

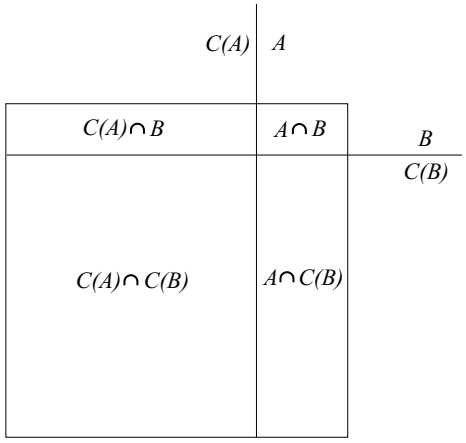

Figure 7.1: If $C(A)$ and $C(B)$ are highly probable then so is $C(A) \cap C(B)$, even if these events are not independent.

and by hypothesis we have that

$$\begin{aligned} \mathbf{P}(A \cap B) + \mathbf{P}(A \cap C(B)) &= \mathbf{P}(A) < \eta_1 && (7.33) \\ \mathbf{P}(A \cap B) + \mathbf{P}(C(A) \cap B) &= \mathbf{P}(B) < \eta_2 && (7.34) \end{aligned}$$

Thus

$$\begin{aligned} \mathbf{P}(C(A) \cap C(B)) &= 1 - \mathbf{P}(A \cap C(B)) - \mathbf{P}(C(A) \cap B) - \mathbf{P}(A \cap B) \\ &= 1 - \mathbf{P}(A) - \mathbf{P}(B) + \mathbf{P}(A \cap B) \qquad (7.35) \\ &\geq 1 - \eta_1 - \eta_2 \end{aligned}$$

By the triangle inequality, the event $C(A) \cap C(B)$ implies the event D defined by $|X \pm Y| \leq \epsilon_1 + \epsilon_2$. From (7.35) it thus follows that $\mathbf{P}(D) \geq 1 - \eta_1 - \eta_2$ and hence $\mathbf{P}(C(D)) < \eta_1 + \eta_2$, i.e.

$$\mathbf{P}(|X \pm Y| > \epsilon_1 + \epsilon_2) < \eta_1 + \eta_2 \qquad (7.36)$$

as required. □

Lemma 7.5. *For any $\epsilon > 0$,*

$$\mathbf{P}(|A_M(k) + B_M(k) + C_M(k)| > \epsilon) \leq \frac{\lambda}{M\epsilon^2} \qquad (7.37)$$

where

$$\lambda = 9(\lambda_A + \lambda_B + \lambda_C) \qquad (7.38)$$

is a finite constant not depending on M.

Proof. Replacing ϵ by $\epsilon/3$ in Corollary 7.1 and applying Lemma 7.4 to the pair $A_M(k)$ and $B_M(k)$, we obtain

$$\mathbf{P}\left(|A_M(k) + B_M(k)| > \frac{\epsilon}{3} + \frac{\epsilon}{3}\right) \leq \frac{\lambda_A}{M(\epsilon/3)^2} + \frac{\lambda_B}{M(\epsilon/3)^2} \qquad (7.39)$$

which is equivalent to

$$\mathbf{P}\left(|A_M(k) + B_M(k)| > \frac{2\epsilon}{3}\right) \leq \frac{9(\lambda_A + \lambda_B)}{M\epsilon^2} \tag{7.40}$$

A further application of Lemma 7.4 to the pair $A_M(k) + B_M(k)$ and $C_M(k)$ then leads to

$$\mathbf{P}(|A_M(k) + B_M(k) + C_M(k)| > \epsilon) \leq \frac{9(\lambda_A + \lambda_B + \lambda_C)}{M\epsilon^2} \tag{7.41}$$

as required. □

By Lemma 7.5 we obtain the following

Corollary 7.2. *For any $\epsilon > 0$,*

$$\mathbf{P}(\left|\gamma_M(k) - \left(\mathrm{Var}(r) + A(M,k)\delta_M(k) + o(\delta_M(k))\right)\right| > \epsilon) \leq \frac{\lambda}{M\epsilon^2} \tag{7.42}$$

as $M \to \infty$ where λ is a finite constant not depending on M.

Proof. By (2.20),

$$\gamma_M(k) - \left(\mathrm{Var}(r) + A(M,k)\delta_M(k) + o(\delta_M(k))\right) = A_M(k) + B_M(k) + C_M(k) \tag{7.43}$$

and the result follows by Lemma 7.5 □

Proof of Theorem 1.1

Proof. Let $\kappa > 0$ and apply Corollary 7.2 with

$$\epsilon = \frac{1}{M^{1/2-\kappa}} \tag{7.44}$$

Then

$$\mathbf{P}\left(\left|\gamma_M(k) - \left(\mathrm{Var}(r) + A(M,k)\delta_M(k) + o(\delta_M(k))\right)\right| > \frac{1}{M^{1/2-\kappa}}\right) \leq \frac{\lambda}{M^{2\kappa}} \tag{7.45}$$

and hence

$$\mathbf{P}\left(\left|\gamma_M(k) - \left(\mathrm{Var}(r) + A(M,k)\delta_M(k) + o(\delta_M(k))\right)\right| < \frac{1}{M^{1/2-\kappa}}\right) \geq 1 - \frac{\lambda}{M^{2\kappa}} \tag{7.46}$$

Thus, for every $\kappa > 0$

$$\gamma_M(k) = \mathrm{Var}(r) + A(M,k)\delta_M(k)) + o(\delta_M(k)) + O\left(\frac{1}{M^{1/2-\kappa}}\right) \tag{7.47}$$

with probability greater than $1 - O(1/M^{2\kappa})$ as $M \to \infty$, as required. □

7.5 Proof of Theorem 1.2

Before addressing the proof of Algorithm 1 we first consider whether or not the gradients $A(M,k)$ are independent of the near neighbour index k. If not, then at first sight the linear regression technique employed by Algorithm 1 is not justified.

However, subject to a fairly weak condition on the asymptotic behaviour of nearest neighbour distances we show that any k–dependence regarding the $A(M,k)$ can be tolerated by proving that the intercept Γ computed by the Gamma test is *approximately* equal to $\mathrm{Var}(r)$ with an error that converges to zero as $M \to \infty$.

We aim to prove Theorem 1.2 which states that, subject to the condition that for some fixed $p \geq 1$ there exists a positive constant $c < 1$ with

$$\delta_M(1) \leq c\delta_M(p) \qquad \text{for all} \quad M \geq 1 \tag{7.48}$$

then the number Γ returned by Algorithm 1 converges in probability to $\mathrm{Var}(r)$ as $M \to \infty$.

In order to justify the linear regression technique employed by Algorithm 1, some spread of the points $\delta_M(1), \ldots, \delta_M(p)$ is clearly required. Condition (7.48) is a weak requirement in this direction – it ensures that $\delta_M(p) - \delta_M(1) \geq (1/c - 1)\delta_M(1)$ where $(1/c - 1) > 0$. In particular, if there exists some $p > 1$ such that the mean squared distance to the pth nearest neighbour is at least *twice* the mean squared distance to the first nearest neighbour, then (7.48) is satisfied with $c = 1/2$. In fact, it is difficult to imagine a set of points that are distributed in such a way that some p having this property does not exist.

Note that if (7.48) fails, we can still use Theorem 1.1 to estimate $\mathrm{Var}(r)$ by performing the crude Gamma test algorithm on $(\delta_M(1), \gamma_M(1))$ for increasing M.

In order to prove Theorem 1.2 we need the following lemma.

Lemma 7.6. *For $1 \leq k \leq p$ let $Y_k = v + a_k X_k + \Delta_k$ where $0 < X_1 \leq \ldots \leq X_p$ are such that $X_1 < X_p$, and let $Y = d + cX$ be the least squares regression line for the points (X_k, Y_k). Then*

$$|d - v| \leq p(p-1)(A_{\max} X_p + \Delta_{\max}) \left(\frac{X_p}{X_p - X_1} \right) \tag{7.49}$$

where $A_{\max} = \max |a_k|$ and $\Delta_{\max} = \max |\Delta_k|$ for $1 \leq k \leq p$.

Proof. The values c and d are defined to be those that minimise the function

$$F(c,d) = \sum_{k=1}^{p} \left((d + cX_k) - (v + a_k X_k + \Delta_k) \right)^2 \tag{7.50}$$

Setting the partial derivatives of $F(c,d)$ to zero leads to

$$\begin{aligned} 2\sum_{k=1}^{p}\big((d+cX_k)-(v+a_kX_k+\Delta_k)\big)X_k=0 \\ 2\sum_{k=1}^{p}\big((d+cX_k)-(v+a_kX_k+\Delta_k)\big)=0 \end{aligned} \tag{7.51}$$

from which we obtain

$$\begin{aligned} c\sum_{k=1}^{p}X_k^2+d\sum_{k=1}^{p}X_k=\sum_{k=1}^{p}(v+a_kX_k+\Delta_k)X_k \\ c\sum_{k=1}^{p}X_k+d\sum_{k=1}^{p}1=\sum_{k=1}^{p}(v+a_kX_k+\Delta_k) \end{aligned} \tag{7.52}$$

Let $S=\sum_{k=1}^{p}X_k$ and $T=\sum_{k=1}^{p}X_k^2$ so that

$$\begin{aligned} cT+dS=vS+\sum_{k=1}^{p}(a_kX_k+\Delta_k)X_k \\ cS+dp=vp+\sum_{k=1}^{p}(a_kX_k+\Delta_k) \end{aligned} \tag{7.53}$$

Solving for d we obtain

$$d(S^2-pT)=v(S^2-pT)+S\sum_{k=1}^{p}(a_kX_k+\Delta_k)X_k-T\sum_{k=1}^{p}(a_kX_k+\Delta_k) \tag{7.54}$$

and hence

$$(d-v)(S^2-pT)=S\sum_{j=1}^{p}(a_jX_j+\Delta_j)X_j-T\sum_{j=1}^{p}(a_jX_j+\Delta_j) \tag{7.55}$$

Substituting for S and T,

$$(d-v)\sum_{i,j}X_i(X_i-X_j)=\sum_{i,j}X_i(X_i-X_j)(a_jX_j+\Delta_j) \tag{7.56}$$

and since

$$\begin{aligned} \sum_{i,j}X_i(X_i-X_j)&=\sum_{i<j}X_i(X_i-X_j)+\sum_{i>j}X_i(X_i-X_j) \\ &=\sum_{i<j}X_i(X_i-X_j)+\sum_{i<j}X_j(X_j-X_i) \\ &=\sum_{i<j}(X_i^2-2X_iX_j+X_j^2) \\ &=\sum_{i<j}(X_i-X_j)^2 \end{aligned} \tag{7.57}$$

it follows that

$$d - v = \frac{\sum_{i,j} X_i(X_i - X_j)(a_j X_j + \Delta_j)}{\sum_{i<j}(X_i - X_j)^2} \tag{7.58}$$

The denominator is the sum of positive terms and is therefore at least as large as its largest term $(X_p - X_1)^2$. By hypothesis, this term is *strictly positive* so the denominator is bounded below by $(X_p - X_1)^2 > 0$. Using the triangle inequality on the numerator we replace each term $X_i(X_i - X_j)$ by the maximum $X_p(X_p - X_1)$ of such terms, and each $|a_j X_j + \Delta_j|$ by $(A_{\max} X_p + \Delta_{\max})$. Since there are at most $p(p-1)$ non–zero terms in the numerator, the result follows on cancelling a factor of $X_p - X_1$. □

Proof of Theorem 1.2

Proof. By Theorem 1.1 we have that

$$\gamma_M(k) = \mathrm{Var}(r) + A(M,k)\delta_M(k) + o(\delta_M(k)) + O\left(\frac{1}{M^{1/2-\kappa}}\right) \tag{7.59}$$

with probability greater than $1 - O(1/M^{2\kappa})$ as $M \to \infty$. Furthermore, by condition (7.48) it follows that $\delta_M(1) < \delta_M(p)$. Hence the conditions of Lemma 7.6 are satisfied by the pairs $(X_k, Y_k) = (\delta_M(k), \gamma_M(k))$ with $a_k = A(M,k)$, $\Delta_k = o(\delta_M(k)) + O(1/M^{1/2-\kappa})$, $v = \mathrm{Var}(r)$ and $d = \Gamma$ so

$$|\Gamma - \mathrm{Var}(r)| \leq p(p-1)(A\delta_M(p) + \Delta)\left(\frac{\delta_M(p)}{\delta_M(p) - \delta_M(1)}\right) \tag{7.60}$$

where $A = \max_k(A(M,k))$ and $\Delta = o(\delta_M(p)) + O(1/M^{1/2-\kappa})$.

By condition (7.48), $\delta_M(p) - \delta_M(1) \geq (1-c)\delta_M(p)$ for some constant $c > 0$ so

$$\frac{\delta_M(p)}{\delta_M(p) - \delta_M(1)} = O(1) \quad \text{as} \quad M \to \infty \tag{7.61}$$

Furthermore, since $|A(M,k)| \leq \frac{1}{2}b_1^2 < \infty$ for every $1 \leq k \leq p$ we also have $A\delta_M(p) = O(\delta_M(p))$ as $M \to \infty$ so

$$|\Gamma - \mathrm{Var}(r)| = O(\delta_M(p)) + O(1/M^{1/2-\kappa}) \tag{7.62}$$

with probability greater than $1 - O(1/M^{2\kappa})$ as $M \to \infty$. Thus we conclude that

$$\Gamma \to \mathrm{Var}(r) \quad \text{in probability as} \quad M \to \infty \tag{7.63}$$

as required. □

7.6 Proof of Theorem 1.3

We obtain a proof of Theorem 1.3 by showing that condition (7.48) holds for sampling distributions which satisfy the conditions of Theorem 3.2.

Proof. Taking $\alpha = 2$ in Theorem 3.2, the expected value of $\delta_M(k)$ is given by

$$\mathcal{E}(\delta_M(k)) = \frac{c(m,k,\phi)}{M^{2/m}} \tag{7.64}$$

to first order in M, where

$$c(m,k,\phi) = V_m^{-2/m} \frac{\Gamma(k+2/m)}{\Gamma(k)} \int_C \phi(x)^{1-2/m}\, dx \tag{7.65}$$

Hence for all $m > 0$ we have that

$$\frac{\delta_M(1)}{\delta_M(p)} \sim \frac{\Gamma(1+2/m)\Gamma(p)}{\Gamma(p+2/m)} < 1 \quad \text{as} \quad M \to \infty \tag{7.66}$$

and the result follows by Theorem 1.2. □

7.7 The gradient $A(M,k)$ of the asymptotic linearity relation

The approximate asymptotic linearity relation between $\gamma_M(k)$ and $\delta_M(k)$ is expressed in Theorem 1.1 as

$$\gamma_M(k) \sim \mathrm{Var}(r) + A(M,k)\delta_M(k) \quad \text{as} \quad M \to \infty \tag{7.67}$$

where the convergence is in probability and $A(M,k)$ is defined by

$$A(M,k) = \frac{\mathcal{E}_\phi\left(\left((X_{N[i,k]} - X_i)\cdot \nabla f(X_i)\right)^2\right)}{2\mathcal{E}_\phi(|X_{N[i,k]} - X_i|^2)} < \infty \tag{7.68}$$

Let θ_i denote the angle between the vectors $\nabla f(X_i)$ and $X_{N[i,k]} - X_i$ so that

$$\left((X_{N[i,k]} - X_i)\cdot \nabla f(X_i)\right)^2 = |X_{N[i,k]} - X_i|^2 |\nabla f(X_i)|^2 \cos^2\theta_i \tag{7.69}$$

By (7.68) we see that $A(M,k)$ depends on the expectation $\mathcal{E}_\phi\left(|\nabla f(X_i)|^2 \cos^2\theta_i\right)$ and also on the degree of dependence between $|X_{N[i,k]} - X_i|$ and $|\nabla f(X_i)|^2 \cos^2\theta_i$.

Since the vector $X_{N[i,k]} - X_i$ depends only on the sampling distribution it is reasonable to suppose that its direction and magnitude are independent of the function f, in which case

$$A(M,k) = \frac{1}{2}\mathcal{E}_\phi(|\nabla f(X_i)|^2 \cos^2\theta_i) \tag{7.70}$$

Whether or not this is true depends on the directional properties of the sampling density ϕ about the points of C. If ϕ results in an *isotropic* distribution of density in local neighbourhoods of the points of C then the near neighbour distances and the near neighbour directions are independent and (7.70) holds.

Consider the case where C is of positive measure in $\mathbb{R}^m$ and the sampling density ϕ is smooth and strictly positive on C. As the number of points M becomes large the neighbourhoods about each interior point shrink and the local density becomes essentially constant and hence isotropic. Provided the boundary of C forms a set of measure zero then no problems arise. In this case the distribution of the θ_i is independent of k so given M sufficiently large, (7.70) is independent of k for all bounded k. In the case where C is a chaotic attractor of zero measure in $\mathbb{R}^m$ then again, if the local density is isotropic at all sufficiently small scales of measurement there are no problems. However, this cannot be assumed for an arbitrary chaotic attractor.

In addition to (7.70), suppose further that $|\nabla f(X_i)|^2$ is independent of $\cos^2\theta_i$ so that

$$A(M,k) = \frac{1}{2}\mathcal{E}_\phi(|\nabla f(X_i)|^2)\mathcal{E}_\phi(\cos^2\theta_i) \tag{7.71}$$

If $m = 1$ then θ_i takes only the values 0 and π and hence $\mathcal{E}_\phi(\cos^2\theta_i) = 1$. If $m \geq 2$ then one might assume that the θ_i are uniformly distributed over $[-\pi, \pi]$, in which case $\mathcal{E}_\phi(\cos^2\theta_i) = 1/2$. Asymptotically we then have

$$A(M,k) = \begin{cases} \frac{1}{2}\mathcal{E}_\phi(|\nabla f(X_i)|^2) & \text{if} \quad m = 1 \\ \frac{1}{4}\mathcal{E}_\phi(|\nabla f(X_i)|^2) & \text{if} \quad m \geq 2 \end{cases} \tag{7.72}$$

Under these circumstances the slope of the regression line computed by the Gamma test provides an estimate of the mean squared gradient of the unknown function f (relative to the sampling distribution) and may thus be considered as a crude indicator of the complexity of the surface defined by f.

7.8 Experimental results

7.8.1 The pth nearest neighbour condition

We examine whether condition (7.48) is likely to hold. To establish the validity of Algorithm 1 we need to show that the ratio $\delta_M(1)/\delta_M(p)$ is strictly less than 1 for all sufficiently large M.

Taking $p = 10$ we compute the ratios $\delta_M(1)/\delta_M(p)$ as M increases from 1000 to 100000 for a uniform distribution and from 1000 to 200000 for the Hénon map (3.165) with $a = 1.4$ and $b = 0.3$, both in steps of 1000. Figures 7.2 and 7.3 show the plots of

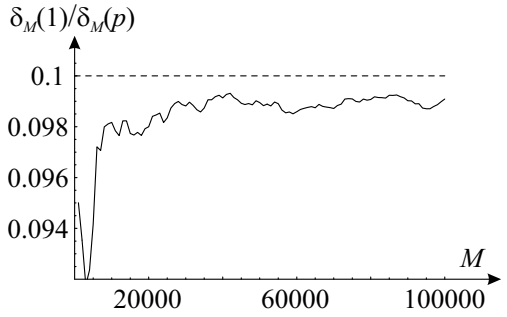

Figure 7.2: Ratio of $\delta_M(1)/\delta_M(p)$ as M increases for the uniform case ($p = 10$).

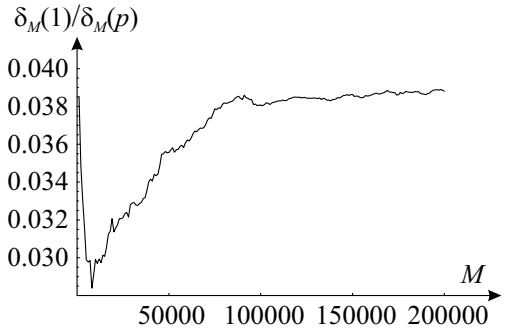

Figure 7.3: Ratio of $\delta_M(1)/\delta_M(p)$ as M increases for the Hénon Map ($p = 10$).

$\delta_M(1)/\delta_M(p)$ against M for the uniform distribution and the Hénon map respectively. The graphs suggest that the ratio $\delta_M(1)/\delta_M(p)$ remains bounded away from 1 in both cases.

If the sampling distribution Φ satisfies the conditions of Theorem 3.2, then taking $\alpha = 2$ in Theorem 3.2 and applying Corollary 6.1 we see that for all $m \geq 1$,

$$\frac{\delta_M(1)}{\delta_M(p)} \sim \frac{\Gamma(1 + 2/m)\Gamma(p)}{\Gamma(p + 2/m)} < 1 \tag{7.73}$$

with probability approaching one as $M \to \infty$. This ensures that (7.48) holds with probability approaching one as $M \to \infty$, and a further application of Lemma 7.4 in the proof of Theorem 1.2 then ensures that the conclusion of Theorem 1.2 holds under these conditions.

Taking $m = 2$ and $p = 10$ in (7.73) we see that $\delta_M(1)/\delta_M(p)$ in the uniform case should be asymptotically equal to 0.1 as $M \to \infty$, shown as a dashed line in Figure 7.2. It is interesting to note that although we have not been able to prove a version of Theorem 3.2 for sets of non–integral dimension, taking $m = 1.26$ for the Hénon map (which is approximately equal to the fractal dimension of its attractor) the theoretical value of $\delta_M(1)/\delta_M(p)$ for $p = 10$ as determined by (7.73) is equal to 0.035. This is close the experimental values shown in Figure 7.3.

7.8.2 Directional distributions

We now investigate whether we can replace the general definition of $A(M, k)$ given in (1.46) by the simpler definitions (7.70) or (7.72). If the direction of the nearest neighbour vectors are uniformly distributed then we may conclude that (7.72) holds.

For a set of $M = 100000$ points, first for a uniform distribution and secondly for the Hénon map (3.165) with $a = 1.4$ and $b = 0.3$, we compute the angle ω_i ($-\pi < \omega_i \leq \pi$) between the kth nearest neighbour vector $\boldsymbol{x}_{N[i,k]} - \boldsymbol{x}_i$ and the horizontal axis, then plot a histogram of the angle probabilities using 100 bins. The histograms for $k = 1$

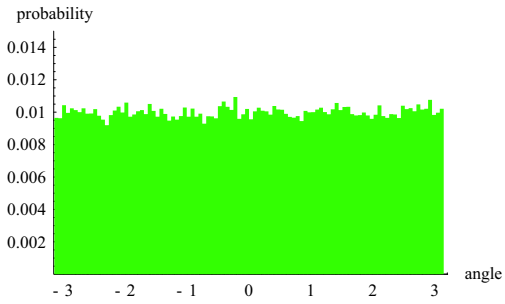

Figure 7.4: Angle histogram for the uniform distribution.

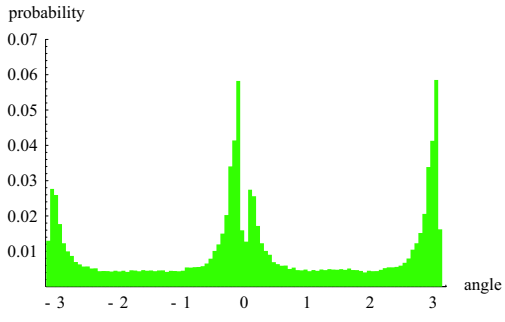

Figure 7.5: Angle histogram for the Hénon Map.

are shown in Figures 7.4 and 7.5 respectively. For $k = 2$ to 10 the histograms were virtually identical.

In the uniform case we see that the direction of the nearest neighbour vectors are approximately uniformly distributed in the range $[-\pi, \pi]$ and hence the (local) density is isotropic.

Figure 7.5 illustrates that the distribution of near neighbour directions on the Hénon map is anisotropic. There are two strongly preferred directions at $\omega \approx -7.2°$ and $\omega \approx 172.8°$ respectively, each with probabilities of approximately 0.058. There are also two subsidiary preferred directions at $\omega \approx -176.4°$ and $\omega \approx 3.6°$ respectively, each with probabilities of approximately 0.028. If this experiment is repeated with data localised to a small region of the attractor, the histogram can be markedly different although remaining anisotropic. Thus although anisotropy appears to be ubiquitous over the Hénon attractor, its precise nature is not scale invariant.

7.8.3 The gradients $A(M, k)$

We define the function $f : \mathbb{R}^2 \to \mathbb{R}$ to be $f(x, y) = x^2 + y^2$ so that $\nabla f = (2x, 2y)$, and generate values of $A(M, k)$ according to (1.46), substituting empirical means for expectations.

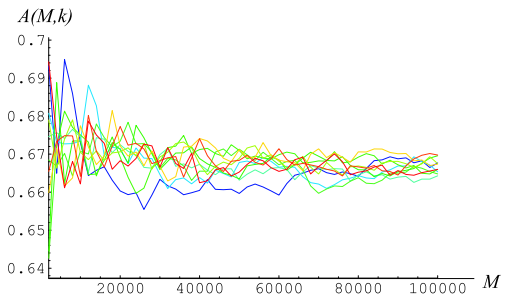

Figure 7.6: Graph of $A(M, k)$ as M increases for the uniform distribution.

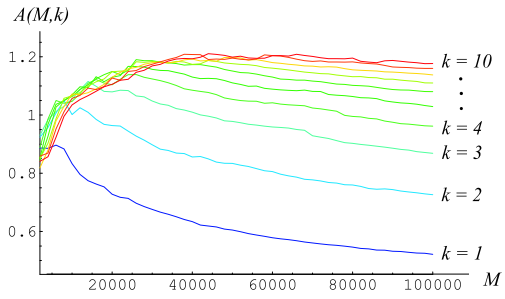

Figure 7.7: Graph of $A(M, k)$ as M increases for the Hénon Map.

For k in the range $1 \leq k \leq 10$ we compute $A(M,k)$ as M increases from 2000 to 100000 in steps of 2000. Figures 7.6 and 7.7 show plots of $A(M,k)$ against M for the uniform distribution and the Hénon map with $a = 1.4$ and $b = 0.3$ respectively.

In the uniform case the gradients $A(M,k)$ appear to be independent of k and converge to a stable value of approximately 0.67 as M increases. Note that

$$\int_0^1 \int_0^1 |\nabla(f(x,y))|^2 \, dx \, dy = \frac{8}{3} \tag{7.74}$$

so in this case the asymptotic value of the $A(M,k)$ is approximately equal to the expected value $\frac{1}{4}\mathcal{E}_\phi(|\nabla f|^2)$ over the input space $[0,1]^2$, which agrees with (7.72).

In the case of the Hénon map we see that the gradients $A(M,k)$ do indeed depend on k. We remark that each one appears to converge to a stable value as M increases.

7.9 Summary

The main goal of this work has now been accomplished. We have assembled the results developed in Chapters 3, 4, 5 and 6 to generate a proof of the Gamma test. As it stands this proof covers the case of a smooth positive sampling density over a compact convex body in $\mathbb{R}^m$, which should be adequate for many practical applications. We should note that the decomposition strategy of the proof means that we are unable to take account of cancellation of errors which may occur between the various sums. Thus the error terms in (7.62) represent a 'worst case' analysis that is often pessimistic in practical applications, for which the convergence may be much faster (see [Tsui 1999]).

It seems remarkable that such a simple and useful algorithm should require so much mathematical machinery in order to provide a formal justification but given the complexities one is forced to confront, it also seems unlikely that a very much simpler analysis can be established. On the contrary, one might regard this work as an initial foray into what promises to be a rich and complex set of questions as one seeks to further generalise the formal justification to cover the applications to chaotic dynamical systems. However, such work must be left for the future.

Fortunately there are now many interesting results which we can readily derive from the techniques so far developed. In the next chapter we shall apply these ideas to prove a potentially useful generalisation of the Gamma test first described in [Durrant 2001].

Chapter 8

The Extended Gamma test

8.1 Introduction

The Gamma test is a technique for estimating the second moment $\mathcal{E}(r^2)$ of the noise distribution Ψ. We now show that under the hypothesis of a *symmetric* noise distribution (i.e. where all odd moments are zero), similar ideas may be applied to estimate the higher order moments of the noise distribution. In some circumstances these estimates may be used to reconstruct the noise distribution itself.

The most computationally expensive aspect of the Gamma test algorithm is to compute the nearest neighbour lists of the input points, which can be achieved in a time of order $O(M \log M)$. As we shall see, the statistics used to estimate the higher order moments of the noise distribution are also defined relative to the nearest neighbour lists so computing these higher order estimates involves very little extra computational cost.

Let $m_j = \mathcal{E}(r^j)$ denote the jth moment of the noise distribution and suppose that $m_1 = 0$ so that $m_2 = \mathrm{Var}(r)$. Our aim is to estimate m_j for $j = 2, 3, \ldots$.

Let $h \geq 2$ be an even integer and define

$$G_h = \mathcal{E}\left(\frac{1}{2}(r_{N[i,k]} - r_i)^h\right) \tag{8.1}$$

which we write as

$$G_h = \frac{1}{2}\sum_{j=0}^{h}\binom{h}{j} m_{h-j} m_j \tag{8.2}$$

Since $m_1 = 0$ the first few values of G_h are given by

$$\begin{aligned} G_0 &= 1 \\ G_2 &= m_2 \\ G_4 &= m_4 + 3m_2^2 \\ G_6 &= m_6 + 15m_4m_2 - 10m_3^2 \\ G_8 &= m_8 + 28m_6m_2 - 56m_5m_3 + 35m_4^2 \\ G_{10} &= m_{10} + 45m_8m_2 - 120m_7m_3 + 210m_6m_4 - 126m_5^2 \\ G_{12} &= m_{12} + 66m_{10}m_2 - 220m_9m_3 + 495m_8m_4 - 792m_7m_5 + 924m_6^2 \end{aligned} \tag{8.3}$$

Using techniques analogous to those employed by the Gamma test we seek to estimate G_h for $h = 2, 4, \ldots$ and then use (8.3) to compute the corresponding estimates of m_j for $j = 2, 3, \ldots$.

We immediately see that there are more unknown variables m_j in (8.3) than there are equations. However, if we make the assumption that the noise distribution is *symmetric* then $m_j = 0$ for all j odd in which case (8.3) becomes

$$\begin{aligned} G_0 &= 1 \\ G_2 &= m_2 \\ G_4 &= m_4 + 3m_2^2 \\ G_6 &= m_6 + 15m_4m_2 \\ G_8 &= m_8 + 28m_6m_2 + 35m_4^2 \\ G_{10} &= m_{10} + 45m_8m_2 + 210m_6m_4 \\ G_{12} &= m_{12} + 66m_{10}m_2 + 495m_8m_4 + 924m_6^2 \end{aligned} \tag{8.4}$$

If we can estimate G_h for $h = 2, 4, \ldots$ then using (8.4) we can successively compute estimates for the moments m_j of symmetric noise distributions.

We call equations (8.3) and (8.4) the *constraining equations*. These equations were first developed and experimentally tested in [Durrant 2001]. In this chapter we use the techniques developed in the preceding chapters to provide a formal justification for the application of the Extended Gamma test and the constraining equations to symmetric noise reconstruction.

8.2 Statement of the theorem

The following theorem is a generalisation of Theorem 1.1 and shows that we can use linear regression to estimate G_h for $h = 2, 4, \ldots$.

Theorem 8.1. *Let the conditions of Theorem 1.1 be satisfied and suppose further that the noise distribution Ψ is symmetric so that $m_j = 0$ for j odd. Let $h \geq 2$ be an even integer and suppose that $m_j < \infty$ for all $j \leq 2h$. Define*

$$\gamma_M(k, h) = \frac{1}{2M} \sum_{i=1}^{M} |y_{N[i,k]} - y_i|^h \tag{8.5}$$

and

$$\delta_M(k) = \frac{1}{M}\sum_{i=1}^{M} |x_{N[i,k]} - x_i|^2 \tag{8.6}$$

Then for every $\kappa > 0$,

$$\gamma_M(k,h) = G_h + G_{h-2}A(M,k,h)\delta_M(k) + o(\delta_M(k)) + O\left(\frac{1}{M^{1/2-\kappa}}\right) \tag{8.7}$$

with probability greater than $1 - O(1/M^{2\kappa})$ *as* $M \to \infty$ *where*

$$A(M,k,h) = \frac{h(h-1)\mathcal{E}_\phi\left(\left((X_{N[i,k]} - X_i)\cdot\nabla f(X_i)\right)^2\right)}{2\mathcal{E}_\phi(|X_{N[i,k]} - X_i|^2)} \tag{8.8}$$

which satisfies

$$0 \le A(M,k,h) \le \frac{1}{2}h(h-1)b_1^2 < \infty \tag{8.9}$$

where b_1 *is the upper bound on* ∇f *over* C.

8.3 Decomposition of the problem

As in Chapter 2 we write

$$\delta_M(k) = \frac{1}{M}\sum_{i=1}^{M} |X_{N[i,k]} - X_i|^2 \tag{8.10}$$

$$\gamma_M(k,h) = \frac{1}{2M}\sum_{i=1}^{M} (Y_{N[i,k]} - Y_i)^h \tag{8.11}$$

where X and Y are random samples on the probability spaces C^M and $C^M \times \mathbb{R}^M$ respectively and consider

$$\frac{1}{2}(Y_{N[i,k]} - Y_i)^h = \frac{1}{2}\left((R_{N[i,k]} - R_i) + (f(X_{N[i,k]}) - f(X_i))\right)^h \tag{8.12}$$

Writing $\Delta_i = f(X_{N[i,k]}) - f(X_i)$ and expanding the right hand side of (8.12) we obtain

$$\frac{1}{2}(Y_{N[i,k]} - Y_i)^h = \frac{1}{2}\sum_{j=0}^{h}\binom{h}{k}(R_{N[i,k]} - R_i)^{h-j}\Delta_i^j \tag{8.13}$$

Following Chapter 2 we apply the second mean value theorem to f so that

$$\Delta_i = (X_{N[i,k]} - X_i)\cdot\nabla f(X_i) + T_f(X_i, X_{N[i,k]}) \tag{8.14}$$

where

$$T_f(X_i, X_{N[i,k]}) = (X_{N[i,k]} - X_i)^{\mathrm{T}} Hf(\Xi_i)(X_{N[i,k]} - X_i) \tag{8.15}$$

for some point Ξ_i on the line segment joining $X_{N[i,k]}$ and X_i. As before we note that the value of $Hf(\Xi_i)$ is uniquely determined by the points $X_{N[i,k]}$ and X_i, even though the intermediate point Ξ_i may not be.

We write

$$\Delta_i^j = \sum_{l=0}^{j} \binom{j}{l} \left((X_{N[i,k]} - X_i) \cdot \nabla f(X_i)\right)^{j-l} T_f(X_i, X_{N[i,k]})^l \tag{8.16}$$

and substitute this into (8.13). Since $|\nabla f(X_i)| \leq b_1$ and $|Hf(\Xi_i)| \leq b_2$ we see that with the exception of the cases where $(h, j) = (0, 0), (1, 0), (1, 1)$ and $(2, 0)$ all terms in the resulting expression are of order $O(|X_{N[i,k]} - X_i|^3)$ as $M \to \infty$.

Summing both sides of the resulting expression over $1 \leq i \leq M$ then dividing by M, we thus arrive at an expression for $\gamma_M(k)$ given by

$$\gamma_M(k, h) = \widetilde{A}_M(k, h) + B_M(k, h) + \widetilde{C}_M(k, h) + o(\delta_M(k)) \quad \text{as} \quad M \to \infty \tag{8.17}$$

where

$$\widetilde{A}_M(k, h) = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{2} \left(R_{N[i,k]} - R_i\right)^h \tag{8.18}$$

$$B_M(k, h) = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{2} h \left(R_{N[i,k]} - R_i\right)^{h-1} \Delta_i^j \tag{8.19}$$

$$\widetilde{C}_M(k, h) = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{4} h(h-1) \left(R_{N[i,k]} - R_i\right)^{h-1} \left((X_{N[i,k]} - X_i) \cdot \nabla f(X_i)\right)^2 \tag{8.20}$$

Next we define

$$A_M(k, h) = \frac{1}{M} \sum_{i=1}^{M} \left(\frac{1}{2} \left(R_{N[i,k]} - R_i\right)^h - G_h \right) \tag{8.21}$$

$$C_M(k, h) = \frac{1}{M} \sum_{i=1}^{M} \Bigg(\frac{1}{4} h(h-1) \left(R_{N[i,k]} - R_i\right)^{h-1} \left((X_{N[i,k]} - X_i) \cdot \nabla f(X_i)\right)^2 - G_{h-2} A(M, k, h) |X_{N[i,k]} - X_i|^2 \Bigg) \tag{8.22}$$

where

$$A(M, k, h) = \frac{h(h-1)\mathcal{E}_\phi \left(\left((X_{N[i,k]} - X_i) \cdot \nabla f(X_i)\right)^2 \right)}{2\mathcal{E}_\phi(|X_{N[i,k]} - X_i|^2)} \tag{8.23}$$

so that $\widetilde{A}_M(k,h) = A_M(k,h) + G_h$ and $\widetilde{C}_M(k) = C_M(k) + G_{h-2}A(M,k,h)\delta_M(k)$. Substituting these into (8.17) we obtain

$$\begin{aligned}\gamma_M(k,h) = G_h + G_{h-2}A(M,k,h)\delta_M(k) + o(\delta_M(k)) \\ + A_M(k,h) + B_M(k,h) + C_M(k,h)\end{aligned} \tag{8.24}$$

and we think of $A_M(k,h)$, $B_M(k,h)$ and $C_M(k,h)$ as random variables on the product space $C^M \times \mathbb{R}^M$.

8.4 Expected value of $A_M(k,h)$, $B_M(k,h)$ and $C_M(k,h)$

Provided $\mathcal{E}(r^j) < \infty$ for all $0 \leq j \leq h$ then by construction the expected value of both $A_M(k,h)$ and $C_M(k,h)$ is zero.

Consider

$$\mathcal{E}(B_M(k,h)) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{2}h\mathcal{E}_\psi\left(\left(R_{N[i,k]} - R_i\right)^{h-1}\right)\mathcal{E}_\phi\left(\Delta_i\right) \tag{8.25}$$

Since $R_{N[i,k]}$ and R_i are independent for $i \neq N[i,k]$ the first expectation in (8.25) can be written as

$$\mathcal{E}((R_{N[i,k]} - R_i)^{h-1}) = \sum_{j=0}^{h-1}\binom{h-1}{j}\mathcal{E}(R^j_{N[i,k]})\mathcal{E}(R_i^{h-j-1}) \tag{8.26}$$

We need this expected value to be identically zero and the only way to ensure this is to require that the distribution Ψ of the noise variables R_i is *symmetric*, so that all of its odd moments are zero.

If the odd moments of R_i are all zero then since h even it follows that $h-j-1$ is odd whenever j is even. Hence for each $0 \leq j \leq h-1$ it follows that at least one of $\mathcal{E}(R^j_{N[i,k]})$ and $\mathcal{E}(R_i^{h-j-1})$ is equal to zero and provided $\mathcal{E}(R^j_{N[i,k]}) < \infty$ for all $0 \leq j \leq h-1$ we thus have that $\mathcal{E}((R_{N[i,k]} - R_i)^{h-1}) = 0$.

By hypothesis, $|(X_{N[i,k]} - X_i)\cdot\nabla f(X_i)| \leq c_1 b_1$ and $|T_f(X_i, X_{N[i,k]})| \leq c_1^2 b_2$ for any random sample $X \in C^M$ so the second expectation in (8.25) is uniformly bounded and it follows that $\mathcal{E}(B_M(k,h)) = 0$ provided Ψ is symmetric.

8.5 Upper bounds on variance

Lemma 8.1. *The variance of term $A_M(k,h)$ satisfies*

$$\mathrm{Var}(A_M(k,h)) \leq \frac{\frac{1}{2}(kK(m)+1)(G_{2h} - 4G_h^2)}{M} \tag{8.27}$$

where $K(m)$ is the maximum kissing number in $\mathbb{R}^m$.

Proof. Define the random variables

$$g_i(R) = \frac{1}{2}(R_{N[i,k]} - R_i)^h - G_h \tag{8.28}$$

on the space of random noise samples $\mathbb{R}^M$ so that

$$A_M(k,h) = \frac{1}{M}\sum_{i=1}^{M} g_i(R) \tag{8.29}$$

By Theorem 4.1, for any indexing structure imposed by some $X \in C^M$ we have that $(g_1, \ldots, g_M)$ is a random sample of identically distributed L–dependent random variables with $L = 2kK(m) + 1$. By Theorem 5.1 it thus follows that

$$\mathrm{Var}(A_M(k,h)) \leq \frac{2(kK(m)+1)\,\mathrm{Var}(g_i)}{M} \tag{8.30}$$

By definition of G_h we have that $\mathcal{E}(g_i) = 0$ so $\mathrm{Var}(g_i) = \mathcal{E}(g_i^2)$. Hence

$$\begin{aligned}
\mathrm{Var}(g_i) &= \mathcal{E}\left(\left(\frac{1}{2}(R_{N[i,k]} - R_i)^h - G_h\right)^2\right) \\
&= \mathcal{E}\left(\frac{1}{4}(R_{N[i,k]} - R_i)^{2h} - G_h(R_{N[i,k]} - R_i)^h + G_h^2\right) \\
&= \mathcal{E}\left(\frac{1}{4}G_{2h} - G_h^2\right)
\end{aligned} \tag{8.31}$$

so by (8.30) we obtain

$$\mathrm{Var}(A_M(k,h)) \leq \frac{\frac{1}{2}(kK(m)+1)(G_{2h} - 4G_h^2)}{M} \tag{8.32}$$

as required. □

Lemma 8.2. *The variance of term $B_M(k,h)$ satisfies*

$$\mathrm{Var}(B_M(k,h)) \leq \frac{(kK(m)+1)h^2c_1^2(b_1+c_1b_2)^2G_{2h-2}}{M} \tag{8.33}$$

where $K(m)$ is the maximum kissing number in $\mathbb{R}^m$.

Proof. Define the random variables

$$g_i(R) = \frac{1}{2}h(R_{N[i,k]} - R_i)^{h-1} \tag{8.34}$$

$$h_i(X) = (X_{N[i,k]} - X_i)\cdot\nabla f(X_i) + T_f(X_i, X_{N[i,k]}) \tag{8.35}$$

on $\mathbb{R}^M$ and C^M respectively so that

$$B_M(k,h) = \frac{1}{M}\sum_{i=1}^{M} g_i(R)h_i(X) \tag{8.36}$$

Since $\mathcal{E}(B_M(k,h)) = 0$ we have that $\mathrm{Var}(B_M(k,h)) = \mathcal{E}(B_M(k,h)^2)$ and computing this over all pairs $(X,R) \in C^M \times \mathbb{R}^M$ we obtain

$$\mathrm{Var}(B_M(k,h)) = \frac{1}{M^2}\sum_{i,j=1}^{M} \mathcal{E}(h_i(X)h_j(X)g_i(R)g_j(R)) \tag{8.37}$$

which we write as

$$\mathrm{Var}(B_M(k,h)) = \frac{1}{M^2}\sum_{i,j=1}^{M} \mathcal{E}_\phi\big(h_i(X)h_j(X)\mathcal{E}_\psi(g_i(R)g_j(R)\big|X)\big) \tag{8.38}$$

Clearly,

$$\begin{aligned}\mathrm{Var}(B_M(k,h)) &\leq \frac{1}{M^2}\sum_{i,j=1}^{M} \big|\mathcal{E}_\phi\big(h_i(X)h_j(X)\mathcal{E}_\psi(g_i(R)g_j(R)\,\big|\,X)\big)\big| \\ &\leq \frac{1}{M^2}\sum_{i,j=1}^{M} \mathcal{E}_\phi\big(|h_i(X)|\,|h_j(X)|\,|\mathcal{E}_\psi(g_i(R)g_j(R)\,\big|\,X)|\big)\end{aligned} \tag{8.39}$$

By Theorem 4.1, for any $X \in C^M$ the $g_1(R),\ldots,g_M(R)$ are identically distributed L–dependent random variables with $L = 2(kK(m)+1)$ where $K(m)$ is the maximum kissing number in $\mathbb{R}^m$. Hence, for any fixed i it follows that $\mathcal{E}_\psi(g_i(R)g_j(R)\big|X) \neq 0$ for at most L indices j. Furthermore, since $|ab| \leq \frac{1}{2}(a^2+b^2)$ for any pair of real numbers a and b and since $g_i(R)$ and $g_j(R)$ are identically distributed, we have that

$$|\mathcal{E}_\psi(g_i(R)g_j(R)\big|X)| \leq \mathcal{E}_\psi(|g_i(R)g_j(R)|\big|X) \leq \mathcal{E}_\psi(g_i^2(R)\,\big|\,X) \tag{8.40}$$

Thus, since

$$\mathcal{E}_\psi(g_i^2\,\big|\,X) = \mathcal{E}_\psi\left(\frac{1}{4}h^2(R_{N[i,k]}-R_i)^{2h-2}\right) = \frac{1}{2}hG_{2h-2} < \infty \tag{8.41}$$

and since $|h_i(X)| \leq c_1b_1 + c_1^2b_2$ for all $X \in C^M$ it follows by (8.39) that

$$\mathrm{Var}(B_M(k,h)) \leq \frac{(kK(m)+1)h^2c_1^2(b_1+c_1b_2)^2G_{2h-2}}{M} \tag{8.42}$$

as required. □

Lemma 8.3 (The $C_M(k,h)$ term). *The variance of term $C_M(k,h)$ satisfies*

$$\mathrm{Var}(C_M(k,h)) \leq \frac{\lambda_C}{M} \tag{8.43}$$

where

$$\begin{aligned}\lambda_C = \frac{1}{4}h^2(h-1)^2c_1^4b_1^4\Big(&(kK(m)+1)G_{2h-4} \\ &+ 4(1+(k+2)(3k+2)(1+k(k+1)K(m)))G_{h-2}^2\Big)\end{aligned} \tag{8.44}$$

and $K(m)$ is the maximum kissing number in $\mathbb{R}^m$.

Proof. Define the random variables

$$S_i(X,R) = \frac{1}{4}h(h-1)(R_{N[i,k]}-R_i)^{h-2}\left((X_{N[i,k}-X_i)\cdot\nabla f(X_i)\right)^2 \quad (8.45)$$

$$T_i(X) = G_{h-2}A(M,k,h)|X_{N[i,k]}-X_i|^2 \quad (8.46)$$

on $C^M\times\mathbb{R}^M$ and C^M respectively where

$$A(M,k,h)=\frac{h(h-1)\mathcal{E}_\phi\left(\left((X_{N[i,k]}-X_i)\cdot\nabla f(X_i)\right)^2\right)}{2\mathcal{E}_\phi(|X_{N[i,k]}-X_i|^2)} \quad (8.47)$$

so that

$$C_M(k,h)=\frac{1}{M}\sum_{i=1}^{M}(S_i-T_i) \quad (8.48)$$

By definition of $A(M,k,h)$ we have that $\mathcal{E}(C_M(k,h))=0$. Thus $\mathrm{Var}(C_M(k,h))=\mathcal{E}(C_M(k,h)^2)$ and we write

$$\begin{aligned}\mathrm{Var}(C_M(k,h)) &= \frac{1}{M^2}\sum_{i,j=1}^{M}\mathcal{E}\big((S_i-T_i)(S_j-T_j)\big)\\ &= \frac{1}{M^2}\sum_{i,j=1}^{M}\mathcal{E}(S_iS_j+T_iT_j-2S_iT_j)\\ &= \frac{1}{M^2}\sum_{i,j=1}^{M}\big(\mathcal{E}(S_iS_j)+\mathcal{E}(T_iT_j)-2\mathcal{E}(S_iT_j)\big) \quad (8.49)\end{aligned}$$

Since h is even and $h\geq 2$ it follows that $S_i\geq 0$ and $T_i\geq 0$ for every $X\in C^M$ and $R\in\mathbb{R}^M$. Furthermore, since $|(X_{N[i,k]}-X_i)\cdot\nabla f(X_i)|^2\leq|X_{N[i,k}-X_i)|^2b_1^2$ we have that $A(M,k,h)\leq\frac{1}{2}h(h-1)b_1^2$ and since $|X_{N[i,k]}-X_i|\leq c_1$ it follows that $T_i\leq\frac{1}{2}h(h-1)G_{h-2}c_1^2b_1^2$.

Thus we obtain

$$0\leq\mathcal{E}(S_iT_j)\leq\frac{1}{2}h(h-1)G_{h-2}c_1^2b_1^2\mathcal{E}(S_i) \quad (8.50)$$

and since

$$0\leq\mathcal{E}(S_i)\leq\frac{1}{2}h(h-1)\mathcal{E}\left(\frac{1}{2}(R_{N[i,k]}-R_i)^{h-2}\right)c_1^2b_1^2 \quad (8.51)$$

then by definition of G_{h-2} we see that

$$0\leq\mathcal{E}(S_iT_j)\leq\frac{1}{4}h^2(h-1)^2G_{h-2}^2c_1^4b_1^4<\infty \quad (8.52)$$

Hence by (8.49) it follows that

$$\mathrm{Var}(C_M(k,h)) \leq \frac{1}{M^2}\sum_{i,j=1}^{M}\big(\mathcal{E}(S_iS_j)+\mathcal{E}(T_iT_j)\big) \tag{8.53}$$

We define

$$\bar{S}_M = \frac{1}{M}\sum_{i=1}^{M} S_i \qquad \text{and} \qquad \bar{T}_M = \frac{1}{M}\sum_{i=1}^{M} T_i \tag{8.54}$$

so that

$$\mathrm{Var}(C_M(k,h)) \leq \mathrm{Var}(\bar{S}_M) + \mathrm{Var}(\bar{T}_M) \tag{8.55}$$

and consider each term separately. Regarding the first term on the right hand side of (8.55) we define the random variables

$$g_i(R) = \frac{1}{2}(R_{N[i,k]} - R_i)^{h-2} \tag{8.56}$$

$$h_i(X) = \frac{1}{2}h(h-1)((X_{N[i,k]} - X_i)\cdot\nabla f(X_i))^2 \tag{8.57}$$

on $\mathbb{R}^M$ and C^M respectively so that

$$\bar{S}_M = \frac{1}{M}\sum_{i=1}^{M} g_i(R)h_i(X) \tag{8.58}$$

Although the expected value of $\bar{S}_M$ is not zero it is certainly true that $\mathrm{Var}(\bar{S}_M) \leq \mathcal{E}(\bar{S}_M^2)$ so we have that

$$\mathrm{Var}(\bar{S}_M) \leq \frac{1}{M^2}\sum_{i=1}^{M}\mathcal{E}((g_i(R)g_j(R))(h_i(X)h_j(X))) \tag{8.59}$$

where the expectation is taken over all $X \in C^M$ and $R \in \mathbb{R}^M$. By hypothesis both h_i and h_j are completely independent of R while g_i and g_j may depend on X. Thus we write

$$\mathrm{Var}(\bar{S}_M) = \frac{1}{M^2}\sum_{i,j=1}^{M}\mathcal{E}_\phi\big(h_i(X)h_j(X)\mathcal{E}_\psi(g_i(R)g_j(R)\,\big|\,X)\big) \tag{8.60}$$

from which it follows easily that

$$\mathrm{Var}(\bar{S}_M) \leq \frac{1}{M^2}\sum_{i,j=1}^{M}\mathcal{E}_\phi\big(|h_i(X)|\,|h_j(X)|\,|\mathcal{E}_\psi(g_i(R)g_j(R)\,\big|\,X)|\big) \tag{8.61}$$

By Theorem 4.1, for any $X \in C^M$ the $g_1(R),\ldots,g_M(R)$ are identically distributed L–dependent random variables with $L = 2(kK(m)+1)$ where $K(m)$ is the maximum

kissing number in $\mathbb{R}^m$. Hence, for any fixed i it follows that $\mathcal{E}_\psi(g_i(R)g_j(R)\big|X) \neq 0$ for at most L indices j. Furthermore, since $|ab| \leq \frac{1}{2}(a^2+b^2)$ for any pair of real numbers a and b and since $g_i(R)$ and $g_j(R)$ are identically distributed, we have that

$$|\mathcal{E}_\psi(g_i(R)g_j(R)\big|X)| \leq \mathcal{E}_\psi(|g_i(R)g_j(R)|\big|X) \leq \mathcal{E}_\psi(g_i^2(R)\,\big|\,X) \tag{8.62}$$

Thus, since

$$\mathcal{E}_\psi(g_i^2(R)\,\big|\,X) = \mathcal{E}_\psi\left(\frac{1}{4}(R_{N[i,k]} - R_i)^{2h-4}\right) = \frac{1}{2}G_{2h-4} \tag{8.63}$$

and since $|h_i| \leq \frac{1}{2}h(h-1)c_1^2b_1^2$ for every $X \in C^M$ it follows by (8.61) that

$$\mathrm{Var}(\bar{S}_M) \leq \frac{\frac{1}{4}(kK(m)+1)h^2(h-1)^2c_1^4b_1^4G_{2h-4}}{M} \tag{8.64}$$

Turning our attention to the second term on the right hand side of (8.55) we define the random variable

$$h_i(X) = G_{h-2}A(M,k,h)|X_{N[i,k]} - X_i|^2 \tag{8.65}$$

on C^M so that

$$\bar{T}_M = \frac{1}{M}\sum_{i=1}^{M} h_i(X) \tag{8.66}$$

Since $0 \leq A(M,k,h) \leq \frac{1}{2}h(h-1)b_1^2$ and $|X_{N[i,k]} - X_i| \leq c_1$ it follows that

$$||h|| = \sup\big\{|h_i(X)| \,:\, X \in C^M\big\} \leq \frac{1}{2}G_{h-2}h(h-1)c_1^2b_1^2 < \infty \tag{8.67}$$

Thus we may apply Theorem 6.1 to the sample mean $\bar{T}_M$ of the h_i so we have that

$$\mathrm{Var}(\bar{T}_M) \leq \frac{2(1+(k+2)(3k+2)(1+k(k+1)K(m)))G_{h-2}h(h-1)c_1^2b_1^2}{M}(\mathcal{E}(|h_1|^2))^{1/2} \tag{8.68}$$

where $K(m)$ is the maximum kissing number in $\mathbb{R}^m$.

Since $\mathcal{E}(|h_1(X)|^2)^{1/2} \leq \frac{1}{2}G_{h-2}h(h-1)c_1^2b_1^2$ for every $X \in C^M$ it thus follows that

$$\mathrm{Var}(\bar{T}_M) \leq \frac{(1+(k+2)(3k+2)(1+k(k+1)K(m)))G_{h-2}^2h^2(h-1)^2c_1^4b_1^4}{M} \tag{8.69}$$

and finally, substituting (8.64) and (8.69) into (8.55) we obtain

$$\mathrm{Var}(C_M(k,h)) \leq \frac{\lambda_C}{M} \tag{8.70}$$

where

$$\begin{aligned}\lambda_C = \frac{1}{4}h^2(h-1)^2c_1^4b_1^4\Big(&(kK(m)+1)G_{2h-4} \\ &+ 4(1+(k+2)(3k+2)(1+k(k+1)K(m)))G_{h-2}^2\Big)\end{aligned} \tag{8.71}$$

as required. □

8.6 Probabilistic upper bounds on $A_M(k,h)$, $B_M(k,h)$ and $C_M(k,h)$

Since the expected values of $A_M(k,h)$, $B_M(k,h)$ and $C_M(k,h)$ are zero, by Lemma 8.1, Lemma 8.2, Lemma 8.3 and Chebyshev's inequality we obtain the following.

Corollary 8.1. *For every $\epsilon > 0$,*

$$\begin{aligned}\mathbf{P}(|A_M(k,h)| > \epsilon) &\leq \frac{\lambda_A}{M\epsilon^2}\\ \mathbf{P}(|B_M(k,h)| > \epsilon) &\leq \frac{\lambda_B}{M\epsilon^2}\\ \mathbf{P}(|C_M(k,h)| > \epsilon) &\leq \frac{\lambda_C}{M\epsilon^2}\end{aligned}$$

where

$$\begin{aligned}\lambda_A &= \frac{1}{2}(kK(m)+1)(G_{2h}-4G_h^2)\\ \lambda_B &= (kK(m)+1)h^2c_1^2(b_1+c_1b_2)^2G_{2h-2}\\ \lambda_C &= \frac{1}{4}h^2(h-1)^2c_1^4b_1^4\Big((kK(m)+1)G_{2h-4}\\ &\qquad +4(1+(k+2)(3k+2)(1+k(k+1)K(m)))G_{h-2}^2\Big)\end{aligned}$$

are finite constants not depending on M.

8.7 Proof of Theorem 8.1

Applying Lemma 7.4 to the results of Corollary 8.1 we obtain the following.

Lemma 8.4. *For any $\epsilon > 0$,*

$$\mathbf{P}(|A_M(k,h)+B_M(k,h)+C_M(k,h)| > \epsilon) \leq \frac{\lambda}{M\epsilon^2} \tag{8.72}$$

where

$$\lambda = 9(\lambda_A+\lambda_B+\lambda_C) \tag{8.73}$$

is a finite constant not depending on M.

Corollary 8.2. *For any $\epsilon > 0$,*

$$\mathbf{P}(\big|\gamma_M(k,h) - \big(G_h + G_{h-2}A(M,k,h)\delta_M(k) + o(\delta_M(k))\big)\big| > \epsilon) \leq \frac{\lambda}{M\epsilon^2} \tag{8.74}$$

as $M \to \infty$ where λ is a finite constant not depending on M.

Proof. By (8.24) we have that

$$\begin{aligned}&A_M(k,h)+B_M(k,h)+C_M(k,h)\\ &\qquad = \gamma_M(k,h) - \big(G_h + G_{h-2}A(M,k,h)\delta_M(k) + o(\delta_M(k))\big)\end{aligned} \tag{8.75}$$

and the result follows by Lemma 8.4. □

Proof of Theorem 8.1

Proof. Let $\kappa > 0$ and apply Corollary 8.2 with

$$\epsilon = \frac{1}{M^{1/2-\kappa}} \tag{8.76}$$

Then

$$\mathbf{P}\left(\left|\gamma_M(k,h) - \left(G_h + G_{h-2}A(M,k,h)\delta_M(k) + o(\delta_M(k))\right)\right| > \frac{1}{M^{1/2-\kappa}}\right) \leq \frac{\lambda}{M^{2\kappa}} \tag{8.77}$$

so that

$$\mathbf{P}\left(\left|\gamma_M(k,h) - \left(G_h + G_{h-2}A(M,k,h)\delta_M(k) + o(\delta_M(k))\right)\right| \leq \frac{1}{M^{1/2-\kappa}}\right) > 1 - \frac{\lambda}{M^{2\kappa}} \tag{8.78}$$

Hence, for every $\kappa > 0$

$$\gamma_M(k,h) = G_h + G_{h-2}A(M,k,h)\delta_M(k) + o(\delta_M(k)) + O\left(\frac{1}{M^{1/2-\kappa}}\right) \tag{8.79}$$

with probability greater than $1 - O(1/M^{2\kappa})$ as $M \to \infty$, as required. □

8.8 The extended algorithm for symmetric noise distributions

The extended Gamma test algorithm for symmetric noise distributions is shown in Algorithm 2.

8.9 Proof of the extended Gamma test algorithm

We use Theorem 8.1 to show that the numbers Γ_h returned by Algorithm 2 converge in probability to G_h as $M \to \infty$.

Theorem 8.2. *Subject to the condition that for some fixed $p \geq 1$ there exists a positive constant $c = c(m) < 1$ such that*

$$\delta_M(1) \leq c\delta_M(p) \tag{8.80}$$

for all sufficiently large M, the number Γ_h returned by Algorithm 2 converges in probability to G_h as $M \to \infty$.

Procedure *Extended Gamma Test* (data)
{data is an array of points (x_i, y_i), $(1 \le i \le M)$ where $x \in \mathbb{R}^m$ and $y \in \mathbb{R}$}
for $i = 1$ to M **do**
  for $k = 1$ to p **do**
    compute $N[i,k]$ where $x_{N[i,k]}$ is the k^{th} nearest neighbour of x_i.
  end for
end for
{If multiple outputs do the remainder for each output}
for $k = 1$ to p **do**
  compute $\delta_M(k)$ as in (1.40)
end for
for $h = 2$ to $2q$, $h = h + 2$ **do**
  for $k = 1$ to p **do**
    Compute $\gamma_M(k,h)$ as in (8.5)
  end for
  Perform linear regression on $\{(\delta_M(k), \gamma_M(k,h)), 1 \le k \le p\}$
  Record the intercept Γ_h
end for
Compute successive estimates for $m_2, m_4, \ldots, m_{2q}$ according to equations (8.4), replacing G_h by Γ_h.

Algorithm 2: The extended Gamma test algorithm.

Proof. By Theorem 8.1 we have that for every $\kappa > 0$,

$$\gamma_M(k,h) = G_h + G_{h-2}A(M,k,h)\delta_M(k) + o(\delta_M(k)) + O\left(\frac{1}{M^{1/2-\kappa}}\right) \tag{8.81}$$

with probability greater than $1 - O(1/M^{2\kappa})$ as $M \to \infty$. Furthermore, by condition (8.80) it follows that $\delta_M(1) < \delta_M(p)$. Hence the conditions of Lemma 7.6 are satisfied by the pairs $(X_k, Y_k) = (\delta_M(k), \gamma_M(k,h))$ with $a_k = G_{h-2}A(M,k,h)$, $\Delta_k = o(\delta_M(k)) + O(1/M^{1/2-\kappa})$, $v = G_h$ and $d = \Gamma_h$ so

$$|\Gamma_h - G_h| \le p(p-1)(A\delta_M(p) + \Delta)\left(\frac{\delta_M(p)}{\delta_M(p) - \delta_M(1)}\right) \tag{8.82}$$

where $A = \max_k(G_{h-2}A(M,k,h))$ and $\Delta = o(\delta_M(p)) + O(1/M^{1/2-\kappa})$. Condition (8.80) implies that $\delta_M(p) - \delta_M(1) \ge (1-c)\delta_M(p)$ for some constant $c > 0$ so

$$\frac{\delta_M(p)}{\delta_M(p) - \delta_M(1)} = O(1) \quad \text{as} \quad M \to \infty \tag{8.83}$$

Furthermore, since $G_{h-2} < \infty$ and $0 \le A(M,k) \le \frac{1}{2}h(h-1)b_1^2 < \infty$ for every $1 \le k \le p$ we also have $A\delta_M(p) = O(\delta_M(p))$ as $M \to \infty$. Hence

$$|\Gamma_h - G_h| = O(\delta_M(p)) + O(1/M^{1/2-\kappa}) \tag{8.84}$$

with probability greater than $1 - O(1/M^{2\kappa})$ as $M \to \infty$ and we conclude that

$$\Gamma_h \to G_h \quad \text{in probability as} \quad M \to \infty \tag{8.85}$$

as required. □

Theorem 8.3. *Let C be a compact convex body in $\mathbb{R}^m$ and let ϕ be a smooth positive sampling density on C. Then the number Γ_h returned by Algorithm 2 converges in probability to G_h as $M \to \infty$.*

Proof. Using an identical argument to that used in the proof Theorem 1.3, since $\delta_M(1)$ and $\delta_M(p)$ satisfy (8.80) the result follows immediately by Theorem 8.2. □

8.10 Summary

In this chapter we have provided a formal justification for the method of symmetric noise distribution reconstruction illustrated in [Durrant 2001]. The idea that a *perfect model* is one for which the output error distribution on unknown input data is identical to the noise distribution was (to our knowledge) first proposed in [Durrant 2001]. If there were a way to reconstruct the noise distribution then knowledge of this distribution could conceivably be used in model construction. Instead of training the non-parametric model to achieve an error variance equal to the noise variance one could aim to develop training methods such that some measure of the difference between the error *distribution* and the noise *distribution* was minimised. Hence the idea of reconstructing the noise distribution may eventually lead to interesting practical applications for non-parametric model construction.

Chapter 9

Conclusion

9.1 The significance of the Gamma test proof

The proof of the Gamma test established in this thesis covers the case of a smooth positive sampling density over a compact convex body in $\mathbb{R}^m$, which is sufficient to include many practical applications. From a purely theoretical standpoint the proof is of value since it assures us that the basic methodology is sound. Moreover, the techniques and theorems we have developed for the analysis of near neighbour relationships may well find other interesting applications.

However, it must said that in particular applications of the Gamma test, where often no theoretical analysis exists, the extent to which theoretical preconditions are satisfied may be unverifiable. In practice it is often the case that the easiest method of determining the utility of the Gamma test is simply to try it!

Thus we should be aware that the degree to which models discovered by tools such as the Gamma test are truly scientific depends on the context. In such situations we may be motivated by purely pragmatic considerations:

> "A model is 'good' just as long as it is predicting well. When it stops predicting well, we just try to build another model." [Jones *et al.* 2001]

9.2 Generalising the proof

We have remarked that the applicability of Gamma test seems much wider than the class of situations considered here. In particular, it is often effective when applied to chaotic dynamical systems based on smooth underlying relationships, and it is natural to ask which aspects of the proof might possibly be generalised to cover these other cases of interest.

With regard to the work of Chapter 3, results such as Conjecture 3.2 seem to raise several quite difficult issues related to Hausdorff measures. However, examination of the proof structure shown in Figure 1.9 reveals that the critical application of Theorem 3.2 is to establish condition (7.48) - a fact that became apparent only some time after the work of Chapter 3 had been completed. Nevertheless, generalisations of Theorem 3.2 based on Conjecture 3.2 may be one way to prove condition (7.48) for the ergodic sampling of a chaotic attractor.

Condition (7.48) has every appearance of being a purely geometrical constraint and might possibly be proved to hold for *any* bounded sequence of sampling points. A first step in eliminating the arguments of Chapter 3 from a proof of the Gamma test might therefore be to address the following problem.

Conjecture 9.1. *There exists some integer $q(m)$ such that for any sequence of points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ in $\mathbb{R}^m$ and for every $p > q(m)$,*

$$\limsup_{M \to \infty} \frac{\delta_M(1)}{\delta_M(p)} < 1 \tag{9.1}$$

If this conjecture could be proved then the work of Chapter 3 would not be required in a proof of the Gamma test, even though it is of considerable interest in its own right.

Suppose that condition (7.48) could be established in some alternative manner. Then we ask which other parts of the proof would carry through to more general classes of sampling distributions? Certainly the work of Chapter 5 on L-dependent variables is independent of the particular method of sampling, because its application to the proof of the Gamma test hinges on Theorem 4.1 which is a purely geometric result. It is less obvious how much of Chapter 6 can be carried through to the more general case because the integrals involving the sampling density would have to be carefully defined in terms of an appropriate Hausdorff measure. However, it is quite conceivable that this work could be done.

Yet another area of generalisation relates to the type of convergence for which the various results may be shown to hold. We have focused on the relatively weak notion of convergence in probability. It is certainly possible that this could be generalised to the stronger notion of 'convergence with probability one', and perhaps even further.

9.3 Implications for further work

9.3.1 The gradient $A(M, k)$

Based on the discussion of section 7.7 and the experimental results presented in 7.8.3, we make the following conjectures regarding the asymptotic behaviour of the gradient $A(M, k)$ as $M \to \infty$.

Conjecture 9.2. *If the input points X_i are identically distributed according to a sampling distribution that satisfies the conditions of Theorem 3.2, there exists some finite constant $A > 0$, independent of M and k, such that*

$$A(M,k) \to A \quad \textit{in probability as} \quad M \to \infty \tag{9.2}$$

Conjecture 9.3. *For some more general class of sampling distributions than those satisfying the conditions of Theorem 3.2, there exists some finite constant $A(k) > 0$, independent of M, such that*

$$A(M,k) \to A(k) \quad \textit{in probability as} \quad M \to \infty \tag{9.3}$$

9.3.2 Near neighbour distance distributions

In Chapter 3 we determined the asymptotic behaviour of the moments of the kth nearest neighbour distance distribution of M points using the technique of 'boundary shrinking' first suggested by W.M. Schmidt. It would be of some interest to further characterise this distribution. This work may have some interesting consequences in the theoretical study of point processes and also in more practical applications – for example, estimates based on near neighbour distances have been employed in tests of uniformity.

9.3.3 Questions relating to near neighbour geometry

In Chapter 4 we examined some theoretical questions relating to near neighbour geometry and the result of Theorem 4.1 has clear relevance to coding theory. As a further illustration of the 'boundary shrinking' technique we also calculated the expected number of connected components in the first nearest neighbour graph of a uniform sampling distribution. It might be useful to have explicit formulae for the probabilities that any point is the kth nearest neighbor of exactly $0, 1, 2, \ldots$ other points. This question is addressed by [Henze 1987].

9.3.4 L-dependent random variables

In Chapter 5 we have given a simple and coherent theory for an interesting class of dependent variables and concluded with a simple proof of the associated Central Limit theorem. What is probably of most interest here is the conceptual framework, which can easily be identified for particular applications.

9.3.5 Functions of a point and its k nearest neighbours

The work of Chapter 6 generalises the earlier work of [Bickel and Breiman 1983] on functions of a point and the distance to its first nearest neighbour. This quite technically demanding analysis seemed unavoidable and owes much to their work. It is

clearly capable of further generalisation to a similar theory (and central limit theorem) for sums of bounded functions of the form $h_i(X) = h(X_{N[i,k]}, X_{N[i,k-1]} \ldots, X_i)$, but such specialised and demanding work seems only worthwhile if prompted by a significant application.

In [Penrose and Yukich 2001] a central limit theorem is proved for functionals of various types of random point sets, including sets obtained by selecting points *uniformly* at random from a fixed set in $\mathbb{R}^m$. The functionals in question must be strongly stabilising, satisfy a uniform bounded moments condition and be polynomially bounded (see [Penrose and Yukich 2001] for details). It is not clear whether these conditions would impose any further conditions on our (unknown) smooth function f, beyond those given in (1.4).

9.3.6 Noise reconstruction

In Chapter 8 we provide a theoretical analysis which justifies using a combination of an extension of the Gamma test and certain constraining equations to estimate the even moments of a symmetric noise distribution, a technique first described and tested experimentally in [Durrant 2001]. It is not clear how this analysis could be extended to encompass non–symmetric noise distributions.

9.4 Final conclusions

The Gamma test has been in the public domain since 1995 and many illustrations of its utility in non-linear modelling have been published by the Cardiff group. A software tool *winGamma*™ based on the the technique has been commercially available[1] since 1998. Nevertheless, despite the huge amount of empirical work on non-linear modelling currently being produced, apart from that mentioned almost none to date has used the Gamma test. Exceptions are a group at Glamorgan University interested in property price prediction, and a group at the Naval Research Laboratory, Washington working on chaotic dynamics.

Given the demonstrated effectiveness of the Gamma test, for example in feature selection and signal processing, we find the lack of interest in the method rather puzzling, particularly given the dearth of alternative techniques. If linear regression cannot be applied to the problem at hand, the Gamma test seems to represent the best approach to true non-linear regression currently available. Thus we entertain the hope that the work of this thesis, when eventually published in papers and book form, will serve to commend the Gamma test to the wider non-linear modelling and statistical community.

[1] Under license from Cardiff University.

Bibliography

[Artin 1964] E. Artin. *The Gamma Function*. Holt, Rinehart and Winston, 1964.

[Baldi and Rinott 1989] P. Baldi and Y. Rinott. On normal approximations of distributions in terms of dependency graphs. *The Annals of Probability*, 17(4):1646–1650, 1989.

[Bates and Watts 1988] D. M. Bates and D. G. Watts. *Nonlinear regression analysis and its applications*. J. Wiley & Sons, 1988. ISBN 0-8194-1845-5.

[Beardwood *et al.* 1959] J. Beardwood, J. H. Halton and J. M. Hammersley. The shortest path through many points. *Proc. Camb. Phil. Soc*, 55:299–327, 1959.

[Bentley 1975] J. L. Bentley. Multidimensional binary search trees used for associative search. *Comm. ACM*, 18:309–517, 1975.

[Bickel and Breiman 1983] P. J. Bickel and L. Breiman. Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. *The Annals of Probability*, 11(1):185–214, 1983.

[Billingsley 1979] P. Billingsley. *Probability and Measure*. J. Wiley & Sons, 1979. ISBN 0-4710-3173-9.

[Birkhoff 1927] G. D. Birkhoff. On the periodic motions of dynamical systems. *Acta Mathematica*, 50:359–379, 1927.

[Chuzhanova *et al.* 1998] Nadia A. Chuzhanova, Antonia J. Jones and S. Margetts. Feature selection for genetic sequence classification. *Bioinformatics*, 14(2):139–143, 1998.

[Cox 1981] T. F. Cox. Reflexive nearest neighbors. *Biometrics*, 37:367–369, 1981.

[de Oliveira 1999] Ana Guedes de Oliveira. *Synchronization of Chaos and Applications to Secure Communications*. PhD thesis, Department of Computing, Imperial College of Science, Technology and Medicine, University of London, UK, 1999.

[Durrant 2001] P. J. Durrant. winGamma^{TM} *: a non-linear data analysis and modelling tool with applications to flood prediction.* PhD thesis, Department of Computer Science, Cardiff University, Wales, UK, 2001.

[Eppstein *et al.* 1997] D. Eppstein, M. S. Patterson and F. F. Yao. On nearest–neighbor graphs. *Discrete Comput. Geom.*, 17:263–282, 1997.

[Falconer 1990] K. J. Falconer. *Fractal geometry: mathematical foundations and applications.* J. Wiley & Sons, 1990. ISBN 0-4719-2287-0.

[Feller 1971] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. J. Wiley & Sons, second edition, 1971. ISBN 0-471-25709-5.

[Hénon 1976] M. Hénon. A two dimensional mapping with a strange attractor. *Comm. Math. Phys*, 50(69), 1976.

[Henze 1986] N. Henze. On the probability that a random point is the jth nearest neighbour of its own kth nearest neighbour. *J. Appl. Prob.*, 23:221–226, 1986.

[Henze 1987] N. Henze. On the fraction of random points with specified nearest-neighbour interrelations and degree of attraction. *Adv. Appl. Prob.*, 19:873–895, 1987.

[Johnson *et al.* 1996] D. S. Johnson, L. A. McGeogh and E. E. Rothberg. Asymptotic experimental analysis for the Held-Karp traveling salesman bound. In *Proceedings of the seventh Annual ACM–SIAM Symposium on Discrete Algorithms*, pages 341–350, 1996.

[Jones *et al.* 2001] Antonia J. Jones, Dafydd Evans, Steve Margetts and Peter M. Durrant. The Gamma test. In Ruhul Sarker, Hussein Abbass, and Charles Newton, editors, *Heuristic and Optimization for Knowledge Discovery*, chapter IX. Idea Group Publishing, Hershey, PA., 2001. ISBN 1-9307-0826-2.

[Jones *et al.* 2002] A. J. Jones, A. P. M Tsui and A. G. Oliveira. Neural models of arbitrary chaotic systems: construction and the role of time delayed feedback in control and synchronization. *Complexity International*, 09, 2002. URL: http://www.csu.edu.au/ci/vol09/tsui01/.

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

[Kendall and Stuart 1963] M. G. Kendall and A. Stuart. *The Advanced Theory of Statistics, Volume 1: Distribution Theory.* Griffin, 1963.

[Končar 1997] N. Končar. *Optimisation methodologies for direct inverse neurocontrol.* PhD thesis, Department of Computing, Imperial College of Science, Technology and Medicine, University of London, UK, 1997.

[Noether 1970] G. E. Noether. A central limit theorem with nonparametric applications. *The Annals of Mathematical Statistics*, 41(5):1753–1755, 1970.

[Ott 1993] E. Ott. *Chaos in Dynamical Systems*. Cambridge University Press, 1993. ISBN 0-521-43799-7.

[Penrose and Yukich 2001] M. D. Penrose and J. E. Yukich. Central limit theorems for some graphs in computational geometry. *Ann. Appl. Prob.*, 11(4):1005–1041, 2001.

[Percus and Martin 1996] A. G. Percus and O. C. Martin. Finite size and dimensional dependence in the euclidean travelling salesman problem. *Physical Review Letters*, 76(8):1188–1191, 1996.

[Percus and Martin 1998] A. G. Percus and O. C. Martin. Scaling universalities of kth nearest neighbor distances on closed manifolds. *Adv. Appl. Math.*, 21:424–436, 1998.

[Petrovskaya and Leontovich 1982] M. B. Petrovskaya and A. M. Leontovich. The central limit theorem for a sequence of random variables with a slowly growing number of dependencies. *Theory Probab. Appl.*, 27:815–825, 1982.

[Pi and Peterson 1994] H. Pi and C. Peterson. Finding the embedding dimension and variable dependencies in time series. *Neural Computation*, 6(3):509–520, 1994.

[Pickard 1982] D. K. Pickard. Isolated nearest neighbors. *J. Appl. Prob.*, 19:444–449, 1982.

[Steele 1981] J. M. Steele. Subadditive euclidean functinals and non-linear growth in geometric probability. *Ann. Prob*, 9:365–376, 1981.

[Stefánsson *et al.* 1997] Aðalbjörn Stefánsson, N. Končar and Antonia J. Jones. A note on the Gamma test. *Neural Computing & Applications*, 5:131–133, 1997.

[Stein 1977] D. Stein. *Scheduling Dial–a–Ride transportation systems: an asymptotic approach*. PhD thesis, Harvard University, Cambridge, MA, 1977.

[Stein 1986] C. Stein. Approximate computation of expectations. *IMS, Hayward, California*, 1986.

[Takens 1981] F. Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence*, volume 898 of *Lecture Notes in Mathematics*, pages 366–381. Springer-Verlag, 1981.

[Tsui *et al.* 2002] A. P. M. Tsui, A. J. Jones and A. G. de Oliveira. The construction of smooth models using irregular embeddings determined by a Gamma test analysis. *Neural Computing and Applications*, 10:318–329, 2002.

[Tsui 1999] A. P. M. Tsui. *Smooth Data Modelling and Stimulus-Response via Stabilisation of Neural Chaos*. PhD thesis, Department of Computing, Imperial College of Science, Technology and Medicine, University of London, UK, 1999.

[Valenzuela and Jones 1997] Christine L. Valenzuela and Antonia J. Jones. Estimating the Held–Karp lower bound for the geometric travelling salesman problem. *European Journal of Operational Research*, 102(1):157–175, 1997.

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

[Yukich 1998] J. E. Yukich. *Probability Theory of Classical Euclidean Optimization Problems*. Springer-Verlag, 1998. ISBN 3-540-63666-8.

[Zeger and Gersho 1994] K. Zeger and A. Gersho. The number of nearest neighbors in a Euclidean code. *IEEE Trans. Inform. Theory*, 40(5):1647–1649, 1994.