TKK Dissertations in Information and Computer Science
Espoo 2010 TKK-ICS-D18

ADVANCES IN THE THEORY OF NEAREST NEIGHBOR DISTRIBUTIONS

Elia Liitiäinen

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Faculty of Information and Natural Sciences for public examination and debate in Auditorium T2 at Aalto University School of Science and Technology (Espoo, Finland) on the 22th of October, 2010, at 12 o'clock noon.

Aalto University School of Science and Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science

Aalto-yliopiston teknillinen korkeakoulu
Informaatio- ja luonnontieteiden tiedekunta
Tietojenkäsittelytieteen laitos

Distribution:
Aalto University School of Science and Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science
P.O.Box 15400
FI-00076 Aalto
FINLAND
Tel. +358-9-470-23272
Fax. +358-9-470-23369

ISBN 978-952-60-3304-4 (Print)
ISBN 978-952-60-3305-1 (Online)
ISSN 1797-5050 (Print)
ISSN 1797-5069 (Online)

Multiprint Oy
Espoo 2010

Abstract

Liitiäinen, E. (2010): **Advances in the Theory of Nearest Neighbor Distributions.** Doctoral thesis, Aalto University School of Science and Technology, Dissertations in Information and Computer Science, TKK-ICS-D18, Espoo, Finland.

Keywords: nearest neighbor, computational geometry, entropy, residual variance, correlation, nonparametric

A large part of non-parametric statistical techniques are in one way or another related to the geometric properties of random point sets. This connection is present both in the design of estimators and theoretical convergence studies. One such relation between geometry and probability occurs in the application of non-parametric techniques for computing information theoretic entropies: it has been shown that the moments of the nearest neighbor distance distributions for a set of independent identically distributed random variables are asymptotically characterized by the Rényi entropies of the underlying probability density. As entropy estimation is a problem of major importance, this connection motivates an extensive study of nearest neighbor distances and distributions.

In this thesis, new results in the theory of nearest neighbor distributions are derived using both geometric and probabilistic proof techniques. The emphasis is on results that are useful for finite samples and not only in the asymptotic limit of an infinite sample.

Previously, in the literature it has been shown that after imposing sufficient regularity assumptions, the moments of the nearest neighbor distances can be approximated by invoking a Taylor series argument providing the connection to the Rényi entropies. However, the theoretical results provide limited understanding to the nature of the error in the approximation. As a central result of the thesis, it is shown that if the random points take values in a compact set (e.g. according to the uniform distribution), then under sufficient regularity, a higher order moment expansion is possible. Asymptotically, the result completely characterizes the error for the original low order approximation.

Instead of striving for exact computation of the moments through a Taylor series expansion, in some cases inequalities are more useful. In the thesis, it is shown that concrete upper and lower bounds can be established under general assumptions. In fact, the upper bounds rely only on a geometric analysis.

The thesis also contains applications to two problems in nonparametric statistics, residual variance and Rényi entropy estimation. A well-established nearest neighbor entropy estimator is analyzed and it is shown that by taking the boundary effect into account, estimation bias can be significantly reduced. Secondly, the convergence properties of a recent residual variance estimator are analyzed.

Tiivistelmä

Liitiäinen, E. (2010): **Kehitysaskeleita Lähimmän Naapurin Jakaumien Teoriassa.** Tohtorin väitöskirja, Aalto-yliopiston teknillinen korkeakoulu, Dissertations in Information and Computer Science, TKK-ICS-D18, Espoo, Suomi.

Avainsanat: lähin naapuri, laskennallinen geometria, entropia, residuaalivarianssi, korrelaatio, epäparametrinen

Suuri osa epäparametrisen tilastotieteen tekniikoista liittyy tavalla tai toisella satunnaisten pistejoukkojen geometrisiin ominaisuuksiin. Tämä yhteys on läsnä sekä estimaattoreiden suunnittelussa että teoreettisessa konvergenssianalyysissä. Yksi tällainen suhde geometrian ja todennäköisyyden välillä esiintyy epäparametristen tekniikoiden sovelluksessa informaatioteoreettisten entropioiden laskentaan: on näytetty, että alla olevan tiheysfunktion Rényi entropiat karakterisoivat asymptoottisesti täysin lähimmän naapurin jakaumien momentit joukolle riippumattomia samoin jakautuneita satunnaismuuttujia.

Tässä väitöskirjassa on johdettu uusia tuloksia lähimmän naapurin jakaumien teoriassa käyttäen sekä geometrisia että todennäköisyysteoreettisia todistustekniikoita. Paino on tuloksissa jotka ovat käyttökelpoisia äärellisille näytejoukoille eivätkä vain äärettömän näytteen asymptoottisella rajalla.

Aiemmassa kirjallisuudessa on näytetty, että asettamalla riittävät säännöllisyysoletukset, lähimmän naapurin jakaumien momentteja voidaan approksimoida käyttäen Taylor-sarja argumenttia, jolloin löydetään yhteys Rényi entropioihin. Kuitenkin kyseiset teoreettiset tulokset antavat rajoitetusti ymmärrystä approksimaatiovirheen luonteesta. Väitöskirjan keskeisenä tuloksena on näytetty, että mikäli satunnaiset pisteet ottavat arvoja kompaktissa joukossa (esim. tasajakauman mukaisesti), niin silloin riittävän säännöllisyyden läsnäollessa korkeamman asteen momenttikehitelmä on mahdollinen. Asymptoottisesti tulos karakterisoi täydellisesti virheen alkuperäisessä alhaisen asteen approksimaatiossa.

Sen sijaan, että pyritään tarkkaan momenttien laskentaan Taylor-sarjan avulla, voidaan joissain tapauksissa käyttää myös epäyhtälöitä. Tässä väitöskirjassa näytetään, että voidaan löytää konkreettiset ala- ja ylärajat yleisillä oletuksilla. Itse asiassa ylärajat voidaan johtaa käyttäen ainoastaan geometrista analyysia.

Väitöskirja sisältää myös sovelluksen kahteen ongelmaan epäparametrisessa tilastotieteessä, jotka ovat residuaalivarianssin ja Rényi entropioiden estimointi. Yhtä vakiintunutta lähimmän naapurin estimaattoria analysoidaan ja näytetään, että ottamalla reunavaikutus huomioon, voidaan estimointiharhaa pienentää merkittävästi. Toiseksi, erään viime aikoina keksityn residuaalivarianssi estimaattorin konvergenssiominaisuuksia on analysoitu.

Preface

This work has been carried out in the Laboratory of Computer and Information Science, which during 2008 was merged into the Department of Computer and Information Science. The department was part of Helsinki University of Technology, which since 2010 became Aalto University School of Science and Technology. I would like to acknowledge the department for the grants that funded the two first years of my research. In addition, I received funding from Helsinki Graduate School in Computer Science and Engineering (HeCSE).

The supervisor of this thesis is Professor Olli Simula. I express my thanks for the support he gave to my research. I am also grateful to my instructors, Docent Francesco Corona and Docent Amaury Lendasse, who guided my steps in the academic world. Moreover, I would like to thank the Environmental and Industrial Machine Learning (EIML) group for providing me a supportive and joyful research environment.

I am deeply grateful to the reviewers of the thesis, Professor Mathew D. Penrose and Doctor Dafydd Evans. They helped to greatly clarify the presentation and correct mistakes in the proofs. Based on their comments, I was not only able to improve the thesis, but I also came to understand better the potential of my own research.

I am grateful to my parents and family for all the support they gave me during these four years. Last, I wish to thank my wife Belle Selene Xia for the love she has brought to my life.

Contents

Frequently Used Symbols

λ	Lebesgue measure
$\mathcal{H}^s$	Hausdorff s-measure
$\|\cdot\|_p$	l^p-norm either for vectors or functions
$\|\cdot\|$	l^2-norm either for vectors or functions
$\|\cdot\|_\infty$	the L^∞-norm of a function
$V_{n,p}$	volume of the unit l^p-ball in $\Re^n$
V_n	volume of the unit Euclidean ball in $\Re^n$
S_n	surface area of the unit Euclidean ball
$\rho(x,A)$	distance between x and a set A
$B(x,r)$	open ball of radius r and center x in the given metric
$S(x,r)$	the surface of an Euclidean ball
$\mathcal{X}_r$	the set of points $x \in \Re^n$ with $\rho(x,\mathcal{X}) \leq r$
$\partial_r\mathcal{X}$	the set $(\partial\mathcal{X})_r \cap \mathcal{X}$
$I(\cdot)$	the indicator function of an event
$\nabla_x f$	gradient vector of f at x
$J_x f$	Jacobian of f at x
$O(\cdot)$	big-oh notation, goes to zero as fast as the given term
$o(\cdot)$	small-oh notation, goes to zero strictly faster than the given term
A^0	interior of a set A
A^C	complement of A
$\bar{A}$	closure of A
$A \setminus B$	the set difference of A and B
$\|\mathcal{A}\|$	cardinality of a set $\mathcal{A}$ (for a number, the absolute value)
$x^{(i)}$	i-th component of a vector x
$[\cdot]$	integer part of the number inside the brackets
$\Gamma(\cdot)$	gamma function
$\phi(\cdot)$	digamma function
diam	diameter of a set

Chapter 1

Introduction

Modern society produces vast amount of data due to the emergence of information technology. This has led to the rise of data-based science including machine learning and data-mining in addition to interesting new developments in traditional statistical techniques. The interest in developing efficient and robust methods for data-based modelling and data exploration is increasing; one might even say that there is an on-going revolution in engineering and science.

While new fields of science have emerged, at the same time the borderline between data-based methods and first principles modelling has grown more vague. For example, neural networks and other machine learning methods have emerged as an alternative modelling tool for real-world systems and on the other hand, data-analysis often brings new insights for modelling.

From the engineer's point of view, these developments can be described using the concept of data engineering as demonstrated in Figure 1.1. When facing a real world problem, the first step is to develop theoretical models for the system at hand. Often the difficulty is that in practice there are many factors that are hard to take into account limiting the scope of theoretical considerations. Once a sufficient understanding is achieved, the data-engineer develops models or in general terms practical solutions based on the data that is available. It is common that even if the theoretical understanding is incomplete, data-based methods are able to find good solutions by using a general enough model. Of course the two steps also interact with each other: valid theoretical principles are supported by data and vice versa.

This thesis can be associated with the first step in the aforementioned process. The goal is to develop new theoretical tools and understanding for statistical estimation and modeling. In the spirit of data engineering, the work is of very general nature: most of the results require relatively weak prior assumptions and the emphasis is on methods that can be used in a large class of problems. The theoretical research was motivated by practical needs, but on the other hand the applying research has on its turn been guided by theoretical considerations.

As a specific task, at the early stages the thesis research was orientated towards

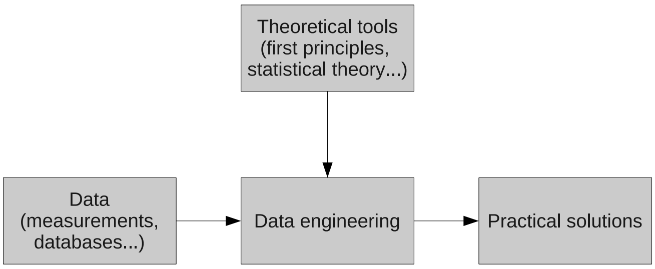

Theoretical tools (first principles, statistical theory...)
Data (measurements, databases...)
Data engineering
Practical solutions

Figure 1.1: The chart demonstrates the concept of data-engineering: theoretical considerations produce a class of possible models, of which the one that fits best the data is chosen using statistical techniques.

providing input selection methods for statistical regression between two random vectors in order to alleviate the effect of the curse of dimensionality. Performing trustworthy input selection is possible only if reliable methods for measuring dependencies between random variables are available. It has turned out that the analysis of nearest neighbor distributions provides a unifying factor behind several important techniques. In the following, the research behind the thesis is introduced in more detail by discussing statistical measures of dependency in order to concretize the reasons behind the study of nearest neighbor distributions and to establish a link to information theory.

1.1 Dependency Measures

A commonly encountered statistical modelling task is the prediction of a target variable Y given a predictor Z. In order to perform the prediction, the underlying relation between Z and Y is modelled using a finite set of observations $(Z_i, Y_i)_{i=1}^M$, where the number of samples M depends on the task at hand. The estimation of the model can be difficult or easy depending on three aspects: the degree of non-linearity present in the relation between Z and Y, the dimensionality (n) of the predictor Z and the number of samples.

To simplify matters, classical statistics often focuses on the case where a linear function from Z to Y is sufficient for a good prediction. In that case n can be relatively large without imposing the requirement of having a very large number of observations available, even though even in linear statistics small dimensionality is preferable. However, when linearity does not hold, things become more complicated as seen from Figure 1.2. When no strong prior assumptions on the non-linearity can be used, resorting to neural networks and basis expansions (e.g. MLP and Gaussian processes [25, 53]) is one of the most important approaches. However, a common denominator of such methods is their sensitivity to the dimensionality n. Currently the seriousness of the problem is debatable, but based e.g. on [25, 67, 66], it is safe to claim that in high dimensional spaces, most non-linear approximators do face problems with approximation capability, stability and computational complexity consequently requiring a large amount of samples in the estimation process.

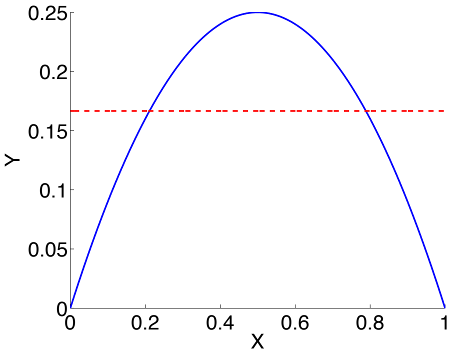


Figure 1.2: The dashed line is the optimal mean square linear fit to the parabola. A linear dependency measure would indicate that X is not a good predictor of Y.

Luckily things are not as gloomy as they might seem to the data engineer who does not have strong prior assumptions. In fact, it is common that even though Z has many components, most of them are not needed for the prediction of Y and a small subset is sufficient. This idea leads to the field of input selection, which vaguely expressed investigates methods for finding a good subset of variables for the prediction of Y. The task decomposes naturally into two subtasks: firstly, an optimization method is used to go through subsets of the components of Z and secondly, a method to evaluate the quality of such subsets is needed.

While the field of input selection arises naturally, it has been investigated surprisingly little. Linear methods, however, are well understood. Considering the random variables

$$X = \begin{bmatrix} Z^{(i_1)} \\ \dots \\ Z^{(i_k)} \end{bmatrix}$$

consisting of components of Z, the most basic linear input selection criterion is the coefficient of determination (R^2-value) of the empirical linear least squares fit between X and Y, even though it is not the only possibility. However, if we look in the univariate case of Figure 1.2, it is seen that no linear method is able to capture the simple parabolic functional dependency. The inability to cope with such non-linear relations suggests the idea of non-linear measures as a possibility for measuring relevance when linearity cannot be assumed similarly as non-linear modelling is a generalization of linear modelling. Of course, the reader familiar with the topic might argue that non-linear modelling can be used to develop such methods; however, as a deep theoretical and practical fact, the two problems should be viewed separately as merely estimating the magnitude of dependencies is easier than building a complete model (see Chapter 6 and [36], for example).

Obviously motivating non-linear dependency measures through input selection for regression is not necessary as linear methods are an important tool in most fields of science. As an example, ecologists and economists use it to seek validation of their theoretical hypothesis. But most of the work in this thesis was in fact originally motivated by the input selection problem largely thanks to the earlier work [1]; later on, more general theoretical ideas then arose as is common in mathematical

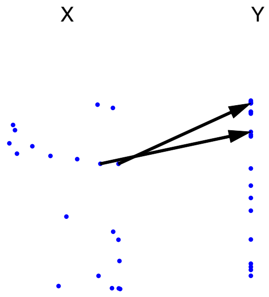


Figure 1.3: In the figure, Y is the first component of X. Points close in the input space tend to be close in the output space as well.

analysis.

1.2 Residual Variance Estimation

Under the definition $m(x) = E[Y|X = x]$ and $r = Y - m(X)$, the target variable Y can be decomposed into

$$Y = m(X) + r.$$

The residual variable r is the part that is not captured by the mean square optimal predictor $m(x)$. In this notation, if the linearity $m(X) = w^T X$ holds, then fitting a linear model into the independent identically distributed (i.i.d) data $(X_i, Y_i)_{i=1}^M$ and computing the coefficient of determination amounts to estimating

$$1 - \frac{\mathrm{Var}[r]}{\mathrm{Var}[Y]}$$

to measure how well X is able to predict Y. The general residual variance estimation problem then suggests itself in a natural way: is it possible to perform the same without assuming linearity? It is an intuitive consequence of the no-free lunch theorem that this is not possible without any prior assumptions. But most real-world phenomena involve at least piecewise smooth functions rendering non-parametric inference of the residual variance a viable option.

Numerous approaches for estimating the residual variance exist; see Chapter 6 for more details on those. However, all of them build at least to some degree on the assumed smoothness of m, which implies that points close to each other in the input space produce similar outputs as demonstrated in Figure 1.3. The most straightforward way to concretize such an idea is to observe that if X_i and $X_{j(i)}$ are two samples close to each other ($j(i)$ is a random index), that is, with $\|X_i - X_{j(i)}\|$ small, then

$$Y_i - Y_{j(i)} \approx (X_i - X_{j(i)})^T \nabla_{X_i} m + r_i - r_{j(i)} \approx r_i - r_{j(i)} \tag{1.1}$$

with $\nabla_{X_i} m$ the gradient of m at the point X_i. Because the approximation is indicated to depend mostly on $r_i - r_{j(i)}$, a natural idea might be to consider ($i > 0$

can be arbitrary as only i.i.d. sampling is discussed here)

$$cE[(Y_i - Y_{j(i)})^2] \approx \frac{c}{M}\sum_{i=1}^{M}(Y_i - Y_{j(i)})^2 \tag{1.2}$$

as an estimate of $\mathrm{Var}[r]$ for some constant $c > 0$. The index $j(i)$ must be chosen in order to ensure

1. The approximation of the expectation in (1.2) is valid.
2. $\|X_i - X_{j(i)}\|$ is small.
3. $E[(r_i - r_{j(i)})^2] \approx c^{-1}\mathrm{Var}[r^2]$.

The validity of point 2 is maximized by choosing $X_{j(i)}$ as the point closest to X_i in the sample $(X_i)_{i=1}^M$ measured in the Euclidean metric. Then $j(i) = N[i,1]$ is called the index of the first nearest neighbor of X_i and $d_{i,1} = \|X_i - X_{j(i)}\|$ is the first nearest neighbor distance. For this choice, point 1 is established in Chapter 4 using a rather standard proof technique. The last point is not obvious, but in [42] it was shown that

$$E[(r_i - r_{j(i)})^2] \approx 2\mathrm{Var}[r]$$

is a valid approximation under certain regularity conditions; this also fixes $c = 1/2$ in Equation (1.2). Unfortunately, those conditions are quite restrictive and generally there is no theoretical guarantees that (1.2) is a good approximation. To solve this problem, the estimator

$$\frac{1}{M}\sum_{i=1}^{M}(Y_i - Y_{N[i,1]})(Y_i - Y_{N[i,2]}) \tag{1.3}$$

has emerged [14, 17, 36] and it is analyzed in detail in Chapter 6 as an application of the theoretical considerations of the thesis. Here $N[i,2]$ refers to the second nearest neighbor of X_i.

In addition to the previous arguments, the choice of using nearest neighbors is motivated by computational efficiency; in fact, by using special data structures, at least in low dimensions the nearest neighbors of each point in $(X_i)_{i=1}^M$ can be found in $O(M \log M)$ time. Nevertheless it should be mentioned that there are other paradigms not directly based on the use of nearest neighbors; the purpose of the thesis is not to review those, even though new tools for that are provided.

1.3 Nearest Neighbor Distributions

In the previous section we saw that whether $Y_i - Y_{N[i,k]}$ is close to $r_i - r_{N[i,k]}$ or not depends on the quantities $(X_i - X_{N[i,k]})^T \nabla_{X_i} m$ if second order terms are neglected. On the other hand

$$E[|(X_i - X_{N[i,k]})^T \nabla_{X_i} m|^2] \leq E[\|\nabla_{X_i} m\|^2 \|X_i - X_{N[i,k]}\|^2] \leq cE[d_{i,k}^2] \tag{1.4}$$

for some $c > 0$ if the gradient is assumed to be a bounded function. By deriving a bound on $E[d_{i,k}^2]$ it is then possible to establish worst-case convergence bounds for the nearest neighbor residual variance estimators (1.2) and (1.3) as demonstrated in Chapter 6. In fact, it is not only due to this particular estimation problem that nearest neighbors are interesting; an analogous consideration applies to many non-parametric statistical methods as well [29].

To bound $E[d_{i,k}^\alpha]$ (introducing $\alpha > 0$ does not complicate matters significantly), it is possible to examine

$$\frac{1}{M}\sum_{i=1}^{M} d_{i,k}^\alpha$$

without any probabilistic arguments. Such an analysis is done in Chapter 2 providing for example the following theorem (among others):

Theorem 1.1. *Suppose that the random vectors $(X_i)_{i=1}^M$ take values in the cube $[0,1]^n$. Then we have for $0 < \alpha \leq n$ and $M > k$,*

$$\frac{1}{M}\sum_{i=1}^{M} d_{i,k}^\alpha \leq (\frac{2^n n^{n/2} k}{M})^{\alpha/n},$$

where distances are measured using the Euclidean norm.

So, in general it can be said that $E[d_{i,k}^\alpha]$ tends to go to zero at least as fast as $M^{-\alpha/n}$. But can it happen that convergence is strictly faster? When the sample $(X_i)_{i=1}^M$ is i.i.d, a negative answer is given in Chapter 2: a constant c_1 with

$$E[d_{1,k}^\alpha] \geq c_1 M^{-\alpha/n}$$

can be found. Moreover, c_1 is found as well, even though the optimality of the bound cannot be claimed.

The upper and lower bounds effectively settle the issue raised in Equation (1.4). However, once we have started in the analysis, it is natural to ask, whether the expectation $E[d_{1,k}^\alpha]$ can be found in some more exact way. It is not hard to find a partial answer in the literature [18, 69]: in general, for i.i.d. variables $(X_i)_{i=1}^M$ with the common density q,

$$M^{\alpha/n} E[d_{1,k}^\alpha] \to V_n^{-\alpha/n} \frac{\Gamma(k+\alpha/n)}{\Gamma(k)} \int_{\Re^n} q(x)^{1-\alpha/n} dx \tag{1.5}$$

in the limit $M \to \infty$ ($\Gamma(\cdot)$ denotes the Gamma function). Effectively this amounts to invoking a locally constant (0th order) approximation of q around X_1 as can be seen from [18]. As a contribution of Chapter 3 and also the central result of the thesis, the possibility of a higher order expansion is demonstrated: under a set of theoretical assumptions, the limit remains the same but an additional term emerges:

$$\begin{aligned} M^{\alpha/n} E[d_{1,k}^\alpha] &= V_n^{-\alpha/n} \frac{\Gamma(k+\alpha/n)}{\Gamma(k)} \int_{\Re^n} q(x)^{1-\alpha/n} dx \\ &\quad + c_2 \frac{\Gamma(k+\alpha/n+1/n)}{\Gamma(k)} M^{-1/n} + o(M^{-1/n}), \end{aligned} \tag{1.6}$$

where c_2 is independent of M and k (see Chapter 3 for details on c_2).

While the discussion started from the bound (1.4), it is in fact true that $(X_i - X_{N[i,k]})^T \nabla_{X_i} m$ has many properties that are not understood only by examining nearest neighbor distances. While the thesis does not comprehensively analyse these, Chapter 6 nevertheless scratches the surface by using the local uniformity of nearest neighbor distributions in order to show that (1.3) has favourable theoretical convergence properties.

1.4 Entropies

While residual variance is a natural measure to estimate how well a given set of variables can be used to predict the target variable, it is still based on the use of a specific cost function (the mean square cost). In many cases, especially when doing data-analysis, it is not necessarily very useful to estimate the relevance in terms of a cost function. Instead, we might want to ask, how much of the randomness in Y is explained by X. If the pair (X, Y) has a joint density $q(x, y)$, then the joint randomness is measured by the differential entropy [58]:

$$H(X,Y) = -\int_{\Re}\int_{\Re^n} q(x,y)\log q(x,y)dxdy; \tag{1.7}$$

the maximal randomness is in fact achieved if the components of (X, Y) are jointly Gaussian and independent of each other. The marginal differential entropy $H(X)$ for X is defined analogously for the marginal distribution $q(x)$ and the conditional entropy is

$$H(Y|X) = H(X,Y) - H(X).$$

Consequently, once an estimate for the entropies $H(X, Y)$ and $H(X)$ is found, then the theoretically attractive measure of relevance $H(Y|X)$ can be computed.

While the information theoretic approach leads naturally to logarithmic entropies when some plausible axioms are accepted, there are however other entropies obtained through relaxing one of the axioms. More specifically, the functions

$$H_\beta(X) = \frac{1}{1-\beta}\log(\int_{\Re^n} q(x)^\beta dx)$$

are called Rényi entropies when $\beta \geq 0$ and while the theoretical foundation behind these is less solid, they are still attractive as measures of randomness in those problems, where deep information theoretic connections are not needed. One observes a connection to nearest neighbor distances: Equation (1.5) yields

$$\frac{n}{\alpha}\log(M^{\alpha/n}\frac{V_n^{\alpha/n}\Gamma(k)}{\Gamma(k+\alpha/n)}E[d_{1,k}^\alpha]) \to H_{1-\alpha/n}(X)$$

showing that a natural finite sample (and well-known [34]) estimate of Rényi entropies is

$$H_{1-\alpha/n}(X) \approx \frac{n}{\alpha}\log(M^{\alpha/n}\frac{V_n^{\alpha/n}\Gamma(k)}{\Gamma(k+\alpha/n)}E[d_{1,k}^\alpha]). \tag{1.8}$$

Because the definition $H_{1-\alpha/n}(Y|X) = H_{1-\alpha/n}(X,Y) - H_{1-\alpha/n}(X)$ still makes sense and is a valid measure of dependency [22] (though the definition is arguable), we have in fact gone through a full circle: the discussion was started from input selection, which (albeit somewhat vaguely) motivated the research on nearest neighbor distributions and finally the asymptotic analysis led to information theory.

The downside of the estimate of Equation (1.8) is its reliance on the low order result (1.5). Since after Chapter 3, the higher order result (1.6) is at our disposal, new possibilities for the estimation emerge. Such techniques are examined in Chapter 5 both theoretically and by simulations. Moreover, not only improvements are suggested, but an analysis of the low-order estimate is provided. Extending the work to the estimation of the differential entropy is done as well.

1.5 Outline and Contributions

In Chapter 2, bounds on the mean nearest neighbor distances on random point sets are derived. The treatment follows our earlier work on the topic in [40]. Originally it came to us as a slight surprise that the closest work to ours is in the field of material physics ([64]), where upper bounds are derived in three dimensions. However, the bounds in [64] do not take the boundary effect into account, and on the other hand, they work only for the first nearest neighbor distance whereas we are interested in k ($k \geq 1$) nearest neighbors. As a second contribution in [40], expected nearest neighbor distances $E[d_{i,k}^{\alpha}]$ are bounded from below using the theory of maximal functions.

Chapter 3 presents a central result of the thesis: an extension to the boundary-corrected expansions of power-weighted mean nearest neighbor distances in our work [38]. In contrast to Chapter 2, the expansion gives exact information in the asymptotic limit of infinite sample size. In this thesis, the boundary correction is shown to hold even if singularities exist at the boundary as long as there are not too many of them.

In Chapter 4, the variance of sums of bounded local functions is analyzed. The theoretical results are rather similar to [16], but somewhat less general. A variance analysis is presented mostly for completeness of the theoretical applications in residual variance and entropy estimation.

In Chapter 5 the boundary corrected expansion is used to derive a bias correction to a class of entropy estimators introduced in [34]. The method was first presented by us in [39] and the treatment here is similar.

Chapter 6 concerns the problem of residual variance estimation as an application of the analysis of nearest neighbor distributions. Recently the product estimator (1.3) has been shown to possess favourable theoretical properties. Using the theory in Chapters 2 and 4, a worst-case convergence analysis is established. We speculated in [41] that the method has a faster rate of convergence than expected by the worst-case analysis. As a theoretical contribution, a formal proof is presented that verifies the hypothesis by exploiting the local uniformity of probability densities.

When reading the thesis, it is best to follow the order of the chapters with the exception that Chapter 6 can be read before 5 depending on the interest of the reader. Moreover, Chapter 6 does not rely on Chapter 3

Chapter 2

Geometry of Nearest Neighbors

2.1 Introduction

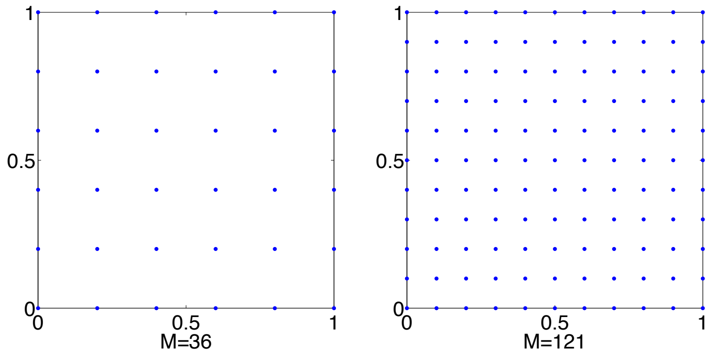
Consider a grid of 36 points in the unit cube $[0,1]^2$ as in Figure 2.1(a). If we take any of the points, then it is easy to convince oneself that the distance to the nearest point in the Euclidean distance is always 0.1. Similarly, in general if M samples are set on an n-dimensional grid, then the nearest neighbor distance is $1/(M^{1/n}-1)$ assuming that $M^{1/n}$ is an integer. Consequently, if a sum over all the first nearest neighbor distances is taken, we have for any $\alpha > 0$,

$$\sum_{i=1}^{M} d_{i,1}^{\alpha} = M^{1-\alpha/n} + o(M^{1-\alpha/n}). \tag{2.1}$$

Is it possible to generalize Equation (2.1) to hold for a larger class of point sets in an inequality form? Based on the grid consideration, one might conjecture vaguely that the left side of Equation (2.1) tends to be of order $M^{1-\alpha/n}$ in most cases. In this chapter, it is shown that in fact, for any set of points in the unit cube $[0,1]^n$, Equation (2.1) generalizes as the inequality

$$\sum_{i=1}^{M} d_{i,k}^{\alpha} \leq c k^{\alpha/n} M^{1-\alpha/n} \tag{2.2}$$

for some constant $c > 0$ independent of M and k. Two geometric proof techniques are used: one based on the concept of instrinsic dimensionality following [32] and another, slightly different technique stemming from material physics [64]. As a theoretical contribution, the extension of [64] to the case $k > 1$ was presented by us in [40] and moreover, by taking the boundary effect into account, a tighter bound was established in a rigorous way. On the other hand, the instrinsic dimension, while often providing bounds far from optimal, generalizes naturally to general metric spaces.



(a) (b)

Figure 2.1: Two point sets with the number of points (M) 36 or 121.

Obviously there does not exist any non-trivial universal lower bounds for the power-weighted sum of distances (2.2) as for example taking M copies of a vector produces a sample with all 1-NN distances equal to zero. To introduce more structure, it is common to examine independent identically distributed (i.i.d) samples in $\Re^n$, which enables the use of probability theoretic tools. Then the direction of the inequality in (2.2) can be reversed on expectation:

$$E[\sum_{i=1}^{M} d_{i,k}^{\alpha}] \geq ck^{\alpha/n} M^{1-\alpha/n} \tag{2.3}$$

for some constant $c > 0$, which depends on n and the common distribution of the points. In fact, the inequality is a consequence of the theory of maximal functions. It seems to appear first time in our work [40].

Figure 2.2 presents an outline of the chapter and it indicates that the sections on upper and lower bounds can be read separately. After the theoretical analysis, in the section related to applications, some known applications of the theory are reviewed.

2.2 Basic Definitions

As the very basic framework in this thesis, we assume that $(X_i)_{i=1}^M$ is a sequence of independent random variables taking values in a metric space $(\mathcal{X}, \rho)$, where $\mathcal{X}$ is the set of elements and ρ the metric on $\mathcal{X}$. Each variable is distributed according to a probability distribution P_i on $\mathcal{X}$. Formally, there is also an underlying probability space $(\Omega, \mathcal{F}, \mathcal{P})$, where the σ-algebra $\mathcal{F}$ defines the set of events and $\mathcal{P}$ is the probability measure.

It is common in statistical literature to require not only independence, but also that the variables $(X_i)_{i=1}^M$ are identically distributed. This is not always done in this thesis, because as a subtle point, dropping this second constraint adds to the generality of the proofs.

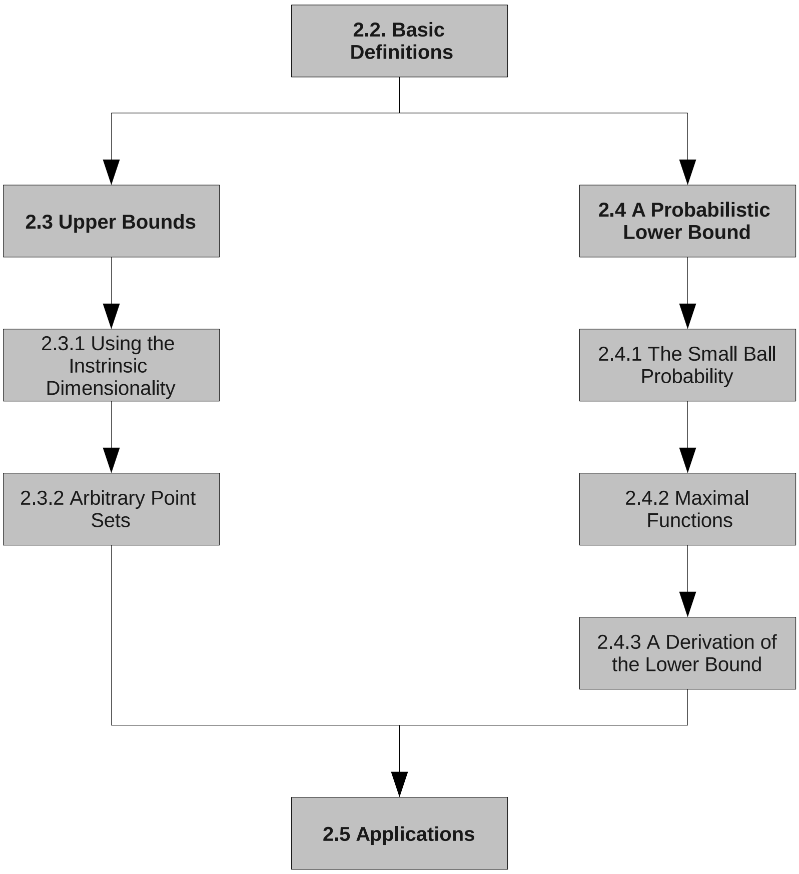

2.2. Basic Definitions
2.3 Upper Bounds
2.4 A Probabilistic Lower Bound
2.3.1 Using the Instrinsic Dimensionality
2.4.1 The Small Ball Probability
2.3.2 Arbitrary Point Sets
2.4.2 Maximal Functions
2.4.3 A Derivation of the Lower Bound
2.5 Applications

Figure 2.2: An outline of Chapter 2.

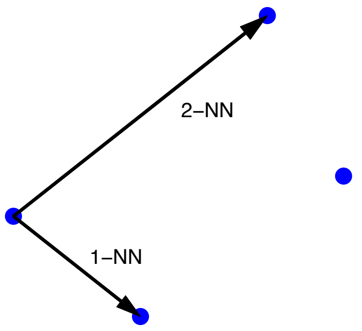


Figure 2.3: 1-NN and 2-NN under the Euclidean distance with $\mathcal{X} = \Re^2$.

The nearest neighbor of a point X_i is defined simply as the point closest to X_i:

$$N[i,1] = \text{argmin}_{1 \leq j \leq M, j \neq i} \rho(X_i, X_j).$$

The k-th nearest neighbor is defined recursively by

$$N[i,k] = \text{argmin}_{1 \leq j \leq M, j \notin \{i, N[i,1], \ldots, N[i,k-1]\}} \rho(X_i, X_j).$$

The concept is intuitive as shown in Figure 2.3 for the Euclidean planar case. Of course, the geometry of the space looks very different with other choices for ρ and $\mathcal{X}$. Especially in high-dimensional vector spaces one must take care as considerations stemming from low-dimensional analysis may lead astray.

Some complications may arise if ties occur, that is, $\rho(X_i, X_j) = \rho(X_i, X_l)$ for some indices $j \neq l$ distinct from i. While for continuous valued random variables, the occurence of two similar distances has zero probability, this is not the case for discrete data and for this reason tie-breaking has deserved attention in the literature [9]. As our purpose is not to delve deeply into this special case due to the additional complications and because we commonly work with distributions possessing a density, the problem of ties is solved by choosing the smallest index among the alternatives.

Using the definition of the k-th nearest neighbor, we may define the corresponding distance by

$$d_{i,k} = \rho(X_i, X_{N[i,k]})$$

and the averaged quantity

$$\delta_{M,k,\alpha} = \frac{1}{M} \sum_{i=1}^{M} d_{i,k}^{\alpha}, \tag{2.4}$$

which is the target of our analysis in this chapter.

Finally, define the distance between $x \in \mathcal{X}$ and a set $\mathcal{C} \subset \mathcal{X}$ by

$$\rho(x, \mathcal{C}) = \inf_{y \in \mathcal{C}} \rho(x, y).$$

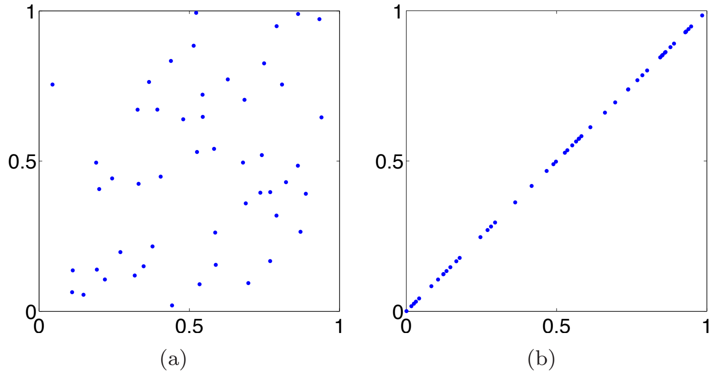


Figure 2.4: (a) presents 50 points sampled according to the uniform distribution on $[0,1]^2$. In (b), uniform sampling was applied on the line $y = x$ to generate the 50 vectors.

If $\mathcal{X} \subset \Re^n$, we also set

$$\mathcal{C}_r = \{x \in \Re^n : \; \exists y \in \mathcal{C} \text{ s.t. } \rho(x,y) \leq r\} \tag{2.5}$$

and $\partial_r \mathcal{C} = (\partial \mathcal{C})_r \cap \mathcal{C}$, where $\partial \mathcal{C}$ refers to the boundary of $\mathcal{C}$.

2.3 Upper Bounds

2.3.1 Using the Instrinsic Dimension

In Figure 2.4 we see two different point sets. In 2.4(a), it is natural to take

$$\mathcal{X} = [0,1]^2 \tag{2.6}$$

as the metric space in which the variables $(X_i)_{i=1}^{50}$ take values. In 2.4(b), the same choice would not describe the situation well as the points are restricted onto the line with unit slope. A better choice is to set instead

$$\mathcal{X} = \{x \in [0,1]^2 : x^{(1)} = x^{(2)}\}. \tag{2.7}$$

Compare now Figures 2.4(a) and 2.4(b) to verify that the nearest neighbor distances are much smaller in (b). Clearly this is due to the fact that the space (2.7) is in some sense smaller than (2.6); in fact, (2.7) is essentially one dimensional being a scaled and rotated version of the interval $[0,1]$. Obviously a good theory should adapt to the special geometric structure in (2.7) and provide significantly smaller bounds for $\delta_{M,k,\alpha}$ compared to the case $[0,1]^2$.

A survey of mathematical literature reveals that there exists various theoretical methods for assessing the dimensionality (and size) of a metric space. To facilitate the choice, the following mathematical observation is useful.

Lemma 2.1. *For any choice of $\mathcal{X}$ and realization $(X_i)_{i=1}^M$, it holds that*

$$B(X_1, d_{1,1}/2) \cap B(X_2, d_{2,1}/2) = \emptyset.$$

Proof. If $x \in B(X_1, d_{1,1}/2) \cap B(X_2, d_{2,1}/2)$, then

$$\rho(X_1, X_2) \leq \rho(X_1, x) + \rho(x, X_2) < \frac{1}{2}d_{1,1} + \frac{1}{2}d_{2,1} \leq \max\{d_{1,1}, d_{2,1}\}$$

with a contradiction. □

Heuristically, Lemma 2.1 implies that the presence of many large 1-NN distances is possible only as long as the corresponding disjoint balls fit inside the metric space. This suggests the use of packing numbers, which bound the maximal number of disjoint balls with given radius inside $\mathcal{X}$.

Definition 2.1. *A set $\mathcal{A} \subset \mathcal{X}$ is an r-packing, if for all distinct points $x, y \in \mathcal{A}$, $\rho(x, y) \geq r$. For $r > 0$, we define the r-packing number as the cardinality of the maximal packing, that is:*

$$N_{packing}(r) = \sup_{\mathcal{A} \textit{ is an } r\textit{-packing}} |\mathcal{A}|.$$

In basic cases, the packing numbers are able to capture the geometric structure of the underlying space rather well as demonstrated by

Example 2.1. *Consider the unit cube $\mathcal{X} = [0,1]^2$ under the Euclidean metric. Let $\{x_1, \ldots, x_l\}$ be an r-packing of $\mathcal{X}$ with $0 < r < 1$. Then the balls $B(x_i, r/2)$ and $B(x_j, r/2)$ are disjoint when $i \neq j$ and also $\lambda(B(x_i, r/2) \cap [0,1]^2) \geq 2^{-4} V_2 r^{-2}$, where λ refers to the Lebesgue measure. We may write*

$$2^{-4} l V_2 r^2 \leq \sum_{i=1}^{l} \lambda(B(x_i, r/2) \cap [0,1]^2) = \lambda([0,1]^2 \cap \cup_{i=1}^{l} B(x_i, r/2)) \leq 1$$

implying the inequality $l \leq [2^4 V_2^{-1} r^{-2}]$ (the operation $[\cdot]$ can be used here because l is an integer) and consequently $N_{packing}(r) \leq [2^4 \pi^{-1} r^{-2}]$. For the line in Equation (2.7), we may consider the Hausdorff 1-measure $\mathcal{H}^1$ (length) to obtain similarly

$$2^{-1} l r \leq \sum_{i=1}^{l} \mathcal{H}^1(B(x_i, r/2) \cap \mathcal{X}) = \mathcal{H}^1(\mathcal{X} \cap \cup_{i=1}^{l} B(x_i, r/2)) \leq \sqrt{2}$$

and $N_{packing}(r) \leq [2^{3/2} r^{-1}]$ when $0 < r < \sqrt{2}$.

Fix a realization of the sample $(X_i)_{i=1}^M$ and a real number $r > 0$. Moreover, assume that $i_1, \ldots, i_l$ indicate points with $d_{i_j,1} \geq r$. Then $(X_{i_j})_{j=1}^l$ is an r-packing of $\mathcal{X}$ and it must be that $l \leq \min\{M, N_{packing}(r)\}$. Thus, we have a concrete method to bound the amount of points with the first nearest neighbor larger than some threshold. In order to apply the considerations to

$$\delta_{M,1,\alpha} = \frac{1}{M} \sum_{i=1}^{M} d_{i,1}^{\alpha}, \tag{2.8}$$

the following lemma is useful.

Lemma 2.2. *Let $\mu(dx)$ be a probability measure on $\Re$. For any bounded measurable function $f : \Re \to [0, \infty)$ and $p > 0$, we have*

$$\int_{\Re} f(x)^p \mu(dx) = p \int_0^\infty r^{p-1} \mu(\{x \in \Re : f(x) > r\}) dr. \tag{2.9}$$

Proof. The proof can be found in [56] in Theorem 8.16; the reference contains also a compact introduction to measure theory. A short derivation of Equation (2.9) is presented here for the sake of completeness: by the right continuity of the integrand

$$\begin{aligned} &p \int_0^\infty r^{p-1} \mu(\{x \in \Re : f(x) > r\}) dr \\ &\quad = p \lim_{h \to 0} h \sum_{i=1}^\infty (hi)^{p-1} \mu(\{x \in \Re : f(x) > h(i+1)\}). \end{aligned}$$

Because partial integration can be applied to sums as well as to continuous integrals, we have

$$\begin{aligned} h \sum_{i=1}^\infty (hi)^{p-1} \mu(\{x \in \Re : f(x) > h(i+1)\}) &= -h \sum_{i=1}^\infty [\mu(\{x \in \Re : f(x) > h(i+2)\}) \\ &\quad - \mu(\{x \in \Re : f(x) > h(i+1)\})] \sum_{j=1}^{i} (hj)^{p-1} \\ &= h^p \sum_{i=1}^\infty \mu(\{x \in \Re : h(i+1) < f(x) \le h(i+2)\}) \sum_{j=1}^{i} j^{p-1}. \end{aligned} \tag{2.10}$$

When i grows, the approximation $\sum_{j=1}^{i} j^{p-1} \approx p^{-1} i^p$ becomes increasingly accurate and on the other hand, terms with a small i can be neglected in the limit $h \to 0$; thus

$$\begin{aligned} &\lim_{h \to 0} h^p \sum_{i=1}^\infty \mu(\{x \in \Re : h(i+1) < f(x) \le h(i+2)\}) \sum_{j=1}^{i} j^{p-1} \\ &\quad = \lim_{h \to 0} p^{-1} \sum_{i=0}^\infty \mu(\{x \in \Re : h(i+1) < f(x) \le h(i+2)\}) (hi)^p \\ &\quad = p^{-1} \int_0^\infty f(x)^p \mu(dx). \end{aligned}$$

□

Applying Lemma 2.2 and the discussion before it to $\delta_{M,1,\alpha}$ (viewing the sum as an integral) yields

$$\begin{aligned} \delta_{M,1,\alpha} &\le \alpha \int_0^{\text{diam}[\mathcal{X}]} r^{\alpha-1} |\{1 \le i \le M : d_{i,1} > r\}| dr \\ &\le \alpha \int_0^{\text{diam}[\mathcal{X}]} r^{\alpha-1} \min\{M, N_{packing}(r)\} dr. \end{aligned} \tag{2.11}$$

It remains to replace $N_{packing}(r)$ with an appropriate expression to generate concrete bounds. Example 2.1 indicates that assuming

$$N_{packing}(r) \leq [cr^{-n}] + 1 \tag{2.12}$$

when $r > 0$ (observe that $N_{packing}(r) \geq 1$) covers a rather large number of finite dimensional spaces. We also note that the well-known packing dimension is the infimum of those $n > 0$ for which the bound (2.12) can be established for some $c > 0$ [65]. Thus, this choice seems appropriate in order to turn (2.11) into the main result of this section

In addition to being just a summary of the discussion this far, the proof contains a generalization to $k > 1$. A proof based on the idea appeared first time in [32].

Theorem 2.1. *Assume that for some constants $C_n, n > 0$, $N_{packing}(r) \leq [C_n r^{-n}] + 1$ when $r > 0$. Then for $0 < \alpha < n$ and $M \geq k(C_n + 1)$,*

$$\delta_{M,k,\alpha} \leq \frac{n}{n-\alpha}\left(\frac{C_n k}{M}\right)^{\alpha/n} - \frac{\alpha C_n^{\alpha/n} k}{(n-\alpha)M} + \frac{C_n^{\alpha/n} k}{M}. \tag{2.13}$$

For $\alpha = n$, we have the bound

$$\delta_{M,k,n} \leq \left(2 + \log(\frac{M}{k})\right)\frac{C_n k}{M}.$$

Proof. As a preliminary observation, the diameter of $\mathcal{X}$ is bounded by $C_n^{1/n}$ due to the fact that the existence of two points $x, y \in \mathcal{X}$ with $\rho(x,y) \geq C_n^{1/n} + \epsilon$ for some $\epsilon > 0$ would indicate that $N_{packing}(C_n^{1/n} + \epsilon) \geq 2$.

Choose arbitrary $r > 0$ and define the set of indices

$$I_r = \{0 < i \leq M : \ d_{i,k} > r\}.$$

When $k = 1$, we know that I_r is an r-packing. For $k > 1$, things are slightly more complicated but it is shown next that I_r can be turned into an r-packing by removing points.

Choose $i_1 \in I_r$ and define the set $I_{r,1} = I_r \setminus \{N[i_1, 1], \ldots, N[i_1, k-1]\}$. Then pick up $i_2 \neq i_1$ ($i_2 \in I_{r,1}$) and set $I_{r,2} = \{i_1\} \cup (I_{r,1} \setminus \{N[i_2, 1], \ldots, N[i_2, k-1]\})$. Correspondingly,

$$I_{r,3} = \{i_1, i_2\} \cup (I_{r,2} \setminus \{N[i_3, 1], \ldots, N[i_3, k-1]\})$$

with $i_3 \neq i_1, i_2$. To explain in words, in each iteration a point is chosen from the active set and its nearest neighbors are removed up to the index $k-1$ (excluding the previously chosen points). Then, this chosen point is added to the set $\{i_j\}$. By repeating the aforementioned procedure as long as possible, we construct the sets $\{I_{r,j}\}_{j=1}^{L}$ for some $L \geq |I_r|/k$. By construction each index in the sequence $(i_j)_{j=1}^{L}$ is in $I_{r,L}$.

Choose now $i, j \in I_{r,L}$ with $i \neq j$ and notice that from the properties of $I_{r,L}$ it follows that $\rho(X_i, X_j) \geq r$ showing that it is an r-packing.

Now the proof proceeds similarly as Equation (2.11) was derived: $I_{r,L}$ contains by construction $L \geq |I_r|/k$ points, which implies that the cardinality of $|I_r|$ is bounded by

$$|I_r| \leq kL \leq kN_{packing}(r) \leq kC_n r^{-n} + k.$$

Under the assumption that $M \geq kC_n$, we have using Lemma 2.2 ($0 < \alpha < n$):

$$\begin{aligned}
\delta_{M,k,\alpha} &= \int_0^\infty \alpha M^{-1} r^{\alpha-1} |I_r| dr \leq \int_0^{C_n^{1/n}} \alpha M^{-1} r^{\alpha-1} \min\{C_n r^{-n} k + k, M\} dr \\
&= C_n^{\alpha/n} k^{\alpha/n} M^{-\alpha/n} + C_n^{\alpha/n} k M^{-1} + \int_{C_n^{1/n} k^{1/n} M^{-1/n}}^{C_n^{1/n}} \alpha C_n k M^{-1} r^{\alpha-1-n} dr \\
&= \frac{n}{n-\alpha} C_n^{\alpha/n} k^{\alpha/n} M^{-\alpha/n} + C_n^{\alpha/n} k M^{-1} - \frac{\alpha C_n^{\alpha/n}}{n-\alpha} k M^{-1}.
\end{aligned}$$

When $\alpha = n$, the calculation is nearly similar, but the integral of r^{-1} produces a logarithm:

$$\begin{aligned}
\delta_{M,k,n} &\leq 2C_n k M^{-1} + \int_{C_n^{1/n} k^{1/n} M^{-1/n}}^{C_n^{1/n}} nkC_n M^{-1} r^{-1} dr \\
&= C_n k M^{-1} (2 + \log \frac{M}{k}).
\end{aligned}$$

□

Example 2.2. *In the case $\mathcal{X} = [0,1]^2$, Example 2.1 and Theorem 2.1 together imply the inequality*

$$\delta_{M,k,1} \leq 8\pi^{-1/2} \left(\frac{k}{M}\right)^{1/2}$$

when $0 < \alpha < n$. For the space in Equation (2.7) similar considerations reveal that

$$\delta_{M,k,1} \leq 2^{3/2} \left(2 + \log \frac{M}{k}\right) \frac{k}{M}.$$

It is of importance to observe that n does not have to be an integer. Thus, Theorem 2.1 applies to many fractal sets of non-integer dimension. Such sets commonly occur for example as attractors of non-linear differential equations.

2.3.2 Arbitrary Point Sets

While Theorem 2.1 adapts well to instrinsic dimensionality, it suffers from suboptimality in the limit $\alpha \to n$, because the term

$$\frac{n}{n-\alpha}$$

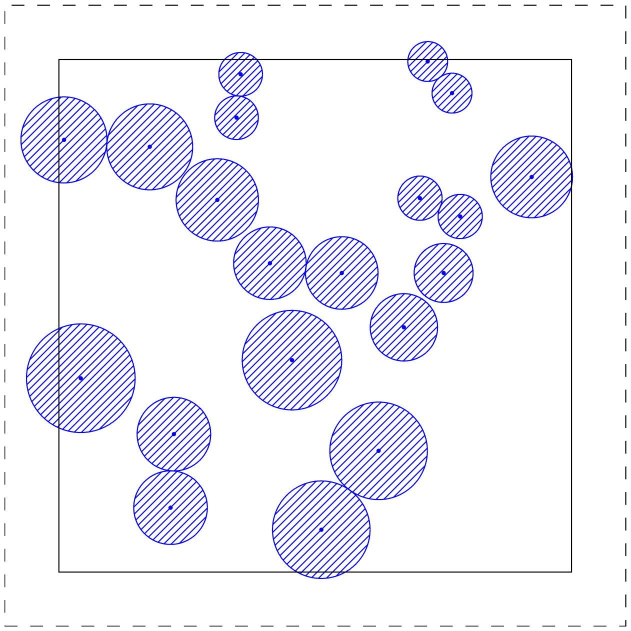


Figure 2.5: If D is the maximal 1-NN distance, then the balls $B(X_i, d_{i,1}/2)$ drawn in the figure belong to the set $[-D/2, 1+D/2]^2$.

approaches infinity providing rather unfavourable constants. Moreover, with $\alpha = n$ one might ask if the additional logarithm can be avoided.

It seems that to address these issues more assumptions are required. At this point the requirement of generality is dropped and the focus is shifted to worst-case bounds on the unit cube $[0,1]^n$ (n is from now on an integer) as stated by

(A1) $\mathcal{X}$ is a subset of $[0,1]^n$ and $\rho(x,y) = \|x-y\|_p$ for some $p \geq 1$ (the l^p-norm).

In contrast to general metric spaces, the concept of volume is now available. It can be employed using the observation that by Lemma 2.1, the sum

$$\sum_{i=1}^{M} d_{i,1}^n$$

is proportional to the volume of $\cup_{i=1}^{M} B(X_i, d_{i,1}/2)$, which is contained in a cube with side length determined by the largest nearest neighbor distance as drawn in Figure 2.5. The generalization to $k > 1$ requires some additional work; moreover, the theory becomes more accurate when large distances are handled separately, which motivates the cut-off in the following lemma.

Lemma 2.3. *Suppose that (A1) holds. Then for any $0 < \alpha \leq n$, $M > k$ and $r > 0$,*

$$\frac{1}{M}\sum_{i=1}^{M} d_{i,k}^{\alpha} I(d_{i,k} \leq r) \leq \left(\frac{2^n \lambda(\mathcal{X}_{r/2})k}{V_{n,p}M}\right)^{\alpha/n} \tag{2.14}$$

with I the indicator function of an event and λ the Lebesgue measure.

Proof. Choose any $x \in \Re^n$. Let us make the counterassumption that there exists $k+1$ points, denoted by $X_{i_1}, \ldots, X_{i_{k+1}}$ (the indices being distinct), such that $x \in B(X_{i_j}, d_{i_j,k}/2)$ (B refers to a ball in the l^p-norm) for $j = 1, \ldots, k+1$. Let $(i_j, i_{j'})$ be the pair that maximizes the distance $\|X_{i_j} - X_{i_{j'}}\|_p$. Then the triangle inequality yields

$$\|X_{i_j} - X_{i_{j'}}\|_p \leq \|X_{i_j} - x\|_p + \|x - X_{i_{j'}}\|_p < \frac{1}{2} d_{i_j,k} + \frac{1}{2} d_{i_{j'},k}.$$

On the other hand, by the definition of the pair $(i_j, i_{j'})$,

$$\begin{aligned} \|X_{i_j} - X_{i_{j'}}\|_p &= \frac{1}{2}\|X_{i_j} - X_{i_{j'}}\|_p + \frac{1}{2}\|X_{i_j} - X_{i_{j'}}\|_p \\ &= \frac{1}{2} \max_{1 \leq j' \leq k+1} \|X_{i_j} - X_{i_{j'}}\|_p + \frac{1}{2} \max_{1 \leq j \leq k+1} \|X_{i_j} - X_{i_{j'}}\|_p \\ &\geq \frac{1}{2} d_{i_j,k} + \frac{1}{2} d_{i_{j'},k} \end{aligned}$$

leading to a contradiction. Thus, we have for the sum of indicator functions

$$\begin{aligned} &\sum_{i=1}^{M} \int_{\Re^n} I(x \in B(X_i, d_{i,k}/2), d_{i,k} \leq r) dx \\ &\qquad = \int_{\mathcal{X}_{r/2}} \sum_{i=1}^{M} I(x \in B(X_i, d_{i,k}/2), d_{i,k} \leq r) dx \leq \lambda(\mathcal{X}_{r/2}) k. \end{aligned}$$

On the other hand,

$$\sum_{i=1}^{M} \int_{\Re^n} I(x \in B(X_i, d_{i,k}/2), d_{i,k} \leq r) dx = 2^{-n} V_{n,p} \sum_{i=1}^{M} d_{i,k}^n I(d_{i,k} \leq r)$$

implies that

$$\frac{1}{M} \sum_{i=1}^{M} d_{i,k}^n I(d_{i,k} \leq r) \leq 2^n V_{n,p}^{-1} \lambda(\mathcal{X}_{r/2}) k M^{-1}.$$

The generalization to $0 < \alpha < n$ is easiest by Jensen's inequality [56]:

$$\frac{1}{M} \sum_{i=1}^{M} d_{i,k}^\alpha I(d_{i,k} \leq r) \leq \left[\frac{1}{M} \sum_{i=1}^{M} d_{i,k}^n I(d_{i,k} \leq r) \right]^{\alpha/n},$$

which implies (2.14). □

By dropping the constraint $d_{i,k} \leq r$, one obtains straightforwardly the following corollary.

Corollary 2.1. *Suppose that (A1) holds. Then we have for $0 < \alpha \leq n$ and $M > k$,*

$$\delta_{M,k,\alpha} \leq \left(\frac{2^n n^{n/p} k}{M} \right)^{\alpha/n}.$$

Proof. The diameter of $\mathcal{X} = [0,1]^n$ in the l_p-norm is $n^{1/p}$, which implies that setting $r = n^{1/p}$ in Lemma 2.3 is equivalent to not having the cut-off at all. Moreover, $\mathcal{X}_{n^{1/p}/2}$ is contained in the ball of radius $n^{1/p}$ and center at the point $(\frac{1}{2}, \dots, \frac{1}{2})$, which has the volume $V_{n,p} n^{n/p}$. □

Compared to Theorem 2.1, the logarithmic factor is avoided for $\alpha = n$. Moreover, when α is close to n, the constant in front of $M^{-\alpha/n}$ is much smaller.

If one wants to proceed to the direction of deriving as tight a bound as possible, then the constraints $d_{i,k} \leq r$ in Lemma 2.3 become useful:

Theorem 2.2. *Suppose that (A1) holds and $0 < \lambda(\mathcal{X}_r) \leq \lambda(\mathcal{X}) + c_1 r$ when $r \leq c_2$ for some constants $c_1, c_2 > 0$. Then for any $0 < \alpha < n$ and $M > k$,*

$$\begin{aligned} \delta_{M,k,\alpha} &\leq \inf_{0 \leq r \leq n^{1/p}} \left(\left(\frac{2^n \lambda(\mathcal{X}_{r/2}) k}{V_{n,p} M} \right)^{\alpha/n} + \frac{2^n n^{n/p} r^{\alpha - n} k}{M} \right) \\ &= (\frac{2^n \lambda(\mathcal{X}) k}{V_{n,p} M})^{\alpha/n} + O(M^{-\frac{\alpha}{n}(\frac{n-\alpha+n\alpha^{-1}}{n-\alpha+1})}). \end{aligned} \tag{2.15}$$

Proof. The proof follows that in [40]. We already know by Lemma 2.3 that

$$\frac{1}{M} \sum_{i=1}^{M} d_{i,k}^{\alpha} I(d_{i,k} \leq r) \leq (\frac{2^n \lambda(\mathcal{X}_{r/2}) k}{V_{n,p} M})^{\alpha/n}.$$

A straightforward application of Corollary 2.1 to the subsample $I = \{1 \leq i \leq M : d_{i,k} > r\}$ yields

$$\frac{1}{M} \sum_{i=1}^{M} I(d_{i,k} > r) d_{i,k}^{\alpha} \leq 2^{\alpha} n^{\alpha/p} |I|^{1-\alpha/n} k^{\alpha/n} M^{-1}.$$

The bound holds also in the special case $|I| \leq k$ as it is always true that $d_{i,k} \leq n^{1/p}$. The first inequality in (2.15) follows now by Chebyshev's inequality and Corollary 2.1:

$$|I| \leq r^{-n} \sum_{i=1}^{M} d_{i,k}^{n} \leq 2^n n^{n/p} r^{-n} k M^{-1},$$

because r can be any value between 0 and $n^{1/p}$. To see the second result, choose $r = M^{-\frac{n-\alpha}{n^2-\alpha n+n}}$ and use the approximation $(1+x)^{\alpha/n} \approx 1 + \frac{\alpha}{n} x$ valid for small x. □

An analogous result to Theorem 2.2 appeared in [64], but the boundary effect was neglected as not significant, which is the case when M is very large.

The condition $\lambda(\mathcal{X}_r) \leq \lambda(\mathcal{X}) + c_1 r$ requires some regularity of the boundary $\partial\mathcal{X}$ and works poorly with fractal boundaries. It is similar to condition C.2 in [18].

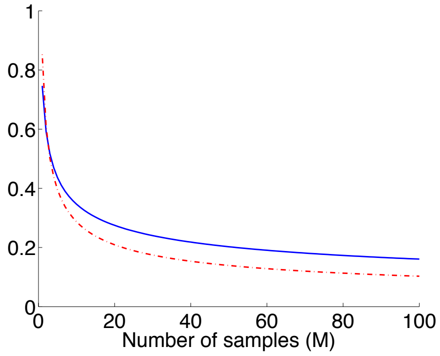


Figure 2.6: A demonstration of the bounds in Corollary 2.1 (the solid line) and Theorem 2.2.

Nevertheless, such a bound holds for many commonly encountered sets; for example, if $\mathcal{X} = [0,1]^n$ we have

$$\lambda(\mathcal{X}_{r/2}) - \lambda(\mathcal{X}) \leq (1+r)^n - 1 = nr + O(r^2)$$

and the open ball $B(0,1)$ satisfies a similar bound. It is clear that the influence of points close to the boundary grows once the dimensionality of the space becomes bigger. To demonstrate the achieved improvement compared to the direct application of Lemma 2.3 with $r = \sqrt{n}$ under the Euclidean metric, both bounds are drawn in Figure 2.6 with $n = 3$, $k = 1$, $p = 2$, $\alpha = 1$ and $\mathcal{X} = [0,1]^3$ using the estimate $\lambda(\mathcal{X}_{r/2}) \leq (r+1)^3$ in (2.15).

The main use of Theorem 2.2 comes from that fact that it is already rather tight and it looks likely that further improvements would require probabilistic arguments as it is hard to improve the geometric method. The request for a probabilistic approach stems from the asymptotic results in [18] and Chapter 3, which the interested reader can compare to Theorem 2.2.

The analysis assumes that the space $\mathcal{X}$ is bounded. However, at least theoretically it is of interest to ask, whether similar bounds hold even if the boundeness condition is replaced by a condition on the moments $E[\|X_i\|^\beta]$ (again, a probabilistic condition is needed). A result to this direction was proven in [36].

2.4 A Probabilistic Lower Bound

2.4.1 The Small Ball Probability

The small ball probability function is often a useful concept when working with nearest neighbors because of its distribution-free properties. Assuming that (A1) holds, in the given l^p-norm it is defined as the probability mass of a ball with radius r and center x:

$$\omega_x(r) = P(X_1 \in B(x,r)), \tag{2.16}$$

which of course makes sense only if the variables $(X_i)_{i=1}^M$ are i.i.d. It is a remarkable fact that the distribution of $\omega_{X_1}(d_{1,k})$ is independent of the common distribution of $(X_i)_{i=1}^M$ or the choice of p as shown by the following theorem, which has been proven for example in [18].

Theorem 2.3. *Assume that (A1) holds and the sample $(X_i)_{i=1}^M$ is i.i.d. with a common density q w.r.t. (with respect to) the Lebesgue measure λ. Then for $\alpha > 0$,*

$$E[\omega_{X_1}(d_{1,k})^\alpha | X_1] = \frac{\Gamma(k+\alpha)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha)},$$

where $\Gamma(\cdot)$ refers to the Gamma function.

Proof. Choose $0 < z < 1$ and set $t = \inf\{s > 0 : \ \omega_{X_1}(s) > z\}$. By continuity, $\omega_{X_1}(t) = z$. Thus $\omega_{X_1}(d_{1,k}) > z$ if and only if there are at most $k-1$ points in the set $B(X_1, t)$. A combinatorial argument yields (with probability one)

$$\begin{aligned} P(\omega_{X_1}(d_{1,k}) > z | X_1) &= \sum_{j=0}^{k-1} \binom{M-1}{j} \omega_{X_1}(t)^j (1 - \omega_{X_1}(t))^{M-j-1} \\ &= \sum_{j=0}^{k-1} \binom{M-1}{j} z^j (1-z)^{M-j-1}. \end{aligned} \tag{2.17}$$

Now, we have using the formula

$$\binom{M-1}{j} = \frac{\Gamma(M)}{\Gamma(M-j)\Gamma(j+1)}$$

and some basic identities for beta functions,

$$\begin{aligned} \binom{M-1}{j} \int_0^1 z^{j+\alpha-1}(1-z)^{M-j-1} dz &= \binom{M-1}{j} \frac{\Gamma(j+\alpha)\Gamma(M-j)}{\Gamma(M+\alpha)} \\ &= \frac{\Gamma(j+\alpha)\Gamma(M)}{\Gamma(j+1)\Gamma(M+\alpha)}. \end{aligned} \tag{2.18}$$

Theorem 8.16 in [56] (to represent the conditional expectation) implies that

$$\begin{aligned} E[\omega_{X_1}(d_{1,k})^\alpha | X_1] &= \alpha \int_0^1 z^{\alpha-1} P(\omega_{X_1}(d_{1,k}) > z | X_1) dz \\ &= \alpha \sum_{j=0}^{k-1} \binom{M-1}{j} \int_0^1 z^{j+\alpha-1} (1-z)^{M-j-1} dz \\ &= \alpha \sum_{j=0}^{k-1} \frac{\Gamma(j+\alpha)\Gamma(M)}{\Gamma(j+1)\Gamma(M+\alpha)} = \frac{\Gamma(k+\alpha)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha)}. \end{aligned} \tag{2.19}$$

The last equality can be proven by an induction argument. □

For a fixed k the small ball probability behaves as $M^{-\alpha}$, because

$$\frac{\Gamma(M)}{\Gamma(M+\alpha)} = M^{-\alpha} + O(M^{-\alpha-1}) \tag{2.20}$$

as shown by

Lemma 2.4. *For any fixed $\sigma > 0$,*

$$\frac{\Gamma(M)}{\Gamma(M+\sigma)} = M^{-\sigma} + O(M^{-\sigma-1}).$$

Proof. The proof here is a shortened version of that in Lemma 3.7 of [13]. For any fixed $\sigma \geq 0$ and all $M > 0$, Stirling's formula for Gamma functions yields the approximation

$$\Gamma(M+\sigma) = \sqrt{\frac{2\pi}{M+\sigma}}(\frac{M+\sigma}{e})^{M+\sigma}[1+O(M^{-1})]. \tag{2.21}$$

Equation (2.21) can be applied to the ratio of two Gamma functions:

$$\frac{\Gamma(M)}{\Gamma(M+\sigma)} = (1+\frac{\sigma}{M})^{-M-\sigma}\sqrt{\frac{M+\sigma}{M}}\frac{1+O(M^{-1})}{1+O(M^{-1})}e^{\sigma}M^{-\sigma}. \tag{2.22}$$

For small $x > 0$, $\log(1+x)$ can be expanded as $\log(1+x) = x + O(x^2)$ yielding

$$\begin{aligned}(1+\frac{\sigma}{M})^{M+\sigma} &= e^{(M+\sigma)\log(1+\sigma M^{-1})} = e^{(M+\sigma)\sigma M^{-1}+\sigma^2 M^{-2}} = e^{\sigma+O(M^{-1})} \\ &= e^{\sigma} + O(M^{-1}). \end{aligned} \tag{2.23}$$

Equation (2.23) substituted into (2.22) yields

$$\frac{\Gamma(M)}{\Gamma(M+\sigma)} = \sqrt{\frac{M+\sigma}{M}}M^{-\sigma}[1+O(M^{-1})].$$

And finally

$$\sqrt{\frac{M+\sigma}{M}} = 1 + O(M^{-1});$$

consequently

$$\frac{\Gamma(M)}{\Gamma(M+\sigma)} = M^{-\sigma} + O(M^{-\sigma-1}).$$

□

2.4.2 Maximal Functions

Suppose that $\mathcal{X} = \Re^n$ with ρ the l^p-metric and that the sample $(X_i)_{i=1}^M$ is i.i.d. possessing a common density q w.r.t. the Lebesgue measure. If we want to analyze $d_{1,k}^{\alpha}$ (notice that by the i.i.d. assumption we may fix the index 1), then the distribution-free properties of the small ball probability function (2.16) are attractive. To reach a transformation of the problem, one may write

$$V_{n,p}^{\alpha/n}d_{1,k}^{\alpha} = \omega_{X_1}(d_{1,k})^{\alpha/n}(\frac{V_{n,p}d_{1,k}^n}{\omega_{X_1}(d_{1,k})})^{\alpha/n} = \omega_{X_1}(d_{1,k})^{\alpha/n}(\frac{\omega_{X_1}(d_{1,k})}{V_{n,p}d_{1,k}^n})^{-\alpha/n}.$$

If the interest is only on bounding the expectation $E[d_{1,k}^{\alpha}]$ from below, then it is useful to observe that

$$\frac{\omega_{X_1}(d_{1,k})}{V_{n,p}d_{1,k}^n} \leq \sup_{0<r<\infty} \frac{\omega_{X_1}(r)}{V_{n,p}r^n}. \tag{2.24}$$

Those familiar with measure theory recognize the expression in the right side of (2.24) as the maximal function of q:

$$\mathcal{M}(x) = \sup_{r>0} \frac{\int_{B_p(x,r)} q(y)dy}{V_{n,p}r^n} = \sup_{r>0} \frac{\omega_x(r)}{V_{n,p}r^n}.$$

The theory of maximal functions is in fact extensive and and appears in most basic treatises of real analysis such as [56]. $\mathcal{M}(x) : \Re^n \to [0,\infty]$ is a positive measurable function (see e.g. [56]). An important goal in the theory of maximal functions is the characterization of the conditions under which $\mathcal{M}$ is integrable. A satisfying answer to this question is given by the following classical result, which tells that finiteness of certain L^p-norms of q is sufficient.

Lemma 2.5. *Choose any $s > 1$. The maximal function of q is bounded by*

$$E[\mathcal{M}(X_i)] \leq 3^{n/s}e^{1/s}(\frac{s^2}{s-1})^{1/s}\|q\|_s\|q\|_{s'}.$$

The scalar s' refers to the conjugate of s found by solving the equation $\frac{1}{s} + \frac{1}{s'} = 1$.

Proof. By Hölder's inequality,

$$E[\mathcal{M}(X_i)] = \int_{\Re^n} q(x)\mathcal{M}(x)dx \leq \|q\|_{s'}\|\mathcal{M}\|_s.$$

The end of the proof relies on the result by Hardy and Littlewood, which can be found for example in Theorem 8.18 of [56]. It says that

$$\|\mathcal{M}\|_s \leq 3^{n/s}e^{1/s}(\frac{s^2}{s-1})^{1/s}\|q\|_s.$$

□

2.4.3 A Derivation of the Lower Bound

Section 2.4.2 gives a lower bound for $E[d_{i,1}^{\alpha}]$ in terms of the maximal function. What remains is an application of Jensen's inequality as demonstrated by

Theorem 2.4. *Suppose that (A1) holds and the variables $(X_i)_{i=1}^M$ are i.i.d. with a common density q w.r.t. the Lebesgue measure. Then the inequality*

$$\begin{aligned} E[d_{1,k}^{\alpha}] &\geq V_{n,p}^{-\alpha/n}E[\mathcal{M}(X_1)]^{-\alpha/n}\frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)} \\ &\geq 3^{-\alpha/2}2^{-\alpha/n}e^{-\alpha/2n}\|q\|_2^{-2\alpha/n}V_{n,p}^{-\alpha/n}\frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)} \end{aligned} \tag{2.25}$$

holds for $\alpha > 0$.

Proof. By Theorem 2.3,

$$\begin{aligned}E[d_{1,k}^{\alpha}] &\geq V_{n,p}^{-\alpha/n} E[\mathcal{M}(X_1)^{-\alpha/n} \omega_{X_1}(d_{1,k})^{\alpha/n}] \\ &= V_{n,p}^{-\alpha/n} E[\mathcal{M}(X_1)^{-\alpha/n}] \frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)}.\end{aligned}$$

The proof is completed by observing the fact that by Jensen's inequality

$$E[\mathcal{M}(X_1)^{-\alpha/n}] \geq E[\mathcal{M}(X_1)]^{-\alpha/n}$$

and applying Lemma 2.5 with $s = 2$. □

The second inequality in (2.25) is of worst-case nature and in fact, some debate about the optimal constant for the bound on maximal functions is still going on. For this reason, it is sometimes a good idea to examine the maximal function directly e.g. when considering the uniform distribution.

Example 2.3. *If (A1) holds and q is uniform on the set $\mathcal{X} \subset \Re^n$, then the first inequality in (2.25) yields*

$$E[d_{1,k}^{\alpha}] \geq V_{n,p}^{-\alpha/n} \lambda(\mathcal{X})^{\alpha/n} \frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)}.$$

In the literature it is known that the inequality becomes an approximate equality in the limit $M \to \infty$. Due to being non-asymptotic, it might be useful in many applications as demonstrated in [35].

One might ask, whether upper bounds in the spirit of Theorem 2.4 are possible. If one tries to proceed into this direction, one probably needs to impose regularity on q.

2.5 Applications

The bounds have importance in discrete and random geometry. Below, we mention some already established applications.

Convergence analysis in non-parametric statistics

In non-parametric statistics, the inequalities of Section 2.3 serve as a useful tool, often allowing a relatively general setting. In fact, the bounds first occured as a method to analyze nearest neighbor classifiers under arbitrary sampling [32]. As another example, in [28] a probabilistic upper bound was used as an important building block in the analysis of the rate of convergence. As one direction of research, one might want to generalize the analysis there to manifolds using the theory in Section 2.3.1.

In Chapter 6, a worst-case rate of convergence is provided for an estimator of residual variance. Again, Section 2.3 turns out to be useful there. The lower bound seems to find less use in this context, but it should be mentioned that in the aforementioned [35] it was employed to demonstrate the curse of dimensionality for a noise variance estimator.

Geometry of high dimensional spaces

Given the i.i.d. sample $(X_i)_{i=1}^M$ on $[0,1]^n$ and a point $q \in \Re^n$, consider the quantity

$$\Delta_q = \frac{\max_{i=1,\ldots,M} \|X_i - q\|_p}{\min_{i=1,\ldots,M} \|X_i - q\|_p}.$$

It is a non-intuitive fact that Δ_q tends to be close to 1 when the dimensionality n is large as shown in [4]; in fact, this holds for any p-norm. Such considerations have deserved attention in the theory of nearest neighbor search [4, 19, 52], because they show that distances lose their significance in high dimensional spaces. The analysis is done in the context that both M and n approach infinity in some appropriate proportion. Interestingly, Δ_q relates to nearest neighbor distances by

$$\Delta_q = \frac{d_{q,M}}{d_{q,1}},$$

where $d_{q,k}$ refers to the nearest neighbor distances of the point q. By Corollary 2.1 and Theorem 2.4,

$$\frac{E[d_{1,M}]}{E[d_{1,1}]} \geq c n^{1/p} V_{n,p}^{1/n} M^{1/n} \tag{2.26}$$

for some $c > 0$ independent of n and M; moreover, it was stated in [21] that

$$\limsup_{n\to\infty} n^{1/p} V_{n,p}^{1/n} \leq 2(ep)^{1/p}.$$

Consequently, if $M^{1/n}$ is large, then the contrast Δ_X tends to be large as well with a high probability given an independent random variable X distributed similarly as the variables $(X_i)_{i=1}^M$; of course, one still has to translate this consideration in terms of Δ_X to take into account that Equation (2.26) has expectations in the fraction. This was done in [21], where new instability results were derived using our original work in [40] contributing to the general thread of research.

Chapter 3

Asymptotic Results for Nearest Neighbors

3.1 Introduction

Suppose that the points $(X_i)_{i=1}^M$ are i.i.d. with some common density q. In Chapter 2 we derived inequalities for the sum

$$\sum_{i=1}^{M} d_{i,1}^{\alpha}.$$

When more probabilistic structure is introduced, it is possible to go deeper and in fact determine the exact asymptotic $M \to \infty$ behavior. Two aspects arise naturally: expectation and variance. The focus here is on the former, whereas the latter is considered in Chapter 4.

As a central result in nearest neighbor analysis, it has been shown that in the Euclidean space with various other assumptions (see Section 3.3),

$$M^{\alpha/n} E[d_{1,k}^{\alpha}] \to V_n^{-\alpha/n} \frac{\Gamma(k+\alpha/n)}{\Gamma(k)} \int_{\mathcal{X}} q(x)^{1-\alpha/n} dx; \tag{3.1}$$

the sufficient conditions for this convergence can be said to be quite well-understood. The result establishes a remarkable connection to information theory as discussed in more detail in Chapter 6. However, Equation (3.1) involves in fact only a low order approximation; for finite M, there are other factors that are significant.

More precisely, there are two sources of error:

1. The derivation of Equation (3.1) is based on a local linearization of q, without taking into account higher order terms in the Taylor expansion.
2. q is taken as smooth on the set $\mathcal{X}$ but not on the whole space $\Re^n$; consequently, the boundaries $\partial\mathcal{X}$ may include points of non-smoothness rendering a linearization argument challenging.

Whether 1 or 2 is more relevant depends on the adopted setting. The mathematical setting here involves assuming that q is smooth and strictly above zero with $\mathcal{X}$ a bounded set (e.g. the uniform distribution in the unit cube). Then, somewhat surprisingly, point 2 is the most relevant source of approximation error. In fact, as the main result of the chapter, it is shown that under sufficient regularity of $\mathcal{X}$ and q,

$$\begin{aligned} M^{\alpha/n} E[d_{1,k}^{\alpha}] &= V_n^{-\alpha/n} \frac{\Gamma(k+\alpha/n)}{\Gamma(k)} \int_{\mathcal{X}} q(x)^{1-\alpha/n} dx \\ &+ (D - V_n^{-\alpha/n-1/n}) \frac{\Gamma(k+\alpha/n+1/n)}{\Gamma(k)} M^{-1/n} \int_{\partial\mathcal{X}} q(x)^{1-\alpha/n-1/n} dS \\ &+ O(M^{-2/n} \log^{2+2\alpha/n+4/n} M). \end{aligned}$$

The second term in the right side shows that the boundary effect contributes a surface integral. The constant D is computed later; it does not depend on k or M. The expansion has appeared before in [38]; in this chapter an extension to that result is provided by adopting more general regularity conditions.

The logarithmic distance has been analyzed in a similar fashion as the expectation (3.1) and it is known that under some regularity,

$$E[\log d_{1,k}] = -\frac{1}{n} \int_{\mathcal{X}} q(x) \log q(x) dx + \frac{1}{n}(\psi(k) - \psi(M) - \log V_n) + o(1), \tag{3.2}$$

where ψ denotes the digamma function. Even though it has not appeared in the literature before, the boundary correction extends straightforwardly to the logarithmic case. The derivation is demonstrated in parallel with the analysis for the α-moments.

The chapter is divided into eight parts as shown in Figure 3.1. In Section 3.2 basic definitions and assumptions are stated. The assumptions exclude unbounded probability measures and require some smoothness of the densities. In Section 3.3 previous work and the main contributions are overviewed and placed in relation to the adopted setting.

In Section 3.4 and 3.5, nearest neighbor distributions and the general proof technique are presented. After that, we proceed to the development of rigorous analysis close to boundaries and in the interior. As common, the theoretical proofs involve general ideas that can find uses in other contexts as well.

3.2 Assumptions and Definitions

Because the boundaries $\mathcal{X}$ have a significant role in a higher order nearest neighbor analysis, regularity assumptions need to be imposed on $\partial\mathcal{X}$. To achieve this, some basic concepts from differential geometry are useful.

A nonempty subset $\mathcal{S} \subset \Re^n$ is called an $(n-1)$-dimensional submanifold, if for each $x \in \mathcal{S}$ there exists $\epsilon > 0$ and a homeomorphism

$$\phi : U \to \mathcal{S} \cap B(x, \epsilon)$$

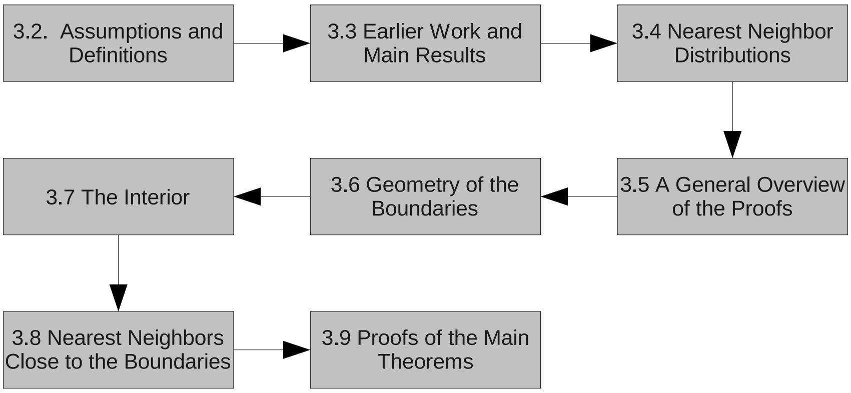

3.2. Assumptions and Definitions
3.3 Earlier Work and Main Results
3.4 Nearest Neighbor Distributions
3.5 A General Overview of the Proofs
3.6 Geometry of the Boundaries
3.7 The Interior
3.8 Nearest Neighbors Close to the Boundaries
3.9 Proofs of the Main Theorems

Figure 3.1: Main parts of Chapter 3.

with U an open subset of $\Re^{n-1}$. Recall that a homeomorphism is a bijection with both ϕ and ϕ^{-1} continuous. This means that the set $\mathcal{S}$ is in a sense $n-1$ dimensional even if it is a subset of $\Re^n$; as a standard example consider for example the surface of a sphere. Because all submanifolds that will be encountered are $n-1$ dimensional, we simply refer to submanifolds without stating the dimensionality explicitly.

A submanifold is said to be twice continuously differentiable, if the local parametrization ϕ can be chosen as twice continuously differentiable on U and the Jacobian $J_y\phi$ has linearly independent columns for all $y \in U$.

Choose $x \in \mathcal{S}$ and $x = \phi(z)$ for some $z \in \Re^{n-1}$. When the Jacobian $J_z\phi$ exists, we may define a subspace G_x as the span of the columns of $J_z\phi$. G_x is the tangent plane at x; for our purposes, another convenient set is the shifted plane

$$T_x = x + G_x = \{y \in \Re^n : \ y = x + \tilde{x} \text{ for some } \tilde{x} \in G_x\}$$

in set aritmetic notation, which will be continuously invoked in this chapter. With two overlapping submanifolds, we use the notation $G_x^{\mathcal{S}}$ to specify that the submanifold $\mathcal{S}$ is meant.

It is possible to show that the tangent plane is invariant with respect to the choice of ϕ as it should be for the definition to make sense. The normal vector $n(x)$ is defined as the unit vector orthogonal to G_x. Notice that there are two possible directions for the normal; this orientation problem will be solved in later sections. The orientation is important as the half-planes

$$\mathcal{U}_x = \{x + y : y \in \Re^n \text{ and } y^T n(x) \leq 0\} \tag{3.3}$$

are used later. In any case the normal is continuous in the sense that regardless of the orientation,

$$\min\{\|n(x) - n(x_n)\|, \|n(x) + n(x_n)\|\} \to 0 \tag{3.4}$$

when $x_n \to x$ in the limit $n \to \infty$.

In our work [38] we required that the boundary $\partial\mathcal{X}$ is a twice continuously differentiable compact submanifold. Such an assumption works well because for any

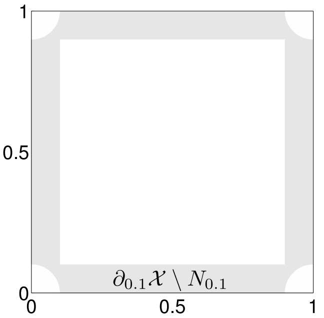


Figure 3.2: The set $\partial_{0.1}\mathcal{X} \setminus N_{0.1}$ when $\mathcal{X} = (0,1)^2$.

$x \in \partial\mathcal{X}$ the boundary can be locally approximated with the tangent plane T_x. However, in addition to the exclusion of most unbounded sets (due to the compactness requirement), many important sets have only a piecewise smooth boundary. For example, the unit cube $(0,1)^n$ has a boundary, which is non-smooth in the intersection of any two faces.

Even if the unit cube does not have a completely smooth boundary, the set of singularities is heuristically speaking small. For example, in the planar $\mathcal{X} = (0,1)^2$ case, the set of singularities (call it N) consists of the four corner points of the rectangle. Then if $\delta > 0$ is small, the amount of points closer than δ to N is small as seen from $\lambda(\partial_\delta(0,1)^2) = 4\delta - 4\delta^2$ and $\lambda(N_\delta \cap (0,1)^2) = \pi\delta^2$ (see Figure 3.2). For n-dimensional hypercubes we can similarly say that $\lambda(\partial_\delta(0,1)^n) \leq 2n\delta$ and $\lambda(N_\delta \cap (0,1)^n) \leq n^2\delta^2$. Corresponding considerations hold in terms of surface area (the Hausdorff $(n-1)$-measure in $\Re^n$):

$$\mathcal{H}^{n-1}(N_\delta \cap \partial(0,1)^n) \leq 2n^2\delta.$$

Because nearest neighbor distributions are of local nature, it is more or less intuitive that once the amount of points close to the problematic points of non-smoothness is small, they have a small effect on nearest neighbor distances. Consequently, it happens that the proof techniques in [38] remain valid albeit with some additional technicalities. The following assumption almost achieves the goal of summarizing the small set of singularities property of the cube.

(A2) $\mathcal{X} \subset \Re^n$ is an open and bounded set with $n \geq 2$, ρ is the Euclidean distance ($\rho(x,y) = \|x - y\|$) and the boundary $\partial\mathcal{X}$ is an $(n-1)$-dimensional submanifold (not necessarily differentiable), which can be represented as the union of l disjoint twice continuously differentiable submanifolds denoted by $\{\mathcal{C}_i\}_{i=1}^l$ and a closed set N. We require that

1. The Hausdorff measure (surface area) $\mathcal{H}^{n-1}(N_\delta \cap \partial\mathcal{X})$ is bounded by

$$\mathcal{H}^{n-1}(N_\delta \cap \partial\mathcal{X}) \leq c\delta$$

for some constant c (depending only on $\mathcal{X}$) and all $0 < \delta < 1$.

2. The volume of the set of points close to N is bounded by

$$\lambda(N_\delta \cap \mathcal{X}) \leq c\delta^2,$$

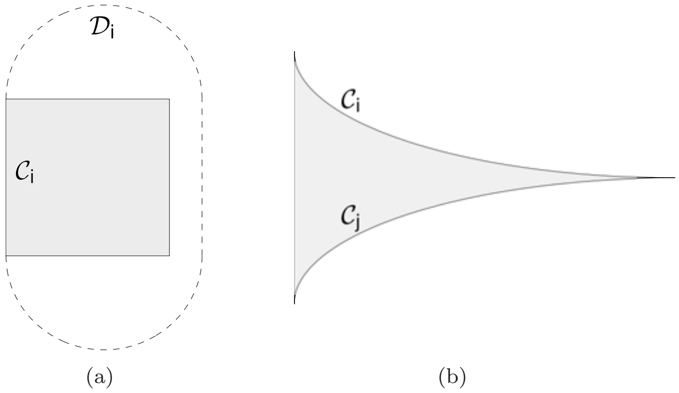


Figure 3.3: (a) The face $\mathcal{C}_i$ denotes here the leftmost line of the rectangle excluding the end points. $\mathcal{D}_i$ completes $\mathcal{C}_i$ into a smooth compact submanifold. (b) Assumption (A2) does not hold, because at the corner points the tangent planes on $\mathcal{C}_i$ and $\mathcal{C}_j$ become parallel.

again for a constant $c > 0$ and $0 < \delta < 1$.

3. There exist compact twice differentiable submanifolds $\{\mathcal{D}_i\}_{i=1}^l$ such that $\mathcal{C}_i = \mathcal{D}_i \cap \partial\mathcal{X} \setminus N$. Moreover, for any pair $\mathcal{D}_i$, $\mathcal{D}_j$ with $i \neq j$, the tangent planes at the intersection of the two submanifolds are not parallel to each other: $G_x^{\mathcal{D}_i} \neq G_x^{\mathcal{D}_j}$.
4. For each i and $x \in \mathcal{C}_i$, there exists $\delta > 0$ such that $B(x, \delta) \cap \mathcal{D}_i \subset \mathcal{C}_i$. This means that each $\mathcal{C}_i$ is open relative to $\mathcal{D}_i$.

The set $\mathcal{X}$ is assumed to be open as a technical detail. For the cube $(0,1)^n$ it is natural to choose

$$\mathcal{C}_{2i} = \{x \in [0,1]^n : \ x^{(i)} = 0 \text{ and } 0 < x^{(j)} < 1 \text{ when } j \neq i\}$$

and

$$\mathcal{C}_{2i-1} = \{x \in [0,1]^n : \ x^{(i)} = 1 \text{ and } 0 < x^{(j)} < 1 \text{ when } j \neq i\}.$$

Then as explained before, points 1 and 2 hold for $N = [0,1]^n \setminus (\cup_{i=1}^{2n} \mathcal{C}_i)$ and point 4 holds as well.

Condition 3 is demonstrated in Figure 3.3(a) for the rectangle $(0,1)^2$. The assumption that each piece of the boundary is a subset of some compact smooth submanifold is useful, because it allows a smooth parametrization of $\partial\mathcal{X}$ even at points of non-smoothness. Figure 3.3(a) shows an extension consisting of two half-circles; the idea generalizes to the n-dimensional hypercube. Because the submanifolds $\{\mathcal{D}_i\}_{i=1}^l$ are allowed to intersect each other outside $\partial\mathcal{X}$, their existence is not such a strong assumption.

In Figure 3.3(b) we find an example where condition 3 in (A2) is not valid, because the tangent planes of the intersecting faces $\mathcal{C}_i$ and $\mathcal{C}_j$ become parallel. To the eye, it is most clear for the rightmost corner point; however, in fact the two other intersections also involve tangent planes becoming parallel.

For the cube $(0,1)^n$ it is on the other hand clear that neighboring faces do not become parallel on points of intersection as they are in fact orthogonal to each other. Thus we have verified that (A2) holds for $(0,1)^n$; in addition, it is clear that it also holds for sets with a smooth, compact boundary. Consequently, it allows a larger class of sets than the setting in [38] even though the case of unbounded probability distributions remains open. Moreover, (A2) is by no means elegant and most likely it is possible to simplify it.

Even though (A2) guarantees a large degree of regularity, it is useful to have the following at disposal as well. It might follow from (A2), but there is no rigorous proof at the moment.

(A3) $\mathcal{X} \subset \Re^n$ and

$$\inf_{x \in \mathcal{X}, 0<r<1} \lambda(B(x,r) \cap \mathcal{X}) r^{-n} > 0.$$

Often all the regularity provided by (A2) is not needed if an analysis of the boundary effect is not the goal. Then, instead of (A2) and (A3) it may be sufficient to work with (A3) and

A2') $\mathcal{X} \subset \Re^n$ is a closed set. Denoting by $\partial_r \mathcal{X}$ the set $\partial_r \mathcal{X} = (\partial \mathcal{X})_r \cap \mathcal{X}$, it holds that

$$\sup_{0<r<1} r^{-1} \lambda(\partial_r \mathcal{X}) < \infty.$$

In contrast to Chapter 2, it is essential in this chapter that $(X_i)_{i=1}^M$ is i.i.d. with a regular common density q. Here two kind of regularity of q is needed: the domain must be regular and the function must be smooth on the domain. (A2)-(A3) ensure the first whereas the second relates to Hölder continuity:

Definition 3.1. *Assuming that $\mathcal{X} \subset \Re^n$ and $0 < \gamma \leq 1$, the set of Hölder continuous functions $H(c,\gamma)$ is defined as the class of bounded scalar valued functions f with the property*

$$|f(x) - f(y)| \leq c\|x-y\|^\gamma \tag{3.5}$$

for all $x, y \in \mathcal{X}$. For $1 < \gamma \leq 2$, $H(c,\gamma)$ is the set of scalar valued functions f on $\mathcal{X}$ such that f is differentiable in the interior $\mathcal{X}^0$ of $\mathcal{X}$, $f \in H(1,c)$ and the gradient ∇f is Hölder continuous (as a vector valued function) on $\mathcal{X}^0$ with the exponent $\gamma - [\gamma]$.

The following states the regularity condition on q:

(A4) The variables $(X_i)_{i=1}^M$ are i.i.d. possessing a common density q w.r.t. the Lebesgue measure λ on the closure $\bar{\mathcal{X}}$ with $\mathcal{X} \subset \Re^n$ an open set and $q \in H(c_1,\gamma)$ for some $0 < \gamma \leq 2$ and $c_1 > 0$. Moreover, we assume that $q \geq c_2$ on $\mathcal{X}$ for some constant $c_2 > 0$.

The closure is taken as the domain of definition for q to ensure it to be well-defined on $\partial \mathcal{X}$.

The following definitions are needed:

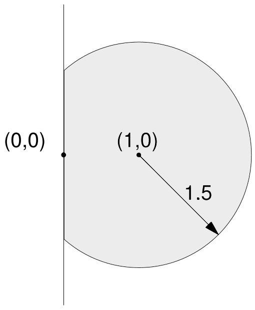


Figure 3.4: $W(1.5)$ is computed as the area of the gray region.

Definition 3.2. *The half-plane of points with a positive first coordinate is denoted by H:*

$$H = \{(s, x) : \ s \in \Re, x \in \Re^{n-1} \text{ and } s \geq 0\}.$$

Obviously, (s, x) is understood as a concatenation. W is defined as the volume of the intersection between a ball of radius r centered at $(1, 0)$ and H:

$$W(r) = \lambda(B((1, 0), r) \cap H).$$

$W(r)$ is used to model the behavior of $\omega_x(r)$ (Equation (2.16)) close to the boundaries of $\mathcal{X}$. In words, H denotes the set of points right from the y-axis and $W(r)$ is just the area of the intersection between the ball centered at $(1, 0)$ and H. This is demonstrated in Figure 3.4. It is obvious that $W(r) = V_n r^n$ when $0 < r < 1$, but for $r > 1$ the situation is more complicated as shown by

Example 3.1. *When $n = 3$, the function W is given by*

$$W(r) = \frac{4}{3}\pi r^3$$

when $0 < r < 1$ and

$$W(r) = \frac{2}{3}\pi r^3 - \frac{1}{3}\pi + \pi r^2$$

for $r > 1$.

3.3 Earlier Work and Main Results

An introduction to general random local geometry can be found in the book [47]. One can also follow individual references; a review of recent developments with applications can be found in [50]. To understand the thesis, it is not necessary to go deeply into that direction, but it is good to know that in [69] such theories were used to compute asymptotic limits for average power-weighted nearest neighbor distances. For example, the following was proven:

Theorem 3.1. *Suppose that the i.i.d. vectors $(X_i)_{i=1}^M$ take values in a convex polyhedron $\mathcal{X} \subset \Re^n$. Then if the common density q is bounded from below and above on $\mathcal{X}$, we have for any $\alpha > 0$,*

$$M^{\alpha/n} E[d_{1,k}^{\alpha}] \to V_n^{-\alpha/n} \frac{\Gamma(k+\alpha/n)}{\Gamma(k)} \int_{\mathcal{X}} q(x)^{1-\alpha/n} dx \tag{3.6}$$

as $M \to \infty$.

Alternatively, the asymptotic limit (3.6) can also be established if $(X_i)_{i=1}^M$ is i.i.d, $0 < \alpha < n$ and

$$\int_{\Re^n} q(x)^{1-\alpha/n} dx < \infty$$

instead of the restrictive convexity and positivity requirements. Theorem 3.1 can be viewed as the most general law of large numbers for nearest neighbor distances providing the asymptotic limit when $0 < \alpha < n$ and thus establishing the connection to information theory through Rényi entropies as discussed in more detail in Chapter 5.

Similar generality has not been achieved for $\alpha > n$; probably it is because in that case

$$\int_{\Re^n} q(x)^{1-\alpha/n} dx$$

is often unbounded (e.g. Gaussians). Nevertheless the convexity condition in Theorem 3.1 seems restrictive and can be relaxed to some degree as long as q is bounded from below by a constant larger than zero.

Many possible ways to extend Theorem 3.1 to different directions exist. One can for example prove central limit theorems (e.g. [48]) or try to work under as weak assumptions as possible. Here the focus is on analyzing the rate of convergence w.r.t. M of which [69] does not as such provide much information. [18] provides an approximation in the order of magnitude notation together with imposing relatively few restrictions on $\mathcal{X}$:

Theorem 3.2. *Suppose that $(A2'), (A3)$ and $(A4)$ hold with $0 < \gamma \leq 1$ in (A4). Then, keeping any k, $\alpha > 0$ and $\rho > 0$ fixed,*

$$E[d_{1,k}^{\alpha}] = V_n^{-\alpha/n} \frac{\Gamma(M)\Gamma(k+\alpha/n)}{\Gamma(M+\alpha/n)\Gamma(k)} \int_{\mathcal{X}} q(x)^{1-\alpha/n} dx + o(M^{-\alpha/n-\gamma/n+\rho}), \tag{3.7}$$

where the remainder term goes to zero faster than $M^{-\alpha/n-\gamma/n+\rho}$ with respect to M.

To interpret Theorem 3.2, recall Equation (2.20):

$$\frac{\Gamma(M)}{\Gamma(M+\alpha/n)} = M^{-\alpha/n} + O(M^{-\alpha/n-1}).$$

In the special case of bounded (from below and above) and smooth densities, Theorem 3.2 improves on 3.1, because convexity is not needed anymore. It also implies the asymptotic rate of convergence $O(M^{-\alpha/n-\gamma/n+\rho})$ to the limit and there are good reasons to believe that the actual rate is in fact $O(M^{-\alpha/n-\gamma/n})$.

Theorem 3.2 is a first-order approximation as it is based on approximating q locally by a constant and neglecting the boundary effect. Of the various higher order terms, it is of interest to ask which ones are dominant, and how much can be computed in closed-form. The following theorem shows the perhaps surprising result that in the presence of boundaries, it is the boundary effect that causes the largest error in the approximation. Moreover, the related higher order term takes a simple form in the final expansion.

Theorem 3.3. *Suppose that (A2)-(A4) hold with $1 < \gamma \leq 2$ in (A4). Then for fixed k,*

$$\begin{aligned} E[d_{1,k}^{\alpha}] =& V_n^{-\alpha/n} \frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)} \int_{\mathcal{X}} q(x)^{1-\alpha/n} dx \\ &+ (D - V_n^{-\alpha/n-1/n}) \frac{\Gamma(k+\alpha/n+1/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n+1/n)} \int_{\partial\mathcal{X}} q(x)^{1-\alpha/n-1/n} dS \\ &+ O(M^{-\gamma/n-\alpha/n} \log^{2+2\alpha/n+4/n} M), \end{aligned} \tag{3.8}$$

where the constant D is the integral (recall Definition 3.2)

$$D = \frac{1}{n} \int_0^1 a^{-\alpha-2} W(a^{-1})^{-\alpha/n-1/n-1} W'(a^{-1}) da$$

with W' the derivative of W.

Theorem 3.3 appeared first time in our work [38]. However, as mentioned in Section 3.2, instead of (A2) it was assumed that $\partial\mathcal{X}$ is a compact twice differentiable submanifold or a polytope; both are special cases of (A2). To the best of our knowledge, comparable earlier results are few; one should mention the work in [51], which can be used for a similar higher order expansion for minimal spanning trees with the simplification $\mathcal{X} = [0,1]^2$ and $q = 1$ on the unit square (uniform distribution).

It is remarkable that the second term in the right side of Equation (3.8) captures the boundary effect in a simple way as a surface integral. The surface integral is rigorously understood to be taken w.r.t. the Hausdorff measure $\mathcal{H}^{n-1}$ on $\partial\mathcal{X}$, but the identification

$$dS = \sqrt{|(J_z\phi)^T J_z\phi|} dz$$

for a local parameterization ϕ can be used to compute such integrals.

The constant D can be represented by the change of variable $y = a^{-1}$ and partial integration:

$$\begin{aligned} D =& \frac{-1}{\alpha+1} \int_1^{\infty} r^{\alpha} \frac{d(W(r)^{-\alpha/n-1/n})}{dr} dr = \frac{1}{\alpha+1} W(1)^{-\alpha/n-1/n} \\ &+ \frac{\alpha}{\alpha+1} \int_1^{\infty} r^{\alpha-1} W(r)^{-\alpha/n-1/n} dr \geq \frac{V_n^{-\alpha/n-1/n}}{\alpha+1} (1 + \alpha \int_1^{\infty} r^{-2} dr) \\ =& V_n^{-\alpha/n-1/n}. \end{aligned}$$

In the last inequality, the fact $W(r) \leq V_n r^n$ was employed. It follows that $D - V_n^{-\alpha/n-1/n} \geq 0$ in Equation (3.8) implying that the boundary effect increases nearest neighbor distances. It is difficult to associate D with a geometric interpretation because it arises from analytical considerations, but at least D can be computed numerically when necessary.

There is also an unspecified error term in Theorem 3.3. Most likely the correct order of magnitude is $M^{-\gamma/n-\alpha/n}$ without the logarithmic factor, but showing it would make the proof more complicated and in any case the Big-Oh notation guarantees small error only for large M. The small sample case is examined via experiments when necessary.

Example 3.2. *As a concrete example, let us analyze uniformly distributed points in the unit ball with $\alpha = 1$. In such a case $q = V_3^{-1}$ and Example 3.1 together with a numerical evaluation gives*

$$D \approx 0.42.$$

Setting $k = 1$, Equation (3.8) takes the form

$$\begin{aligned} E[d_{1,1}] \approx &\Gamma(4/3)M^{-1/3} + 3DV_3^{2/3}\Gamma(5/3)M^{-2/3} - 3\Gamma(5/3)M^{-2/3} \\ &+ O(M^{-1}\log^4 M). \end{aligned}$$

While the moments of $d_{1,k}$ are definitely very interesting, an equally interesting quantity is $\log d_{1,k}$ due to its close connection to the differential entropy ([30]). The theory behind the expectation $E[\log d_{1,k}]$ differs from $E[d_{1,k}^\alpha]$ only in technical details even though the extension was not done in [38]. A modification of Theorem 3.3 to fit into this case results in

Theorem 3.4. *Suppose that Assumptions (A2)-(A4) hold with $1 < \gamma \leq 2$ in (A4). Then for a fixed k,*

$$\begin{aligned} E[\log d_{1,k}] =&C_1(M,k)\frac{\Gamma(M)}{\Gamma(M+1/n)}\int_{\partial\mathcal{X}} q(x)^{1-1/n}dS - n^{-1}\int_{\mathcal{X}} q(x)\log q(x)dx \\ &+ C_2(M,k) + O(M^{-\gamma/n}\log^{3+4/n} M). \end{aligned}$$

The variables $C_1(M,k)$ and $C_2(M,k)$ are (ψ refers to the digamma function)

$$\begin{aligned} C_1(M,k) =& \frac{V_n^{-1/n}\Gamma(k+1/n)\log V_n}{n\Gamma(k)} + \frac{V_n^{-1/n}\psi(M+1/n)\Gamma(k+1/n)}{n\Gamma(k)} \\ &+ D_1(\psi(k+1/n) - \frac{\psi(M+1/n)\Gamma(k+1/n)}{\Gamma(k)}) - \frac{V_n^{-1/n}\psi(k+1/n)}{n} \\ &+ \frac{D_2\Gamma(k+1/n)}{\Gamma(k)} \\ C_2(M,k) =& \frac{1}{n}(\psi(k) - \psi(M) - \log V_n) \\ D_1 =& \frac{1}{n^2}\int_0^1 a^{-2}W(a^{-1})^{-1/n-1}W'(a^{-1})da \\ D_2 =& -\frac{1}{n^2}\int_0^1 a^{-2}W(a^{-1})^{-1/n-1}W'(a^{-1})(\log W(a^{-1}) + n\log a)da. \end{aligned}$$

Informally, Theorem 3.4 is obtained from Theorem 3.3 by taking derivative w.r.t. α at the point $\alpha = 0$. But a rigorous derivation requires more work.

3.4 Nearest Neighbor Distributions

Denote by $d\omega_x(r)$ the Lebesgue-Stieltjes measure of $\omega_x(r)$ (see e.g. [56]). Moreover, we define

$$S_{x,k}(t) = \{(x_1, \ldots, x_k) \in \Re^{n\times k} : \ 0 < \|x_1 - x\| < \ldots < \|x_k - x\| < t\} \tag{3.9}$$

for any fixed $x \in \Re^n$ and $k > 0$. In Chapter 2, we have analyzed nearest neighbor distributions mostly based on geometric arguments. However, it turns out that the nearest neighbor distribution has an expression, which allows an elaborate analysis. The following theorem is of course by no means novel (see [18]).

Theorem 3.5. *Suppose that the points $(X_i)_{i=1}^M$ are i.i.d. with a common density q. Then the distribution of the nearest neighbors of a point $x \in \Re^n$ is characterized by (f is a bounded measurable function on $\Re^{n\times(k+1)}$)*

$$\begin{aligned}&E[f(X_1, X_{N[1,1]}, \ldots, X_{N[1,k]})|X_1 = x]\\ &= k!\binom{M-1}{k}\int_{S_{x,k}(\infty)}(1-\omega_x(\|x_{1,k}-x\|))^{M-k-1} f \prod_{i=1}^{k} q(x_{1,i})dx_{1,1}, \ldots, dx_{1,k}.\end{aligned}$$

The argument of f was dropped for notational compactness. For functions that depend only on the k-th nearest neighbor distance,

$$E[f(d_{1,k})|X_1 = x] = k\binom{M-1}{k}\int_0^\infty (1-\omega_x(r))^{M-k-1}\omega_x(r)^{k-1} f(r)d\omega_x(r).$$

Proof. Choose $(x_{1,1}, \ldots, x_{1,k}) \in S_{x,k}(\infty)$. Then we have

$$\begin{aligned}&P(N[1,1] = 2, \ldots, N[1,k] = k+1|X_1 = x, X_2 = x_{1,1}, \ldots, X_{k+1} = x_{1,k})\\ &\quad = (1-\omega_x(\|x_{1,k}-x\|))^{M-k-1}\end{aligned} \tag{3.10}$$

as this is the probability that no other points lie inside the ball $B(x, \|x_{1,k} - x\|)$. Using Equation (3.10),

$$\begin{aligned}&E[f\prod_{j=1}^{k} I(N[1,j] = j+1)|X_1 = x, X_2 = x_{1,1}, \ldots, X_{k+1} = x_{1,k}]\\ =&P(N[1,1] = 2, \ldots, N[1,k] = k+1|X_1 = x, X_2 = x_{1,1}, \ldots, X_{k+1} = x_{1,k}]\\ &\times f(x, x_{1,1}, \ldots, x_{1,k})\\ =&(1-\omega_x(\|x_{1,k}-x\|))^{M-k-1} f(x, x_{1,1}, \ldots, x_{1,k}).\end{aligned}$$

Because the sample is i.i.d, replacing the set of indices $(2, \ldots, k+1)$ with any other set does not make a difference. Thus, using the tower rule of conditional expectations ($E[\cdot] = E[E[\cdot|\mathcal{F}]]$, see [56]) and the combinatorial fact that the number of different subsets of $\{2,\dots,M\}$ with cardinality k is

$$\binom{M-1}{k},$$

we end up with

$$(k!)^{-1}\binom{M-1}{k}^{-1}E[f|X_1=x]=E[f\prod_{j=1}^{k}I(N[1,j]=j+1)|X_1=x]$$

$$=\int_{S_{x,k}(\infty)}E[f\prod_{j=1}^{k}I(N[1,j]=j+1)|X_1=x,X_2=x_{1,1},\dots,X_{k+1}=x_{1,k}]$$

$$\times\prod_{j=1}^{k}q(x_{1,j})dx_{1,1},\dots,dx_{1,k}$$

$$=\int_{S_{x,k}(\infty)}(1-\omega_x(\|x_{1,k}-x\|))^{M-k-1}f\prod_{i=1}^{k}q(x_{1,i})dx_{1,1},\dots,dx_{1,k}.$$

The second result is obtained by integrating out the first $k-1$ coordinates. □

Under (A3)-(A4), the upper tails of the nearest neighbor distribution approach zero fast.

Lemma 3.1. *If (A3)-(A4) hold, then for* $0<t<1$ *and* $0<k<M$*,*

$$P(d_{1,k}>t|X_1)\le M^k e^{-c(M-k-1)t^n}$$

for a constant $c>0$*, which depends only on* $\mathcal{X}$ *and* q*.*

Proof. Using Equation (2.17), we may estimate

$$P(d_{1,k}>t|X_1)=P(\omega_{X_1}(d_{1,k})>\omega_{X_1}(t)|X_1)\le M^k\int_{\omega_{X_1}(t)}^{1}(1-z)^{M-k-1}dz.$$

Now, we apply the fact that $1-z\le e^{-z}$ for $0\le z\le 1$ and the inequality (implied by (A3)-(A4)),

$$\omega_{X_1}(t)\ge c_1\lambda(B(X_1,t)\cap\mathcal{X})\ge c_2t^n$$

for a constant $c_2>0$ which depends only on $\mathcal{X}$. Thus

$$P(d_{1,k}>t|X_1)\le M^k e^{-c_2(M-k-1)t^n}.$$

□

From Lemma 3.1 it follows that for a fixed k and

$$t_M=M^{-1/n}\log^{2/n}M, \tag{3.11}$$

it holds that

$$P(d_{1,k}>M^{-1/n}\log^{2/n}M|X_1)\le M^{k-c\log M}, \tag{3.12}$$

which approaches zero faster than any polynomial w.r.t. M.

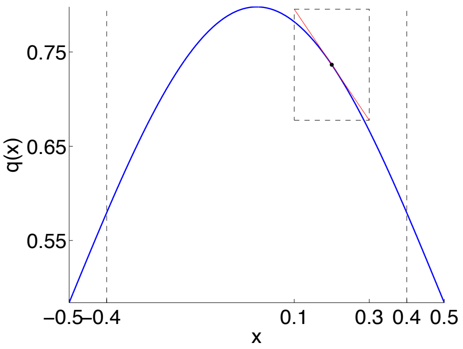


Figure 3.5: In the interior $\mathcal{X} \setminus \partial_{0.1}\mathcal{X}$, q can be expanded inside the ball $B(0.2, 0.1)$, but outside the region bounded by the dashed lines it is no longer the case.

3.5 A General Overview of the Proofs

If we set for example $t_M = M^{-1/n} \log^{2/n} M$, then under (A2)-(A4),

$$E[d_{1,k}^{\alpha}] = E[d_{1,k}^{\alpha} I(d_{1,k} \leq t_M)] + O(M^{-\beta})$$

with $\beta > 0$ any fixed positive number. Thus cutting off nearest neighbor distances at t_M introduces a negligible error term. If $x \in \mathcal{X} \setminus \partial_{t_M}\mathcal{X}$, then under the cut-off the quantity

$$E[d_{1,k}^{\alpha} I(d_{1,k} \leq t_M) | X_1 = x] \tag{3.13}$$

depends only on the values of the density q in the ball $B(x, t_M)$. Because $B(x, t_M) \subset \mathcal{X}$, it is possible to approximate (see Figure 3.5)

$$q(y) = q(x) + (y - x)^T \nabla_x q + O(t_M^2).$$

In Section 3.7 it is shown that because the error in the expansion is of order t_M^2, a local linearization has a small effect on the expectations (3.13). On the other hand, substituting a locally linear density into Theorem 3.5 leads to a simple closed-form expression for the α-moments $E[d_{1,k}^{\alpha}]$.

The linearization argument here is a modification of the analysis in [18] with the difference that in [18] a locally constant approximation of q was used. While the interior $\mathcal{X} \setminus \partial_{t_M}\mathcal{X}$ is relatively easy to handle under sufficient smoothness, difficulties arise when $x \in \partial_{t_M}\mathcal{X}$, because then $B(x, t_M)$ is no longer contained in $\mathcal{X}$ and q cannot be linearized.

As is often the case, to solve the difficulties associated with boundaries, it is useful to consider a simplified case, namely the planar boundary. To do that, we set $\mathcal{X} = [0, 1] \times [-1/2, 1/2]^{n-1}$ and

$$x = (s, 0, \ldots, 0) \qquad (0 < s < t_M). \tag{3.14}$$

Then we ask whether

$$E[d_{1,k}^{\alpha} | X_1 = (s, 0, \ldots, 0)]$$

has some closed-form expression. Unfortunately, this does not seem to be the case because the boundary cut-off introduces a source of non-linearity. To solve the

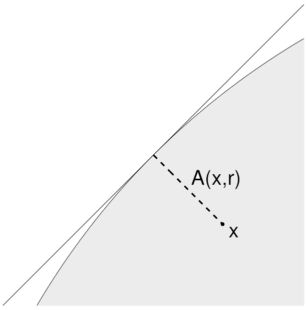


Figure 3.6: The set $A(x, r)$.

problem, the trick is recalling that X_1 does not have to stay fixed. It is shown in Section 3.8.1 that surprisingly

$$\int_0^{t_M} E[d_{1,k}^\alpha I(d_{1,k} \leq t_M)|X_1 = (s, 0, \ldots, 0)]ds \tag{3.15}$$

does have a closed-form expression. In fact, this observation is the most central idea of the chapter; the rest is mostly an application of literature together with additional technicalities.

The main source of technicalities arises from the transition to general $\mathcal{X}$ and q from the uniform distribution. In that process, we would like to end up with integrals of the form (3.15). One approach is to define for each $x \in \partial\mathcal{X} \setminus N$ (the singularities N have to be excluded) and $r > 0$

$$A(x, r) = \{x - sn(x) : \ s \in (0, r]\}, \tag{3.16}$$

see Figure 3.6. $n(x)$ refers here to the outer normal of the set pointing outwards of $\mathcal{X}$; formally the existence of such a normal is shown in Section 3.6.1. The idea is to associate to each $A(x, r)$ a line in the unit cube through a linearization argument; after that q is linearized as well and the problem can be reduced to the uniform unit cube case.

Does every point in $\partial_{t_M}\mathcal{X}$ belong to some $A(x, t_M)$ if (A2) is valid and M large? When points close to the set of singularities N are excluded, an affirmative answer is given. In fact, it is shown in Section 3.8.2 that under (A2)-(A4),

$$\begin{aligned}&\int_{\partial_{t_M}\mathcal{X}} E[d_{1,k}^\alpha I(d_{1,k} \leq t_M)|X_1 = x]q(x)dx \\ &\quad = \int_{\mathcal{X}\cap[\cup_{y\in\partial\mathcal{X}\setminus N} A(y,t_M)]} E[d_{1,k}^\alpha I(d_{1,k} \leq t_M)|X_1 = x]q(x)dx + O(t_M^{2+\alpha}),\end{aligned} \tag{3.17}$$

when $q \in H(c, \gamma)$ with $\gamma \geq 1$. Each point in the set $\mathcal{X} \cap [\cup_{y\in\partial\mathcal{X}\setminus N} A(y, t_M)]$ can be written as

$$y = x - rn(x)$$

for some $r > 0$ suggesting a natural parametrization. In Section 3.6 it is shown that

$$\begin{aligned}
&\int_{\mathcal{X}\cap[\cup_{y\in\partial\mathcal{X}\setminus N}A(y,t_M)]} E[d_{1,k}^{\alpha}I(d_{1,k}\le t_M)|X_1=x]q(x)dx \\
&\quad = \int_{\partial\mathcal{X}\setminus N}\int_0^{t_M} E[d_{1,k}^{\alpha}I(d_{1,k}\le t_M)|X_1=x-rn(x)]q(x-rn(x))drdS \\
&\qquad + O(t_M^{2+\alpha}).
\end{aligned}$$

To simplify further, we utilize

$$\begin{aligned}
&\int_0^{t_M} E[d_{1,k}^{\alpha}I(d_{1,k}\le t_M)|X_1=x-rn(x)]q(x-rn(x))dr \\
&\qquad \approx q(x)\int_0^{t_M} E[d_{1,k}^{\alpha}I(d_{1,k}\le t_M)|X_1=x-rn(x)]dr \qquad (3.18)
\end{aligned}$$

as a good approximation under (A4). If we now compare to the uniform case of Equation (3.15), we see that the integral in (3.18) has the same form except that q cannot be said to be locally uniform and the boundary is not planar. However, it is possible to use

$$q(y) = q(x) + O(t_M) \qquad \text{when } y \in B(x, 2t_M)\cap\mathcal{X}$$

to show that for each $x \in \partial\mathcal{X}$, the points $(X_i)_{i=1}^M$ can be thought to be approximately uniformly distributed inside $B(x, 2t_M)\cap\mathcal{X}$. In addition, as shown in Figure 3.6, the boundary $\partial\mathcal{X}$ can be linearized around x. Together with a linearization of q and $\partial\mathcal{X}$, it is then shown in Section 3.8.2 that (3.18) reduces to the case of uniform points in the unit cube. Of the two arguments, linearization of the boundary is more difficult to handle formally and the proofs in Section 3.6 take more effort.

The chapter proceeds now to the geometric proofs in Section 3.6. In Section 3.7 we analyze the interior $\mathcal{X}\setminus\partial_{t_M}\mathcal{X}$ and after that in Section 3.8 the boundary effect is analyzed first for the unit cube and then in the general case.

3.6 Geometry of the Boundaries

As discussed in Section 3.5, the purpose is using the sets

$$A(x,r) = \{x - sn(x):\ s\in(0,r]\} \qquad (3.19)$$

to reparametrize $\partial_r\mathcal{X}$, but before that it is necessary to fix the direction of the normal $n(x)$. As mentioned earlier, we would like $n(x)$ to point away from $\mathcal{X}$. Then intuitively one would suppose that $A(x,r)\subset\mathcal{X}$, while $A(x,-r)\subset\mathcal{X}^C$ for any small positive r. But rigorously it is not evident that the direction can be fixed this way and considerable effort is taken in Section 3.6.1 to establish the orientation when x is not too close to N.

Despite the difficulties, the orientation can be fixed and the following reparametrization result is proven in Section 3.6.2:

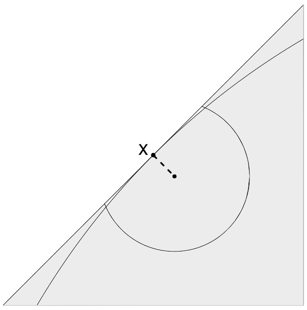


Figure 3.7: The set $\mathcal{X}$ (inside the circular arc) can be locally approximated by $\mathcal{U}_x$ (grey area).

Lemma 3.2. *Suppose that (A2) holds. For any function $f : \Re^n \to \Re$ with $0 \leq f \leq 1$ and a constant $c_1 > 0$ large enough, we have for $0 < r < 1$*

$$\int_{\partial_r \mathcal{X}} f(x)dx = \int_{\partial \mathcal{X} \setminus N_{c_1 r}} \int_0^r f(x - \tilde{r}n(x))d\tilde{r}dS + O(r^2), \tag{3.20}$$

where the outer integral is the surface integral over $\partial \mathcal{X}$. The remainder term $O(r^2)$ can be bounded by $c_2 r^2$ with the constant c_2 depending only on $\mathcal{X}$ and c_1, but not on f.

Lemma 3.2 will be applied to the function

$$f(x) = E[d_{1,k}^{\alpha} I(d_{1,k} \leq t_M)|X_1 = x]q(x).$$

Choose two small numbers $r_1, r_2 > 0$, $x \in \partial \mathcal{X}$ and any $y \in A(x, r_1)$. Then if $\partial \mathcal{X}$ is locally smooth, it is reasonable to assume that the boundary can be well approximated with a plane inside the ball $B(y, r_2)$. As a consequence, under our convention that $n(x)$ points outward, it would seem intuitive that inside $B(y, r_2)$, $\mathcal{X}$ is very similar to the half-plane $\mathcal{U}_x$ of Equation (3.3) as demonstrated in Figure 3.7. A formalization of this idea in Section 3.6.1 in terms of volumes is the second main result of this section:

Lemma 3.3. *Assume that (A2) holds. Then there exists constants $c_1, c_2 > 0$ (depending only on $\mathcal{X}$) such that for any $0 < \delta < 1$, $0 < r_1, r_2 < c_1 \delta$ and $x \in \partial \mathcal{X} \setminus N_\delta$, if we fix $y = x - r_1 n(x)$ and define the sets $\Xi_1 = B(y, r_2) \cap \mathcal{X}$ and $\Xi_2 = B(y, r_2) \cap \mathcal{U}_x$, then we have*

$$\lambda(\Xi_1 \setminus \Xi_2) + \lambda(\Xi_2 \setminus \Xi_1) \leq c_2(r_1^{n+1} + r_2^{n+1}). \tag{3.21}$$

Later on, it turns out that it is exactly the volumes of the intersections that matter when simplifying the general case to the uniform distribution.

3.6.1 Linearization

The first step in establishing the main results in Lemmas 3.2 and 3.3 is showing that for some constant $c > 0$, $n(x)$ can be fixed to ensure that $A(x, r) \subset \mathcal{X}$ when

$r > 0$ is small and $x \in \partial\mathcal{X} \setminus N_{cr}$. We proceed through four intermediate steps, which provide useful technical results. The following lemma is an application of the derivative of inverse functions.

Lemma 3.4. *Suppose that $\mathcal{D} \subset \Re^n$ is an (n-1)-dimensional twice continuously differentiable submanifold and choose any $x \in \mathcal{D}$. Then there exists a local parametrization $\phi : U \to B(x,\epsilon) \cap \mathcal{D}$ and a constant $c_x > 0$ such that if $(x_i)_{i=1}^{\infty} \subset \mathcal{D}$ is a sequence converging to x, then it is possible to choose an integer $i_0 > 0$ with*

$$\|\phi^{-1}(x_i) - \phi^{-1}(x)\| \leq c_x \|x_i - x\|$$

for $i > i_0$.

Proof. Choose a twice continuously differentiable parametrization $\phi : U \to B(x,\epsilon) \cap \mathcal{D}$ with $\phi(0) = x$. Because U is an open set, there exists $\delta_1 > 0$ such that the closure $\overline{B(0,\delta_1)}$ is a subset of U. Notice that

$$(J_y\phi)^T J_y\phi$$

is a continuous matrix valued function with eigenvalues strictly above zero for each fixed $y \in U$. By continuity of the Jacobian, this implies that there exists a constant $c > 0$ such that

$$\inf_{y \in \overline{B(0,\delta_1)}, \|z\|=1} \|(J_y\phi)z\| \geq c. \tag{3.22}$$

Because ϕ is a homeomorphism, $\phi(B(0,\delta_1))$ contains the set $B(x,\delta_2) \cap \mathcal{D}$ for some $\delta_2 > 0$ and consequently $\phi^{-1}(x_i) \in B(0,\delta_1)$ for some i_0 and all $i > i_0$. Thus by Equation (3.22) and the mean value theorem, for some $\xi \in B(0,\delta_1)$ it holds that

$$\begin{aligned}\|x_i - x\| &= \|\phi(\phi^{-1}(x_i)) - \phi(\phi^{-1}(x))\| \\ &= \|J_\xi\phi(\phi^{-1}(x_i) - \phi^{-1}(x))\| \geq c\|\phi^{-1}(x_i) - \phi^{-1}(x)\|\end{aligned}$$

finishing the proof if we take $c = c_x^{-1}$. □

Consider a local parametrization $\phi : U \to B(x,\epsilon) \cap \mathcal{D}$ of a continuously differentiable submanifold $\mathcal{D}$. The tangent plane at the point $x = \phi(u)$ is determined by the span of the Jacobian $J_u\phi$. If $\tilde{x}$ is close to x, then by Lemma 3.4 $\tilde{u} = \phi^{-1}(\tilde{x})$ is also close to u and

$$\begin{aligned}\tilde{x} &= x + J_u\phi(\tilde{u} - u) + O(\|\tilde{u} - u\|^2) \\ &= x + J_u\phi(\tilde{u} - u) + O(\|\tilde{x} - x\|^2) \end{aligned} \tag{3.23}$$

implying that the component of $\tilde{x} - x$ orthogonal to the plane G_x has a length of order $\|\tilde{x} - x\|^2$, because the sum of the first two terms in the right side of (3.23) belongs to T_x. But the length of the projection is in fact given by $|n(x)^T(\tilde{x} - x)|$, showing that

$$n(x)^T(\tilde{x} - x) = n(x)^T J_u\phi(\tilde{u} - u) + O(\|\tilde{x} - x\|^2) = O(\|\tilde{x} - x\|^2).$$

While the logic works around a fixed point x, we would like to extend it to hold for any pair $(x, \tilde{x}) \in \mathcal{D} \times \mathcal{D}$. If $\mathcal{D}$ is assumed to be a compact set, then a proof to this direction is possible.

Lemma 3.5. *Suppose that $\mathcal{D}$ is a twice continuously differentiable (n-1)-dimensional compact submanifold. Then for any sequence $(x_{1,i}, x_{2,i})_{i=1}^{\infty} \subset \mathcal{D} \times \mathcal{D}$ with $\|x_{1,i} - x_{2,i}\| \to 0$ and $x_{1,i} \neq x_{2,i}$,*

$$\sup_{i>0} \frac{|n(x_{1,i})^T(x_{1,i} - x_{2,i})|}{\|x_{1,i} - x_{2,i}\|^2} < \infty. \tag{3.24}$$

Proof. Let us make the counterassumption, that Equation (3.24) goes to infinity for the sequence $(x_{1,i}, x_{2,i})_{i=1}^{\infty}$. By compactness, without losing generality it may be assumed that

$$(x_{1,i}, x_{2,i}) \to (x, x)$$

for some $x \in \mathcal{D}$. Choose $\phi : U \to \mathcal{D} \cap B(x, \epsilon)$ as a local parametrization around x and set (for large i)

$$u_{j,i} = \phi^{-1}(x_{j,i}) \qquad (j = 1, 2).$$

By Lemma 3.4,

$$u_{1,i}, u_{2,i} \to \phi^{-1}(x)$$

and using the fact that $(J_{u_{1,i}}\phi)^T n(x_{1,i}) = 0$, we have

$$\begin{aligned} \frac{n(x_{1,i})^T(x_{1,i} - x_{2,i})}{\|x_{1,i} - x_{2,i}\|^2} &= n(x_{1,i})^T \frac{J_{u_{1,i}}\phi(u_{1,i} - u_{2,i}) + O(\|u_{1,i} - u_{2,i}\|^2)}{\|J_{\phi^{-1}(x)}\phi(u_{1,i} - u_{2,i}) + o(\|u_{1,i} - u_{2,i}\|)\|^2} \\ &= O(1) \end{aligned}$$

leading to a contradiction. □

The following lemma is invoked twice in the forthcoming analysis. It shows that if a ball $B(x, r)$ is not too close to N with $x \in \mathcal{C}_1$ (recall (A2) for the notation), then $B(x, r) \cap \mathcal{D}_1 = B(x, r) \cap \mathcal{C}_1$. We already know by condition 4 in (A2) that for any small $r > 0$ and fixed x this happens, but it is of interest to show that the threshold under which the inclusion is valid does not get arbitrarily small.

Lemma 3.6. *Suppose that (A2) holds and choose one of the submanifolds $\{\mathcal{C}_i\}_{i=1}^{l}$. Then there exists a constant $c > 0$ such for all $0 < r < \delta/5 < c$ and $x \in \mathcal{C}_i \setminus N_\delta$*

$$B(x, r) \cap \mathcal{D}_i = B(x, r) \cap \mathcal{C}_i.$$

Proof. It seems easiest to proceed by a counterassumption: assume that for some sequence $(x_i, y_i, r_i, \delta_i)_{i=1}^{\infty}$ it holds that

1. For each $i > 0$, $x_i \in \mathcal{C}_1 \setminus N_{\delta_i}$ and $y_i \in \mathcal{D}_1 \setminus \mathcal{C}_1$.

2. $\delta_i \to 0$ in the limit $i \to \infty$ and $0 < r_i < \delta_i/5$.

3. $\|x_i - y_i\| < r_i$.

The contradiction is established by examining a path between x_i and y_i to exploit the fact that such a path contains a point in N by (A2). To construct an appropriate path between the two points, a local parametrization on $\mathcal{D}_1$ is used.

By compactness of $\mathcal{D}_1$, it may be assumed that $x_i \to x$ for some $x \in \mathcal{D}_1$ and a local parametrization $\phi : U \to B(x, \epsilon) \cap \mathcal{D}_1$ can be found. When i is large, both x_i and y_i are in the range of ϕ and it makes sense to define $u_i = \phi^{-1}(x_i)$ and $\Delta u_i = \phi^{-1}(x_i) - \phi^{-1}(y_i)$. Consider

$$t_i = \sup\{0 \le t \le 1 : \ \phi(u_i + t\Delta u_i) \in \mathcal{C}_1\}$$

and $z_i = \phi(u_i + t_i \Delta u_i)$. $z_i \notin \mathcal{C}_1$ by condition 4 in (A2), but also $z_i \in \mathcal{D}_1 \cap \partial\mathcal{X}$ by closedness. Consequently, by condition 3, $z_i \in N$ and it remains to show that $\|x_i - z_i\| < 4r_i$. Lemma 3.4 indicates that $(u_i, \Delta u_i) \to (\phi^{-1}(x), 0) = (u, 0)$ with $u = \phi^{-1}(x)$ and by a Taylor expansion

$$\begin{aligned} r_i > \|x_i - y_i\| &= \|(J_{u_i}\phi)\Delta u_i + O(\|\Delta u_i\|^2)\| = \|(J_u\phi)\Delta u_i + o(\|\Delta u_i\|)\| \\ &\ge \frac{1}{2}\|(J_u\phi)\Delta u_i\| \end{aligned}$$

when $i > i_0$ for some threshold $i_0 > 0$. Consequently also

$$\begin{aligned} \|x_i - z_i\| &= \|t_i(J_u\phi)\Delta u_i + o(t_i\|\Delta u_i\|)\| \\ &\le 2\|(J_u\phi)\Delta u_i\| \le 4\|x_i - y_i\| < 4r_i. \end{aligned} \tag{3.25}$$

A contradiction with the counterassumption is established, because $z_i \in N$, while $x_i \in \partial\mathcal{X} \setminus N_{5r_i}$. □

Consider two points on different faces, say $x \in \mathcal{C}_1$ and $y \in \mathcal{C}_2$. For $\|x - y\|$ to be small, it would have to be that the compact submanifolds $\mathcal{D}_1$ and $\mathcal{D}_2$ are close to each other as well. However, it is stated in (A2) that $\mathcal{D}_1$ and $\mathcal{D}_2$ intersect inside $\partial\mathcal{X}$ only in the set N, where they have tangent planes non-parallel to each other. For this reason, it is expected that x and y cannot be too close to each other unless they are close to N.

The heuristic discussion is demonstrated in Figure 3.8, where the point z is a point of non-smoothness. The lower bound $\|x - y\| \ge \min\{\|x - z\|, \|y - z\|\}$ shows that x and y can be made arbitrarily close to each other only if their distance to z approaches zero.

Lemma 3.7. *Suppose that (A2) holds and choose integers $i \neq j$. Then there exists a constant $c > 0$ depending only on $\mathcal{X}$ such that for all $0 < \delta < 1$, $x \in \mathcal{C}_i \setminus N_\delta$ and $y \in \mathcal{C}_j \setminus N_\delta$,*

$$\|x - y\| \ge c\delta.$$

Proof. The proof is divided into three parts, of which the second one is the longest.

1. Counterassumption

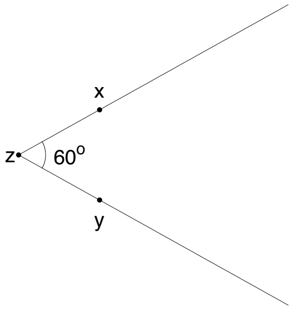


Figure 3.8: The distance between x and y is bounded from below by $\|x - y\| \geq \min\{\|x - z\|, \|y - z\|\}$.

The counterassumption states that there exists a sequence $(x_i, y_i)_{i=1}^{\infty}$ and a strictly increasing sequence of integers $(j_i)_{i=1}^{\infty}$ such that

$$j_i \|x_i - y_i\| \to 0$$

when $i \to \infty$, while $x_i \in \mathcal{C}_1 \setminus N_{j_i^{-1}}$ and $y_i \in \mathcal{C}_2 \setminus N_{j_i^{-1}}$ ($\mathcal{C}_1$ and $\mathcal{C}_2$ may be fixed without compromising the generality of the proof). By compactness, we may assume (by taking an appropriate subsequence) that $(x_i, y_i) \to (x, x)$ for some $x \in \mathcal{D}_1 \cap \mathcal{D}_2 \cap \partial\mathcal{X}$ and consequently by (A2) $x \in N$. Without losing generality, the choice $x = 0$ is made.

2. For some strictly increasing sequence of integers $(i_k)_{k=1}^{\infty}$, $\rho(x_{i_k}, \mathcal{D}_1 \cap \mathcal{D}_2) = O(k^{-1} j_{i_k}^{-1})$.
Let us choose two local parametrizations, $\phi_1 : U_1 \to \mathcal{D}_1 \cap B(0, \epsilon)$ and $\phi_2 : U_2 \to \mathcal{D}_2 \cap B(0, \epsilon)$ with $\phi_1(0) = \phi_2(0) = 0$. By invoking (A2) for large i, there exists a vector $g \in \Re^{n-1}$ of unit norm such that

$$(J_0 \phi_2) g \notin G_0^{\mathcal{D}_1} \tag{3.26}$$

with $G_0^{\mathcal{D}_1}$ the tangent plane of $\mathcal{D}_1$ at 0. Set $u_i = \phi_1^{-1}(x_i)$, $v_i = \phi_2^{-1}(y_i)$ and define the functions f_i around $(0, 0) \in \Re^{n-1} \times \Re$ by

$$f_i(\tilde{u}, \alpha) = \phi_1(u_i + \tilde{u}) - \phi_2(v_i + \alpha g).$$

In the following, it is shown that for any positive number $c > 0$, f_i has a zero point in the ball $B(0, c j_i^{-1}) \subset \Re^n$ for any large i (clearly f_i is defined in the ball once i is large). To understand why the zero points are useful, observe that if $f_i(\tilde{u}_i, \alpha_i) = 0$ for some $(\tilde{u}_i, \alpha_i) \in B(0, c j_i^{-1})$, then $\phi_1(u_i + \tilde{u}_i) = \phi_2(v_i + \alpha_i g)$ and consequently $\phi_1(u_i + \tilde{u}_i) \in \mathcal{D}_1 \cap \mathcal{D}_2$. Moreover, because c can be arbitrary, there exists an increasing sequence of integers $(i_k)_{k=1}^{\infty}$ such that $(\tilde{u}_{i_k}, \alpha_{i_k}) \in B(0, k^{-1} j_{i_k}^{-1})$ with $f(\tilde{u}_{i_k}, \alpha_{i_k}) = 0$ for $k > 0$ showing that $\rho(x_{i_k}, \mathcal{D}_1 \cap \mathcal{D}_2) = O(k^{-1} j_{i_k}^{-1})$.

Fix any $c > 0$. In order to show that eventually f_i has a zero point in $B(0, c j_i^{-1})$, observe that

$$J_{(0,0)} f_i = [J_{u_i} \phi_1, -(J_{v_i} \phi_2) g] = [J_0 \phi_1, -(J_0 \phi_2) g] + O(\|v_i\| + \|u_i\|); \tag{3.27}$$

this holds because the parametrizations are twice continuously differentiable. One notices that both matrices $(J_{0,0}f_i)(J_{0,0}f_i)^T$ and $(J_{0,0}f_i)^T(J_{0,0}f_i)$ have eigenvalues bounded from below by a constant $c_f^2 > 0$ independent of i and the choice of c assuming that $i \geq i_0$, where the threshold i_0 depends on c. In fact, $[J_0\phi_1, -(J_0\phi_2)g]$ is independent of i and non-singular by Equations (3.26) and (3.27) implying that the non-negative eigenvalues of the positive definite matrices $(J_{0,0}f_i)^T(J_{0,0}f_i)$ (and $(J_{0,0}f_i)(J_{0,0}f_i)^T$) can be bounded from below once the remainder term in (3.27) is made small enough in the limit $i \to \infty$.

Now, let $(\tilde{u}_{0,i}, \alpha_{0,i})$ be the point that minimizes $\|f_i\|^2$ on $\overline{B(0, \frac{1}{2}cj_i^{-1})}$ (such a point exists even if not necessarily unique). We assume first that $\|f_i(\tilde{u}_{0,i}, \alpha_{0,i})\| > 0$. With the definition ($\tilde{u}_{0,i}$ is understood as a row vector so a transpose is taken)

$$h_i = \begin{pmatrix} \tilde{u}_{0,i}^T \\ \alpha_{0,i} \end{pmatrix}$$

we have for large i,

$$\|f_i(\tilde{u}_{0,i}, \alpha_{0,i})\| = \|(J_{0,0}f_i)h_i + f_i(0,0) + O(\|\tilde{u}_{0,i}\|^2 + \alpha_{0,i}^2)\| = \|(J_{0,0}f_i)h_i + o(j_i^{-1})\|, \tag{3.28}$$

because

$$f_i(0,0) = \|\phi_1(u_i) - \phi_2(v_i)\| = \|x_i - y_i\| = o(j_i^{-1})$$

by the counterassumption in the first step of the proof. The remainder depends on the choice of c, but more importantly it goes to zero faster than j_i^{-1}.

Moreover, the minimizing point must be at the boundary of $B(0, \frac{1}{2}cj_i^{-1})$, as the gradient of $\|f_i\|^2$ is

$$\begin{aligned} 2f_i(\tilde{u}_{0,i}, \alpha_{0,i})^T J_{\tilde{u}_{0,i},\alpha_{0,i}} f_i &= 2f_i(\tilde{u}_{0,i}, \alpha_{0,i})^T J_{0,0} f_i \\ &\quad + 2f_i(\tilde{u}_{0,i}, \alpha_{0,i})^T (J_{\tilde{u}_{0,i},\alpha_{0,i}} f_i - J_{0,0}f_i) \\ &= 2f_i(\tilde{u}_{0,i}, \alpha_{0,i})^T J_{0,0} f_i + O(j_i^{-1}\|f_i(\tilde{u}_{0,i}, \alpha_{0,i})\|), \end{aligned}$$

which cannot be zero if $f_i(\tilde{u}_{0,i}, \alpha_{0,i})$ is non-zero and i large, because

$$\begin{aligned} \|f_i(\tilde{u}_{0,i}, \alpha_{0,i})^T J_{0,0} f_i\|^2 &= f_i(\tilde{u}_{0,i}, \alpha_{0,i})^T (J_{0,0}f_i)(J_{0,0}f_i)^T f_i(\tilde{u}_{0,i}, \alpha_{0,i}) \\ &\geq c_f^2 \|f_i(\tilde{u}_{0,i}, \alpha_{0,i})\|^2 \end{aligned}$$

and thus no local minimum of f_i may exist in $B(0, \frac{1}{2}cj_i^{-1})$. On the other hand, if h_i is at the boundary, then

$$\|(J_{0,0}f_i)h_i\| \geq \frac{1}{2}c_f c j_i^{-1}$$

and by Equation (3.28), $\|f_i(\tilde{u}_{0,i}, \alpha_{0,i})\| \geq \frac{1}{4}c_f c j_i^{-1}$ (say) for i large enough. Consequently the minimizing point cannot be at the boundary either and it must be that $\|f_i(\tilde{u}_{0,i}, \alpha_{0,i})\| = 0$ once i passes some threshold (which depends on c).

3. Contradiction:
We have proven in step 2 that there exists a sequence $(z_k)_{k=1}^{\infty} \subset \mathcal{D}_1 \cap \mathcal{D}_2$ such that $\|z_i - x_{i_k}\| = o(j_{i_k}^{-1}k^{-1})$. But because $x_{i_k} \in \partial\mathcal{X} \setminus N_{cj_{i_k}^{-1}}$, this contradicts Lemma 3.6 as by that result, $B(x_{i_k}, j_{i_k}^{-1}/5) \cap \mathcal{D}_1 \subset \mathcal{C}_1$ for large k whereas $z_k \in \mathcal{D}_1 \setminus \mathcal{C}_1$. □

Recall that there is two possible orientations for the normal $n(x)$ at the boundary $\partial\mathcal{X}$. Now the orientation problem can be solved. The following result is not only useful in this regard, but it also states that the sets $A(x, r)$ in Equation (3.19) tend to be disjoint for different points x.

Lemma 3.8. *Suppose that (A2) holds. Then, there exists a constant* $0 < c < 1$ *(depending only on the set* $\mathcal{X}$*) such that for all* $0 < \delta < 1$*,*

$$A(x, r) \cap A(y, r) = \emptyset$$

when $y \neq x$*,* $x, y \in \partial\mathcal{X} \setminus N_\delta$ *and* $|r| < c\delta$*. Moreover, when* $|r| < c\delta$*, the orientation of* $n(x)$ *can be chosen in such a way that* $A(x, r) \subset \mathcal{X}$ *and* $A(x, -r) \subset \bar{\mathcal{X}}^C$*.*

Proof. Lemma 3.7 ensures that we may always choose $0 < c < 1$ in such a way that if $x \in \mathcal{C}_i \setminus N_\delta$ and $y \in \mathcal{C}_j \setminus N_\delta$ with $i \neq j$, then $A(x, r) \cap A(y, r) = \emptyset$ for all $0 < r < c\delta$ and $0 < \delta < 1$. Thus, we may restrict ourselves to the case $x, y \in \mathcal{C}_1$.

Let us make the counterassumption that there exist sequences $(x_i, y_i)_{i=1}^\infty \subset \mathcal{C}_1 \times \mathcal{C}_1$ and $(r_{1,i}, r_{2,i})_{i=1}^\infty \to (0, 0)$ such that

$$x_i - y_i = r_{1,i} n(x_i) - r_{2,i} n(y_i)$$

and $x_i \neq y_i$. Then we would have

$$1 = \frac{r_{1,i} n(x_i)^T (x_i - y_i)}{\|x_i - y_i\|^2} - \frac{r_{2,i} n(y_i)^T (x_i - y_i)}{\|x_i - y_i\|^2}$$

leading to a contradiction, because by Lemma 3.5 the right side should go to zero. Thus the first part of the proof is complete.

We know also that if $x \in \partial\mathcal{X} \setminus N_r$, then $A(x, r)$ must either be a subset of $\mathcal{X}$ or its complement $\bar{\mathcal{X}}^C$ because otherwise it would contain points from $\partial\mathcal{X} \setminus N$. To see that this would be contradictory, one should observe that the first part of the proof holds for the sets $A(x, r) \cup \{x\}$ as well.

Let us make the counterassumption that

$$A(x, r) \cup A(x, -r) \subset \bar{\mathcal{X}}^C$$

for some $x \in \mathcal{C}_1 \setminus N_\delta$ and $0 < r < c\delta$.

As a vague heuristic idea, such a situation could occur only if $\mathcal{X}$ somehow resembles the set $\mathcal{X}$ in Figure 3.9 at the cross-point of four faces, which of course does not make sense if $n(x)$ is assumed to be well-defined. In order to implement this consideration, choose any $0 < |t| < |r|$, define the pair of points $(z_{1,t}, z_{2,t})$ by

$$z_{l,t} = x + (-1)^l t n(x) \qquad (l = 1, 2)$$

and choose $\epsilon > 0$ in such a way that $B(z_{l,t}, \epsilon) \subset \bar{\mathcal{X}}^C$ (this is possible because $\bar{\mathcal{X}}^C$ is an open set). By (A3) there exists a sequence $(x_i)_{i=1}^\infty \subset \mathcal{X}$ approaching x (when $i \to \infty$). Then for any $t > 0$, there exists an integer i_t such that the set

$$\{x_{i_t} - s n(x) : s \in [0, t]\} \cup \{x_{i_t} + s n(x) : s \in [0, t]\} \tag{3.29}$$

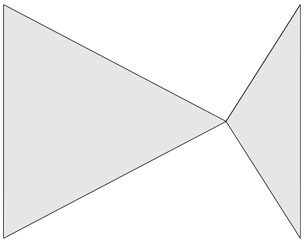


Figure 3.9: $n(x)$ is not defined in any of the corner points including the cross-point of the four faces.

contains two distinct points $(\tilde{z}_{1,t}, \tilde{z}_{2,t})$ on $\mathcal{C}_1$ approaching (x, x) when $t \to 0$. This follows from the fact that for i_t large enough, the set (3.29) intersects both $B(z_{1,t}, \epsilon)$ and $B(z_{2,t}, \epsilon)$ with $x_{i_t} \in \mathcal{X}$ thus containing points from $\mathcal{X}$ and $\bar{\mathcal{X}}^C$. Consequently, (3.29) must also include points in $\partial\mathcal{X}$ and with the choice of c at the beginning, such points are in $\mathcal{C}_1$.

We have arrived into a contradiction with Lemma 3.5 because by definition

$$|\frac{n(x) \cdot (\tilde{z}_{1,t} - \tilde{z}_{2,t})}{\|\tilde{z}_{1,t} - \tilde{z}_{2,t}\|^2}| = \frac{1}{\|\tilde{z}_{1,t} - \tilde{z}_{2,t}\|}$$

and

$$|\frac{(n(x) - n(\tilde{z}_{1,t})) \cdot (\tilde{z}_{1,t} - \tilde{z}_{2,t})}{\|\tilde{z}_{1,t} - \tilde{z}_{2,t}\|^2}| \leq \|n(x) - n(\tilde{z}_{1,t})\| \frac{1}{\|\tilde{z}_{1,t} - \tilde{z}_{2,t}\|}$$

the latter being asymptotically neglible, because we may choose the normals in such a way that

$$\|n(x) - n(\tilde{z}_{1,t})\| \to 0$$

as $t \to 0$.

To finish, we must examine the opposite case

$$A(x, r) \cup A(x, -r) \subset \mathcal{X}.$$

In this case, we may find a sequence $(x_i)_{i=1}^{\infty} \subset \bar{\mathcal{X}}^C$ approaching x. As in the previous step, it can be seen that again

$$\{x_i - sn(x) : s \in [0, t]\} \cup \{x_i + sn(x) : s \in [0, t]\}$$

contains at least two distinct points from $\partial\mathcal{X}$ for arbitrarily small $t > 0$ when i is large enough. Analogously to the previous case, this leads to a contradiction. □

From now on, the outer direction of $n(x)$ that ensures $A(x, -r) \subset \bar{\mathcal{X}}^C$ is always chosen. Next the proof of the linearization argument of Lemma 3.3 is given.

Proof. (Proof of Lemma 3.3)
Let us make the counterassumption that for any $0 < c_1 < 1$, there exists a sequence

$$(x_i, y_i, r_{1,i}, r_{2,i}, \delta_i)_{i=1}^{\infty}$$

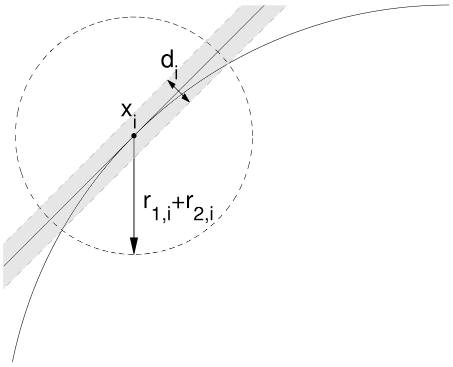


Figure 3.10: The set G_i (grey region).

with $r_{1,i}, r_{2,i} \to 0$, $0 < r_{1,i}, r_{2,i} < c_1\delta_i < 1$, $x_i \in \partial\mathcal{X} \setminus N_{\delta_i}$ and $y_i = x_i - r_{1,i}n(x_i)$ such that the left side of inequality (3.21) exceeds $c_2 r_{1,i}^{n+1} + c_2 r_{2,i}^{n+1}$ for any $c_2 > 0$ when i is large enough. We may assume that $x_i \in \mathcal{D}_1$ for $i > 0$ and $x_i \to x$ with $x \in \mathcal{D}_1$ by choosing an appropriate subsequence (fixing $\mathcal{D}_1$ is only a notational convention). Then there exists a local parametrization $\phi : U \to B(x, \epsilon) \cap \mathcal{D}_1$.

For each $i > 0$, choose an arbitrary point $z_i \in B(x_i, r_{1,i} + r_{2,i}) \cap \mathcal{D}_1$. In the limit $i \to \infty$, the formula for the derivative of inverse functions yields

$$\phi^{-1}(z_i) - \phi^{-1}(x_i) = (J_{\phi^{-1}(x_i)}\phi)^{-1}(z_i - x_i) + O(\|z_i - x_i\|^2) = O(\|z_i - x_i\|), \quad (3.30)$$

where we used $(J_{\phi^{-1}(x_i)}\phi)^{-1} \to (J_{\phi^{-1}(x)}\phi)^{-1}$ in the limit $i \to \infty$. When i is large enough, the fact that $B(x_i, r_{1,i} + r_{2,i}) \subset B(x, \epsilon)$, a Taylor expansion and (3.30) yield

$$z_i = x_i + J_{\phi^{-1}(x_i)}\phi(\phi^{-1}(z_i) - \phi^{-1}(x_i)) + O(\|z_i - x_i\|^2).$$

The first sum in the right side is a point on the tangent plane T_{x_i}; thus the distance between z_i and the set T_{x_i} is

$$\rho(z_i, T_{x_i}) = O(\|z_i - x_i\|^2) \quad (3.31)$$

in the sense that there exists a point in T_{x_i} whose distance to z_i is of magnitude $O(\|z_i - x_i\|^2)$. Set $d_i = \sup_{z \in \partial\mathcal{X} \cap B(x_i, r_{1,i} + r_{2,i})} \rho(z, T_{x_i})$. Because each z_i is arbitrary in the ball $B(x_i, r_{1,i} + r_{2,i})$, we may say that when c_1 is small in order for Lemma 3.7 to hold,

$$d_i = \sup_{z \in \mathcal{C}_1 \cap B(x_i, r_{1,i} + r_{2,i})} \rho(z, T_{x_i}) \le \sup_{z \in \mathcal{D}_1 \cap B(x_i, r_{1,i} + r_{2,i})} \rho(z, T_{x_i}) = O(r_{1,i}^2 + r_{2,i}^2). \quad (3.32)$$

Define the sets (the sum of a vector and a set being defined in the standard way)

$$G_i = \cup_{-d_i \le r \le d_i} [T_{x_i} + rn(x_i)].$$

G_i is demonstrated in Figure 3.10. Then,

$$\partial\mathcal{X} \cap B(y_i, r_{2,i}) \subset G_i \cap B(x_i, r_{1,i} + r_{2,i}) \quad (3.33)$$

and Equation (3.32) implies that

$$\lambda(G_i \cap B(x_i, r_{1,i} + r_{2,i})) = O(r_{1,i}^{n+1} + r_{2,i}^{n+1}). \quad (3.34)$$

We may divide $B(y_i, r_{2,i}) \setminus G_i$ into the sets

$$A_1 = (B(y_i, r_{2,i}) \setminus G_i) \cap \mathcal{U}_{x_i}$$

and

$$A_2 = (B(y_i, r_{2,i}) \setminus G_i) \cap \mathcal{U}_{x_i}^C,$$

both of which are open and convex. Lemma 3.8 implies that for i large enough (again assuming that c_1 is sufficiently small),

$$y_i - \frac{1}{2} r_{2,i} n(x_i) \in A_1 \cap \mathcal{X}.$$

Thus $A_1 \cap \mathcal{X}$ is non-empty and consequently A_1 must be a subset of $\mathcal{X}$ because it does not contain points from $\partial\mathcal{X}$ (as implied by Equation (3.33)). For by convexity, if $A_1 \cap \mathcal{X}^C \neq \emptyset$, then A_1 would contain a boundary point as well. On the other hand, by Lemma 3.8 and the same argument as before, A_2 is in $\mathcal{X}^C$ when i is large enough. Thus, inevitably

$$A_2 \cap \mathcal{X} = \emptyset.$$

We may conclude that for any large i,

$$\begin{aligned} A_1 &= A_1 \cap \mathcal{X} = (A_1 \cap \mathcal{X}) \cup (A_2 \cap \mathcal{X}) \\ &= ((B(y_i, r_{2,i}) \cap \mathcal{U}_{x_i} \cap \mathcal{X}) \setminus G_i) \cup ((B(y_i, r_{2,i}) \cap \mathcal{U}_{x_i}^C \cap \mathcal{X}) \setminus G_i) \\ &= (B(y_i, r_{2,i}) \cap \mathcal{X}) \setminus G_i \end{aligned}$$

and Equation (3.34) leads to a contradiction with the counterassumption at the beginning finishing the proof. □

3.6.2 The Set $\partial_r\mathcal{X}$

We examine the set of points close to the boundary $\partial\mathcal{X}$ defined in Equation (2.5) rigorously. The idea behind the sets $A(x, r)$ becomes evident once it is proven that each point in $\partial_r\mathcal{X}$ belongs to one of such sets.

Lemma 3.9. *If (A2) holds, then there exists a positive number $c_2 > 0$ such that for all $c_1 > c_2$, there is a constant $c_3 > 0$ such that*

$$\lambda(\partial_r\mathcal{X} \setminus \cup_{x \in \partial\mathcal{X} \setminus N_{c_1 r}} A(x, r)) + \lambda(\cup_{x \in \partial\mathcal{X} \setminus N_{c_1 r}} A(x, r) \setminus \partial_r\mathcal{X}) \leq c_3 r^2$$

for $0 < r < 1$.

Proof. We show that any sequence approaching $\partial\mathcal{X}$ eventually belongs to $\partial_r\mathcal{X}$ as long as the elements of the sequence do not get too close to N. Set $(x_i)_{i=1}^{\infty}$ as a sequence with $d_i = \rho(x_i, \mathcal{C}_1) \to 0$ and $x_i \in \mathcal{X} \setminus N_{c_1 d_i}$ for some constant $c_1 > 17$. By compactness, we may assume that $x_i \to x \in \mathcal{D}_1$. The proof proceeds by an application of the implicit function theorem to show that eventually $x_i \in \cup_{y \in \partial\mathcal{X} \setminus N_{c_1 d_i}} A(y, d_i)$.

Choose a local parametrization $\phi : U \to \mathcal{D}_1 \cap B(x, \epsilon)$ ($\phi(0) = x$) at the point x and define the injective mapping $g : U \times [-\delta_1, \delta_1] \to \Re^n$ (the proof of Lemma 3.8 shows injectivity for small $\delta_1 > 0$) by

$$g(u, r) = \phi(u) - r n(\phi(u)).$$

The normal is understood as that of $\mathcal{D}_1$, because $\phi(u)$ may well be outside $\partial\mathcal{X}$ (the orientation of the normal is not relevant here). If it could be shown that the range of g contains an open set with x in it, then x_i would eventually belong to that open set when i is large and it would hold that

$$x_i = g(u_i, r_i) = \phi(u_i) - r_i n(\phi(u_i))$$

for some pair $(u_i, r_i) \in U \times [-\delta_1, \delta_1]$ leaving us the proofs of $r_i \leq d_i$ and $\phi(u_i) \in \partial\mathcal{X} \setminus N_{\frac{1}{2}c_1 d_i}$. Namely, if the latter two hold, then Lemma 3.8 shows that in the limit $i \to \infty$,

$$x_i \in \cup_{y \in \partial\mathcal{X} \setminus N_{\frac{1}{2}c_1 d_i}} A(y, d_i)$$

when c_1 is large enough.

Possibly the easiest way to proceed is by examining the derivative of g at the origin. Let $(v_i(\phi(u)))_{i=1}^{n-1}$ be an orthonormal basis for the tangent space at the point $\phi(u)$ obtained by Gram-Schmidt orthonormalization of the columns of $J_u\phi$. Then, each $v_i(\phi(u))$ is a continuously differentiable function on U. For u close to 0, we obtain

$$n(\phi(u)) = \frac{n(x) - \sum_{i=1}^{n-1} \langle n(x), v_i(\phi(u)) \rangle \, v_i(\phi(u))}{\|n(x) - \sum_{i=1}^{n-1} \langle n(x), v_i(\phi(u)) \rangle \, v_i(\phi(u))\|}.$$

Clearly $n(\phi(u))$ and consequently g is continuously differentiable around the origin, because the denominator is bounded away from zero when u is close enough to 0. Moreover, the Jacobian of g at $(0,0)$ is

$$J_{(0,0)} g = [J_0\phi, -n(x)],$$

which is non-singular rendering the inverse function theorem valid. By the inverse function theorem,

$$g(B((0,0), \delta_2))$$

is open for any small $\delta_2 > 0$ and it contains x. Consequently, as mentioned before there exists an integer i such that x_i belongs to the range of g.

Let (u_i, t_i) be the pair with $g(u_i, t_i) = x_i$; then the inverse function theorem also implies that $(u_i, t_i) \to 0$. For any $0 < \delta_3 < 1$, we may choose a point $y_i \in \mathcal{C}_1$ with $\|x_i - y_i\| \leq (1 + \delta_3) d_i$ and by Lemma 3.5,

$$\begin{aligned}(1+\delta_3)^2 d_i^2 \geq \|x_i - y_i\|^2 &\geq t_i^2 + \|\phi(u_i) - y_i\|^2 - 2t_i(\phi(u_i) - y_i)^T n(\phi(y_i)) \\ &= t_i^2 + \|\phi(u_i) - y_i\|^2 + O(t_i \|\phi(u_i) - y_i\|^2).\end{aligned}$$

Specifically, there exists $i_0 > 0$ such that for all $i > i_0$ the remainder term is at most half in absolute value compared to $\|\phi(u_i) - y_i\|^2$ regardless of how δ_3 is chosen. This implies that $t_i \leq (1 + \delta_3) d_i$ and setting $\delta_3 \to 0$ with i fixed, we have $t_i \leq d_i$.

Before proceeding, we still need to verify that $\phi(u_i) \in \mathcal{C}_1$ as well (it could be outside $\partial\mathcal{X}$). To see this, for each i choose $y_i \in \mathcal{C}_1$ with $\|x_i - y_i\| < 2d_i$. Then

$$\|\phi(u_i) - y_i\| \leq \|\phi(u_i) - x_i\| + \|x_i - y_i\| < 3d_i$$

and also $y_i \in \partial\mathcal{X}\backslash N_{15d_i}$ (recall that $c_1 > 17$) proving by Lemma 3.6 that eventually $\phi(u_i) \in \mathcal{C}_1$, because $\phi(u_i) \in B(y_i, 3d_i) \cap \mathcal{D}_1$.

To summarize, when c_1 is large enough, then for any choice $(x_i)_{i=1}^{\infty}$ with $d_i = \rho(x_i, \partial\mathcal{X}) \to 0$ and $x_i \in \partial\mathcal{X} \setminus N_{c_1 d_i}$, x_i belongs to the set $\cup_{x\in\partial\mathcal{X}\backslash N_{\frac{1}{2}c_1 d_i}} A(x, d_i)$ for $i \geq i_0$, where i_0 is a positive integer. In other words, for small r

$$\partial_r\mathcal{X} \setminus N_{c_1 r} \subset \cup_{x\in\partial\mathcal{X}\backslash N_{\frac{1}{2}c_1 r}} A(x, r). \tag{3.35}$$

It is also true that the inclusion

$$\cup_{x\in\partial\mathcal{X}\backslash N_{c_1 r}} A(x, r) \subset \partial_r\mathcal{X} \tag{3.36}$$

can be assumed to be valid in the small r region. It remains to show that

$$\cup_{x\in N_{c_1 r}\cap\partial\mathcal{X}\backslash N_{\frac{1}{2}c_1 r}} A(x, r)$$

and $N_{c_1 r} \cap \mathcal{X}$ have small volumes. The latter holds by condition 2 in (A2):

$$\lambda(N_{c_1 r} \cap \mathcal{X}) = O(r^2).$$

Lemma 3.8 ensures that $A(x, r) \subset \mathcal{X}$ whenever $x \in \partial\mathcal{X} \setminus N_{\frac{1}{2}c_1 r}$ and

$$\cup_{x\in\partial\mathcal{X}\backslash N_{\frac{1}{2}c_1 r}} A(x, r) \setminus \cup_{x\in\partial\mathcal{X}\backslash N_{c_1 r}} A(x, r) \subset N_{c_1 r} \cap \mathcal{X}.$$

Again by by condition 2 in (A2) the set in the right side has a measure of order $O(r^2)$. □

Lemma 3.8 implies that the sets $A(x, r)$ and $A(z, r)$ are disjoint when $x \neq z$ and they also (mostly) cover $\partial_r\mathcal{X}$. For these two reasons, the reparametrization

$$y = x - tn(x)$$

becomes possible for $y \in \partial_r\mathcal{X}$, $x \in \partial\mathcal{X}$ and $0 \leq t \leq r$. One arrives at the result in Lemma 3.2 as proven next.

Proof. (Proof of Lemma 3.2)
To begin with, let us fix a constant $c > 0$ and choose any positive number $r > 0$. Let $x_0 \in \partial\mathcal{X}$ be a point on one or more of the smooth submanifolds $\{\mathcal{D}_i\}_{i=1}^{l}$, say on $\mathcal{D}_1$. Then, there exists a local parametrization $\phi : U \to B(x_0, \delta_0) \cap \mathcal{D}_1$ with U a bounded open set and we may define $V = \phi^{-1}(B(x_0, \delta_0) \cap \partial\mathcal{X} \cap \mathcal{D}_1 \setminus N_{cr})$. Instead of examining the whole integral (3.20), at this point we restrict ourselves to a local neighborhood:

$$\int_{\cup_{x\in B(x_0,\delta_0)\cap\partial\mathcal{X}\cap\mathcal{D}_1\setminus N_{cr}} A(x,r)} f(y)dy. \tag{3.37}$$

Assuming that c is initially large enough and r sufficiently small, $g(u, t) = \phi(u) - tn(\phi(u))$ can be taken as an injection on $V \times [-r, r]$ by Lemma 3.8 and it has the Jacobian

$$J_{(u,t)}g = \left[J_u\phi - tJ_u n(\phi(u)), -n(\phi(u))\right]. \tag{3.38}$$

The Jacobian is important, because its determinant appears after the change of variables. To simplify the expression (3.38), notice that all submatrices in the expression can be assumed to be bounded and we may use (for the l^2-matrix norm)

$$\sup_{\|D\|,\|E\|\leq 1} |\det(D+\epsilon E)-\det(D)| = O(\epsilon)$$

when ϵ approaches zero to conclude that in absolute value, the determinant of $J_{(u,t)}g$ is

$$|\det(J_{(u,t)}g)| = |\det([J_u\phi, -n(\phi(u))])| + O(t) = \sqrt{\det((J_u\phi)^T J_u\phi)} + O(t)$$

allowing us to get rid of the term $J_u n(\phi(u))$. Under the change of variables $y = \phi(u) - tn(\phi(u))$, Equation (3.37) takes the form

$$\begin{aligned}
&\int_{\cup_{x\in B(x_0,\delta_0)\cap\partial\mathcal{X}\cap\mathcal{D}_1\setminus N_{cr}} A(x,r)} f(y)dy \\
&= \int_V \int_0^r f(\phi(u)-tn(\phi(u)))\sqrt{\det((J_u\phi)^T J_u\phi)}dtdu + O(r^2) \\
&= \int_{B(x_0,\delta_0)\cap\partial\mathcal{X}\cap\mathcal{D}_1\setminus N_{cr}} \int_0^r f(x-tn(x))dtdS + O(r^2); \qquad (3.39)
\end{aligned}$$

the remainder term goes to zero at least as fast as r^2 (with x_0 and δ_0 fixed). In the last equality, the standard definition of surface integrals was invoked.

Even though the local considerations are already convincing, a generalization of (3.39) to the whole set is required. By compactness, each submanifold $\mathcal{D}_1$ can be covered with a finite number of sets of the form $\mathcal{D}_1 \cap B(x_i,\delta_i)$ with corresponding local parametrizations ϕ_i. If we set $S_i = B(x_i,\delta_i)\cap\partial\mathcal{X}\cap\mathcal{D}_1\setminus N_{cr}$, then one can examine each ball separately by replacing f with

$$f_i(x) = I(x \notin \cup_{y\in S_i\setminus\cup_{k=0}^{i-1}S_k} A(y,r))f(x)$$

to take into account the overlap between the sets. By proceeding through all the submanifolds $\{\mathcal{D}_j\}_{j=1}^l$ we are able to cover the whole set $\partial\mathcal{X}\setminus N_{cr}$ (the number of sets in the cover is independent of r and c) and for the resulting functions $f_1,\dots,f_s$,

$$\begin{aligned}
\int_{\partial_r\mathcal{X}} f(y)dy &= \int_{\cup_{x\in\partial\mathcal{X}\setminus N_{cr}} A(x,r)} f(y)dy + O(r^2) \\
&= \sum_{i=0}^{s} \int_{\cup_{x\in S_i\setminus\cup_{k=0}^{i-1}S_k} A(x,r)} f(y)dy + O(r^2) \\
&= \sum_{i=0}^{s} \int_{S_i} \int_0^r f_i(x-tn(x))dtdS + O(r^2) \\
&= \sum_{i=0}^{s} \int_{S_i\setminus\cup_{k=0}^{i-1}S_k} \int_0^r f(x-tn(x))dtdS + O(r^2) \\
&= \int_{\partial\mathcal{X}} \int_0^r f(x-tn(x))dtdS + O(r^2).
\end{aligned}$$

Lemmas 3.7-3.9 and the local analysis in the first part of the proof were applied in the three first equalities, whereas (A2) establishes the last one. □

3.7 The Interior

It was shown in Lemma 3.1 that when considering $E[d_{1,k}^{\alpha}|X_1]$, the cut-off

$$E[d_{1,k}^{\alpha} I(d_{1,k} \leq t_M)|X_1]$$

is possible if we set

$$t_M = M^{-1/n} \log^{2/n} M \tag{3.40}$$

as the difference between the thresholded and original expectations approaches zero fast. After introducing the cut-off, $\mathcal{X} \setminus \partial_{t_M} \mathcal{X}$ can be considered as the interior points of $\mathcal{X}$. The idea is that in this set, the boundaries can be neglected and a linearization of q is possible. In this sense, heuristically speaking $\mathcal{X} \setminus \partial_{t_M} \mathcal{X}$ contains the easy points.

We define the function

$$\begin{aligned} g_M(x,r) &= 1 \text{ if } x \in \mathcal{X} \setminus \partial_{t_M} \mathcal{X} \text{ and } 0 < r < t_M; \\ g_M(x,r) &= 0 \text{ otherwise.} \end{aligned} \tag{3.41}$$

The random variable $g_M(X_1, d_{1,k})$ ensures that X_1 is in the interior and $d_{1,k}$ does not exceed t_M.

Recall Equation (2.17), which says that

$$P(\omega_{X_1}(d_{1,k}) > z|X_1) = \sum_{j=0}^{k-1} \binom{M-1}{j} z^j (1-z)^{M-j-1} \tag{3.42}$$

regardless of q. One immediately sees from Equation (3.42) that if $d_{1,k}$ could be replaced by $\omega_{X_1}(d_{1,k})^{1/n}$, then the problem of estimating the moments $d_{1,k}^{\alpha}$ would reduce into a relatively simple integral. In fact, taking into account the cut-off, we have

Lemma 3.10. *If (A2)-(A4) hold, then for any bounded function* $0 \leq f \leq 1$ *and* $M > 2k$,

$$\begin{aligned} &E[f(X_1)\omega_{X_1}(d_{1,k})^{\alpha/n} g_M(X_1, d_{1,k})] \\ &\qquad = \frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)} \int_{\mathcal{X} \setminus \partial_{t_M} \mathcal{X}} f(x)q(x)dx + R \end{aligned}$$

with

$$|R| \leq M^k e^{-c t_M^n M}$$

for a constant $c > 0$ *independent of* M.

Proof. From Lemma 3.1 we know that when $x \in \mathcal{X} \setminus \partial_{t_M} \mathcal{X}$,

$$\begin{aligned} E[f(X_1)\omega_{X_1}(d_{1,k})^{\alpha/n}(1 - g_M(X_1, d_{1,k}))|X_1 = x] &\leq P(d_{1,k} > t_M | X_1 = x) \\ &\leq M^k e^{-cMt_M^n}. \end{aligned}$$

On the other hand, by Theorem 2.3,

$$\begin{aligned}E[f(X_1)\omega_{X_1}(d_{1,k})^{\alpha/n}|X_1] &= E[\omega_{X_1}(d_{1,k})^{\alpha/n}|X_1]f(X_1)\\ &= \frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)}f(X_1).\end{aligned}$$

□

Using Lemma 3.10 it is possible to evaluate

$$E[d_{1,k}^\alpha g_M(X_1, d_{1,k})]:$$

Lemma 3.11. *Suppose that (A2)-(A4) hold for some $1 \le \gamma \le 2$ and set t_M according to Equation (3.40). Then there exists a constant c_1 (depending on $\mathcal{X}$, q and k, but not on M) such that we have the estimate*

$$\begin{aligned}&|E[d_{1,k}^\alpha g_M(X_1, d_{1,k})] - V_n^{-\alpha/n}\frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)}\int_{\mathcal{X}\backslash\partial_{t_M}\mathcal{X}} q(x)^{1-\alpha/n}dx|\\ &\le c_1 M^{-\gamma/n-\alpha/n}\log^{2\gamma/n+2\alpha/n} M. \qquad (3.43)\end{aligned}$$

Proof. By algebraic manipulation,

$$\begin{aligned}d_{1,k}^\alpha g_M &= V_n^{-\alpha/n}q(X_1)^{-\alpha/n}(V_n q(X_1)d_{1,k}^n)^{\alpha/n}g_M\\ &= V_n^{-\alpha/n}q(X_1)^{-\alpha/n}\omega_{X_1}(d_{1,k})^{\alpha/n}(\frac{V_n q(X_1)d_{1,k}^n - \omega_{X_1}(d_{1,k})}{\omega_{X_1}(d_{1,k})}+1)^{\alpha/n}g_M.\end{aligned} \qquad (3.44)$$

The arguments of g_M were dropped for notational convenience. The Hölder continuity of q implies that

$$|\frac{V_n q(X_1)d_{1,k}^n - \omega_{X_1}(d_{1,k})}{\omega_{X_1}(d_{1,k})}|g_M \le c_1 c_2^{-1} t_M g_M < \frac{1}{2}, \qquad (3.45)$$

where c_1 is the Hölder constant of q and c_2 is the lower bound. By the mean value theorem

$$|(1+x)^{\alpha/n} - 1 - \frac{\alpha}{n}x| \le 2^{\alpha/n+2}\frac{\alpha}{n}|\frac{\alpha}{n}-1|x^2$$

when $0 < x < 1/2$, which together with Equation (3.45) can be applied for estimating the expression (3.44):

$$\begin{aligned}&(\frac{V_n q(X_1)d_{1,k}^n - \omega_{X_1}(d_{1,k})}{\omega_{X_1}(d_{1,k})}+1)^{\alpha/n}g_M\\ &= (1 + \frac{\alpha V_n q(X_1)d_{1,k}^n - \alpha\omega_{X_1}(d_{1,k})}{n\omega_{X_1}(d_{1,k})})g_M + R_1 \qquad (3.46)\end{aligned}$$

with an error term bounded by (the constants are not the best possible due to notational convenience)

$$|R_1| \le \frac{4^{1+\alpha/n}\alpha|n-\alpha|}{n^2}c_1^2 c_2^{-2} t_M^2 \le c_3 t_M^2,$$

where the constant c_3 collects the terms in front of t_M^2. To assess the first term on the right side of Equation (3.46), we observe that if $X_1 \in \mathcal{X} \setminus \partial_{t_M}\mathcal{X}$, then by a Taylor expansion

$$\omega_{X_1}(d_{1,k})g_M = V_n q(X_1) d_{1,k}^n g_M + \int_{B(X_1,d_{1,k})} (x - X_1)^T \nabla_{X_1} q dx \; g_M + R_2 \quad (3.47)$$

with

$$|R_2| \leq c_4 t_M^{n+\gamma}$$

for some $c_4 > 0$. Moreover, by symmetry

$$\int_{B(X_1,d_{1,k})} (x - X_1)^T \nabla_{X_1} q dx = 0 \quad (3.48)$$

and thus by Equation (3.47),

$$|\frac{V_n q(X_1) d_{1,k}^n - \omega_{X_1}(d_{1,k})}{\omega_{X_1}(d_{1,k})}| g_M \leq V_n^{-1} c_2^{-1} c_4 t_M^{\gamma}. \quad (3.49)$$

Because q is bounded, we may also write that

$$\omega_{X_1}(d_{1,k}) \leq c_5 d_{1,k}^n \quad (3.50)$$

for some constant $c_5 > 0$ depending only on q and n. Putting these considerations together, we derive from Equation (3.44) using (3.46), (3.49) and (3.50) that

$$\begin{aligned} d_{1,k}^{\alpha} g_M &= V_n^{-\alpha/n} q(X_1)^{-\alpha/n} \omega_{X_1}(d_{1,k})^{\alpha/n} g_M \\ &+ \frac{\alpha}{n} V_n^{-\alpha/n} q(X_1)^{-\alpha/n} \omega_{X_1}(d_{1,k})^{\alpha/n} \frac{V_n q(X_1) d_{1,k}^n - \omega_{X_1}(d_{1,k})}{\omega_{X_1}(d_{1,k})} g_M \\ &+ V_n^{-\alpha/n} q(X_1)^{-\alpha/n} \omega_{X_1}(d_{1,k})^{\alpha/n} g_M R_1 \\ &= V_n^{-\alpha/n} q(X_1)^{-\alpha/n} \omega_{X_1}(d_{1,k})^{\alpha/n} g_M + R_3 \end{aligned}$$

with

$$|R_3| \leq c_6 t_M^{\gamma+\alpha}.$$

The proof is finalized by an application of Lemma 3.10:

$$E[q(X_1)^{-\alpha/n} \omega_{X_1}(d_{1,k})^{\alpha/n} g_M] = \frac{\Gamma(k + \alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M + \alpha/n)} \int_{\mathcal{X} \setminus \partial_{t_M}\mathcal{X}} q(x)^{1-\alpha/n} dx + R_4$$

with

$$|R_4| \leq \lambda(\mathcal{X}) M^k e^{-ct_M^n M} \leq \lambda(\mathcal{X}) e^{-ck \log^2 M + k \log M},$$

which decays to zero faster than any $M^{-\beta}$ $(\beta > 0)$. □

If $\gamma < 1$, then the symmetry argument in Equation (3.48) is not useful as the gradient does not exist in that case. Nevertheless the error in the expansion is of order t_M^{γ}, which however, converges to zero too slowly in comparison with the terms arising from the boundary effect. For this reason, $1 < \gamma \leq 2$ is necessary to make our proof technique to work.

The analogue of Lemma 3.11 for the logarithm is

Lemma 3.12. *If (A4) holds, then almost surely for any $a \geq 0$,*

$$E[\omega_{X_1}(d_{1,k})^a \log \omega_{X_1}(d_{1,k})|X_1] = \frac{\Gamma(M)}{\Gamma(M+a)}(\psi(k+a) - \frac{\Gamma(k+a)\psi(M+a)}{\Gamma(k)}),$$

where $\psi(\cdot)$ denotes the digamma function.

Proof. Define the function $f(\epsilon)$ by

$$f(\epsilon) = E[\omega_{X_1}(d_{1,k})^{a+\epsilon}|X_1] = \frac{\Gamma(k+a+\epsilon)\Gamma(M)}{\Gamma(k)\Gamma(M+a+\epsilon)}.$$

If $a > 0$, then the derivative of f at $\epsilon = 0$ is $E[\omega_{X_1}(d_{1,k})^a \log \omega_{X_1}(d_{1,k})|X_1]$, because the derivative can be taken inside the integral. The claim is then proved by the definition of digamma functions via derivatives of gamma functions. □

Using Lemma 3.12, an analysis for the logarithmic distances in the interior can be established in a similar way as for the α-moments. The proof is nearly the same, but the final formula is different.

Lemma 3.13. *Suppose that (A2)-(A4) hold for some $1 \leq \gamma \leq 2$ and set t_M according to Equation (3.40). Then there exists a constant c (depending on $\mathcal{X}$ and q, but not on M and k) such that we have the estimate*

$$\begin{aligned}&|nE[g_M(X_1, d_{1,k}) \log d_{1,k}] \\ &\quad - \psi(k) + \psi(M) + \log V_n + \int_{\mathcal{X}\backslash\partial_{t_M}\mathcal{X}} q(x) \log q(x)dx| \\ &\leq cM^{-\gamma/n} \log^{2\gamma/n} M. \end{aligned} \tag{3.51}$$

Proof. When $g_M = 1$, we may write

$$n \log d_{1,k} = \log \omega_{X_1}(d_{1,k}) - \log q(X_1) - \log V_n + \log(1 + \frac{d_{1,k}^n q(X_1)V_n - \omega_{X_1}(d_{1,k})}{\omega_{X_1}(d_{1,k})}).$$

Again, we apply the mean value theorem, this time to the function $\log(1+x) \approx x$:

$$g_M \log(1 + \frac{d_{1,k}^n q(X_1)V_n - \omega_{X_1}(d_{1,k})}{\omega_{X_1}(d_{1,k})}) = g_M \frac{d_{1,k}^n q(X_1)V_n - \omega_{X_1}(d_{1,k})}{\omega_{X_1}(d_{1,k})} + R \tag{3.52}$$

and recalling the symmetry argument (3.49),

$$|R| \leq c_1 t_M^\gamma \tag{3.53}$$

for an appropriate constant $c_1 > 0$. We still have to show that $E[\log \omega_{X_1}(d_{1,k})|X_1]$ is close to $E[I(d_{1,k} \leq t_M) \log \omega_{X_1}(d_{1,k})|X_1]$, that is,

$$|E[I(d_{1,k} > t_M) \log \omega_{X_1}(d_{1,k})|X_1]|$$

is small. This is not too hard as by Lemma 3.1,

$$\begin{aligned}|E[I(d_{1,k} > t_M)\log\omega_{X_1}(d_{1,k})|X_1]| &\leq -P(d_{1,k} > t_M|X_1)\log\omega_{X_1}(t_M)\\ &\leq -e^{-c_2k\log^2 M+k\log M}\log\omega_{X_1}(t_M)\end{aligned}$$

for some $c_2 > 0$, which as stated in the proof of Lemma 3.11 is neglible compared to the other sources of estimation error, because

$$|\log\omega_{X_1}(t_M)| \leq \log c_3 - n\log t_M$$

for some constant $c_3 > 0$. □

In fact, Equation (3.51) comes from (3.43) by taking the derivative w.r.t. α if the error terms are not considered.

3.8 Nearest Neighbors Close to the Boundaries

In Section 3.7 it turned out that when points close to the boundaries of $\mathcal{X}$ are excluded, it is possible to cope with relatively few regularity conditions. For points close to the boundaries the situation looks very different as q can no longer be approximated by a Taylor expansion. Nevertheless a transition to the unit cube is possible by simultaneous linearization of the boundary and the density q as demonstrated in Figure 3.11. As an important point, the same linearization is used for all the points on the dashed line $A(x_1, r)$. The argument is formalized in Section 3.8.2 by utilizing the geometric analysis of Section 3.6 as sketched in Section 3.5.

Under the assumption of uniform points $(X_i)_{i=1}^M$ on $[0,1]\times[-1/2,1/2]^{n-1}$, in Section 3.8.1 the expectations $E[d_{1,k}^\alpha|X_1 = (s,0,\ldots,0)]$ are considered. Even though the nearest neighbor distributions of Section 3.4 do not seem to allow any simple representation of

$$E[d_{1,k}^\alpha|X_1 = (s,0,\ldots,0)],$$

it turns out that somewhat surprisingly,

$$\int_0^{t_M} E[d_{1,k}^\alpha I(d_{1,k} \leq t_M)|X_1 = (s,0,\ldots,0)]ds$$

can be estimated with a high accuracy. This trick then applies straightforwardly in Section 3.8.2.

3.8.1 The Unit Cube

Recalling the function W in Definition 3.2, we have the following important lemma. The proof is based on the use of Fubini's theorem for interchanging the order of integration.

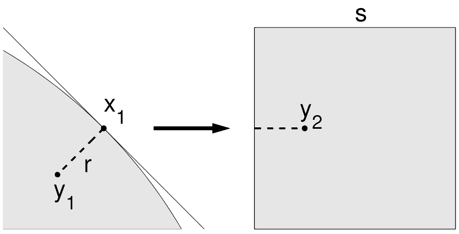


Figure 3.11: From general $\mathcal{X}$, linearization allows a transition to the uniform distribution on the cube with sidelength $s = q(x_1)^{-1/n}$ and the translated point $y_2 = (q(x_1)^{-1/n}r, 0, \dots, 0)$ with the length of the second dashed line given by $q(x_1)^{-1/n}r$.

Lemma 3.14. *Let us define the constant*

$$D = \frac{1}{n}\int_0^1 a^{-\alpha-2}W(a^{-1})^{-\alpha/n-1/n-1}W'(a^{-1})da.$$

If the variables $(X_i)_{i=1}^M$ are uniform on the cube $[0,1]\times[-1/2,1/2]^{n-1}$, we have for any $\alpha > 0$, $t_M = M^{-1/n}\log^{2/n} M$ and $M > 2k$,

$$\int_0^{t_M} E[I(d_{1,k} \le t_M)d_{1,k}^\alpha | X_1 = (s,0,\dots,0)]ds = V_n^{-\alpha/n}\frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)}t_M$$
$$+ (D - V_n^{-\alpha/n-1/n})\frac{\Gamma(k+\alpha/n+1/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n+1/n)} + R.$$

The remainder term is bounded by

$$|R| \le M^k e^{-\frac{1}{4}V_n t_M^n M}(D + V_n^{-\alpha/n} + V_n^{-\alpha/n-1/n}).$$

Proof. Recalling the nearest neighbor distribution in Theorem 3.5, we obtain

$$\int_0^{t_M} E[I(d_{1,k} \le t_M)d_{1,k}^\alpha | X_1 = (s,0,\dots,0)]ds$$
$$= k\binom{M-1}{k}\int_0^{t_M}\int_0^{t_M} r^\alpha \omega_{(s,0)}(r)^{k-1}(1-\omega_{(s,0)}(r))^{M-k-1}\, d\omega_{(s,0)}(r)ds$$
$$= k\binom{M-1}{k}\int_0^{t_M}\int_0^{t_M} t(s,r)drds. \tag{3.54}$$

Because

$$\omega_{(s,0)}(r) = s^n W(\frac{r}{s}),$$

we have

$$t(s,r) = r^\alpha s^{n-1}\omega_{(s,0)}(r)^{k-1}(1-\omega_{(s,0)}(r))^{M-k-1}W'(\frac{r}{s}).$$

The integral (3.54) can be divided into two parts by considering sets with $s > r$ and $s < r$ separately. Because $W(r) = V_n r^n$ when $r < 1$, it is true that for $s > r$,

$$W'(\frac{r}{s}) = \frac{nV_n r^{n-1}}{s^{n-1}}$$

and

$$\begin{aligned} I_1 &= k\binom{M-1}{k}\int_0^{t_M}\int_r^{t_M} t(s,r)dsdr \\ &= nV_n^k k\binom{M-1}{k}\int_0^{t_M}\int_r^{t_M} r^{\alpha+kn-1}(1-V_nr^n)^{M-k-1}dsdr \\ &= nV_n^k k\binom{M-1}{k}\int_0^{t_M} r^{\alpha+kn-1}(t_M-r)(1-V_nr^n)^{M-k-1}dr. \end{aligned}$$

By making the change of variable $y = V_nr^n$, I_1 can be written as

$$\begin{aligned} I_1 =& V_n^{-\alpha/n}t_M k\binom{M-1}{k}\int_0^{V_nt_M^n} y^{\alpha/n+k-1}(1-y)^{M-k-1}dy \\ &- V_n^{-\alpha/n-1/n}k\binom{M-1}{k}\int_0^{V_nt_M^n} y^{\alpha/n+1/n+k-1}(1-y)^{M-k-1}dy. \end{aligned}$$

The integrals from 0 to $V_nt_M^n$ can be extended to integrals from 0 to 1 at the expense of an error term roughly bounded by

$$\begin{aligned} |R_1| &\leq M^k(V_n^{-\alpha/n}+V_n^{-\alpha/n-1/n})\int_{V_nt_M^n}^1 y^{\alpha/n+k-1}(1-y)^{M-k-1}dy \\ &\leq M^k(V_n^{-\alpha/n}+V_n^{-\alpha/n-1/n})e^{-V_nt_M^n(M-k-1)}. \end{aligned}$$

Applying the connection to the beta function as in Equation (2.18), we obtain the final form of I_1:

$$\begin{aligned} I_1 = V_n^{-\alpha/n}t_M\frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)} &- V_n^{-\alpha/n-1/n}\frac{\Gamma(k+\alpha/n+1/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n+1/n)} \\ &+ R_1. \end{aligned}$$

Next we proceed to the slightly more difficult case $r > s$. One possible approach is to make the change of variable $(s,r) = (ar,r)$ to obtain

$$\begin{aligned} I_2 &= k\binom{M-1}{k}\int_0^{t_M}\int_0^r t(s,r)dsdr = k\binom{M-1}{k}\int_0^1\int_0^{t_M} t(ar,r)rdrda \\ &= k\binom{M-1}{k}\int_0^1 a^{kn-1}W(a^{-1})^{k-1}W'(a^{-1}) \\ &\times\int_0^{t_M} r^{\alpha+nk}(1-a^nW(a^{-1})r^n)^{M-k-1}drda \\ &= \frac{k}{n}\binom{M-1}{k}\int_0^1 a^{-\alpha-2}W(a^{-1})^{-\alpha/n-1-1/n}W'(a^{-1}) \\ &\times\int_0^{a^nW(a^{-1})t_M^n} y^{\alpha/n+1/n+k-1}(1-y)^{M-k-1}dyda. \end{aligned}$$

Intuitively, the half-plane can cut at most half of a ball; thus, $W(a^{-1}) \geq \frac{1}{2}V_na^{-n}$ implying

$$a^nW(a^{-1})t_M^n > \frac{1}{2}V_nt_M^n$$

and consequently

$$\begin{aligned} I_2 &= Dk\binom{M-1}{k}\int_0^1 y^{\alpha/n+1/n+k-1}(1-y)^{M-k-1}dy + R_2 \\ &= D\frac{\Gamma(k+\alpha/n+1/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n+1/n)} + R_2 \end{aligned}$$

the error term being bounded by

$$|R_2| \leq DM^k \int_{\frac{1}{2}V_n t_M^n}^1 y^{\alpha/n+1/n+k-1}(1-y)^{M-k-1}dy \leq DM^k e^{-\frac{1}{2}V_n t_M^n(M-k-1)}.$$

□

The calculations for the logarithm are rather similar, but unfortunately slightly more technical.

Lemma 3.15. *If the variables $(X_i)_{i=1}^M$ are uniform on the cube $[0,1]\times[-1/2,1/2]^{n-1}$, we have for any $\alpha > 0$, $M > 2k$ and $t_M = M^{-1/n}\log^{2/n} M$,*

$$\begin{aligned} &\int_0^{t_M} E[I(d_{1,k} \leq t_M)\log d_{1,k}|X_1 = (s,0,\dots,0)]ds \\ &\qquad = C_1\frac{\Gamma(M)}{\Gamma(M+1/n)} + t_M C_2 \end{aligned}$$

with

$$\begin{aligned} C_1 &= \frac{V_n^{-1/n}\Gamma(k+1/n)\log V_n}{n\Gamma(k)} + \frac{V_n^{-1/n}\psi(M+1/n)\Gamma(k+1/n)}{n\Gamma(k)} \\ &\quad + D_1(\psi(k+1/n) - \frac{\Gamma(k+1/n)\psi(M+1/n)}{\Gamma(k)}) - \frac{V_n^{-1/n}\psi(k+1/n)}{n} \\ &\quad + \frac{D_2\Gamma(k+1/n)}{\Gamma(k)} \qquad (3.55) \\ C_2 &= \frac{1}{n}(\psi(k) - \psi(M) - \log V_n) \qquad (3.56) \\ D_1 &= \frac{1}{n^2}\int_0^1 a^{-2}W(a^{-1})^{-1/n-1}W'(a^{-1})da \qquad (3.57) \\ D_2 &= -\frac{1}{n^2}\int_0^1 a^{-2}W(a^{-1})^{-1/n-1}W'(a^{-1})(\log W(a^{-1}) + n\log a)da. \qquad (3.58) \end{aligned}$$

Proof. As in Equation (3.54),

$$\begin{aligned} &\int_0^{t_M} E[I(d_{1,k} \leq t_M)\log d_{1,k}|X_1 = (s,0,\dots,0)]ds \\ &\qquad = k\binom{M-1}{k}\int_0^{t_M}\int_0^{t_M} t(s,r)drds, \end{aligned}$$

but this time

$$t(s,r) = s^{n-1}\omega_{(s,0)}(r)^{k-1}(1-\omega_{(s,0)}(r))^{M-k-1}W'(\frac{r}{s})\log r.$$

Continuing in the same fashion as in the proof of Lemma 3.14, we obtain

$$\begin{aligned}I_1 &= k\binom{M-1}{k}\int_0^{t_M}\int_r^{t_M} t(s,r)dsdr\\ &= nV_n^k k\binom{M-1}{k}\int_0^{t_M}\int_r^{t_M} r^{kn-1}(1-V_nr^n)^{M-k-1}\log r\ dsdr\\ &= nV_n^k k\binom{M-1}{k}\int_0^{t_M} r^{kn-1}(t_M-r)(1-V_nr^n)^{M-k-1}\log r\ dr.\end{aligned}$$

By making the change of variable $y = V_nr^n$, I_1 can be written as

$$\begin{aligned}I_1 =& t_M k\binom{M-1}{k}\int_0^{V_nt_M^n} y^{k-1}(1-y)^{M-k-1}\log(V_n^{-1/n}y^{1/n})dy\\ &- V_n^{-1/n}k\binom{M-1}{k}\int_0^{V_nt_M^n} y^{k+1/n-1}(1-y)^{M-k-1}\log(V_n^{-1/n}y^{1/n})dy.\end{aligned}$$

The integral is extended up to 1 from $V_nt_M^n$; this results in an error term R_1 bounded by

$$|R_1| \leq -cM^ke^{-V_n(M-k-1)t_M^n}\log t_M,$$

where c does not depend on M. Possibly the easiest way to analyze I_1 is via observing that by taking the derivative of Equation (2.17), we have

$$P(\omega_{X_1}(d_{1,k}) \in [z, z+dz]) = k\binom{M-1}{k}z^{k-1}(1-z)^{M-k-1}dz, \tag{3.59}$$

which is equivalent to saying that for bounded measurable functions f,

$$E[f(\omega_{X_1}(d_{1,k}))] = k\binom{M-1}{k}\int_0^1 f(z)z^{k-1}(1-z)^{M-k-1}dz.$$

Using Lemma 3.12 and Equation (3.59), we obtain

$$\begin{aligned}&k\binom{M-1}{k}\int_0^1 y^{k+a-1}(1-y)^{M-k-1}\log y\ dy = E[\omega_{X_1}(d_{1,k})^a\log\omega_{X_1}(d_{1,k})]\\ &= \frac{\Gamma(M)}{\Gamma(M+a)}(\psi(k+a) - \frac{\Gamma(k+a)\psi(M+a)}{\Gamma(k)}).\end{aligned} \tag{3.60}$$

By collecting all the terms and applying Equation (3.60) we calculate

$$\begin{aligned}I_1 &= -\frac{t_M}{n}\log V_n + \frac{t_M}{n}(\psi(k)-\psi(M)) + \frac{V_n^{-1/n}\log V_n}{n}\frac{\Gamma(k+1/n)\Gamma(M)}{\Gamma(k)\Gamma(M+1/n)}\\ &\quad -\frac{V_n^{-1/n}}{n}\psi(k+1/n)\frac{\Gamma(M)}{\Gamma(M+1/n)} + \frac{V_n^{-1/n}}{n}\psi(M+1/n)\frac{\Gamma(k+1/n)\Gamma(M)}{\Gamma(k)\Gamma(M+1/n)}\\ &= \frac{t_M}{n}(\psi(k)-\psi(M)-\log V_n) + (\frac{\Gamma(k+1/n)}{\Gamma(k)}\log V_n - \psi(k+1/n)\\ &\quad +\frac{\psi(M+1/n)\Gamma(k+1/n)}{\Gamma(k)})\frac{V_n^{-1/n}\Gamma(M)}{n\Gamma(M+1/n)}.\end{aligned}$$

We move to examine the term I_2 corresponding to $r > s$, which now takes the form

$$\begin{aligned}
I_2 &= k\binom{M-1}{k}\int_0^{t_M}\int_0^{r} t(s,r)dsdr = k\binom{M-1}{k}\int_0^{1}\int_0^{t_M} t(ar,r)rdrda \\
&= k\binom{M-1}{k}\int_0^1 a^{kn-1}W(a^{-1})^{k-1}W'(a^{-1}) \\
&\quad\times\int_0^{t_M} r^{nk}(1-a^nW(a^{-1})r^n)^{M-k-1}\log r\ drda \\
&= \frac{k}{n^2}\binom{M-1}{k}\int_0^1 a^{-2}W(a^{-1})^{-1/n-1}W'(a^{-1}) \\
&\quad\times\int_0^{a^nW(a^{-1})t_M^n} y^{k+1/n-1}(1-y)^{M-k-1}(\log y-\log(a^nW(a^{-1})))dyda.
\end{aligned}$$

Again there is no problem in extending $a^nW(a^{-1})t_M^n$ to 1. The terms dependent on a are denoted by D_1 and D_2:

$$\begin{aligned}
D_1 &= \frac{1}{n^2}\int_0^1 a^{-2}W(a^{-1})^{-1/n-1}W'(a^{-1})da \\
D_2 &= -\frac{1}{n^2}\int_0^1 a^{-2}W(a^{-1})^{-1/n-1}W'(a^{-1})(\log W(a^{-1})+n\log a)da.
\end{aligned}$$

Moreover, the integrals over y are already calculated in Equation (3.60). Thus

$$I_2 = D_1(\psi(k+1/n) - \frac{\Gamma(k+1/n)\psi(M+1/n)}{\Gamma(k)})\frac{\Gamma(M)}{\Gamma(M+1/n)} + D_2\frac{\Gamma(M)\Gamma(k+1/n)}{\Gamma(M+1/n)\Gamma(k)}.$$

□

3.8.2 Smooth Sets

The following lemma connects sampling and the total variation distance allowing a transition between two measures. The proof is taken from [43].

Lemma 3.16. *Suppose that $(X_i)_{i=1}^M$ is i.i.d. with respect to a probability density q. Assume now that $\tilde{q}$ is another density function; the L^1-distance between the densities is*

$$\|q-\tilde{q}\|_1 = \int_{\Re^n}|q(x)-\tilde{q}(x)|dx.$$

Then, there exists (formally by extending the underlying probability space $(\Omega,\mathcal{F},P)$) another i.i.d. sample $(\tilde{X}_i)_{i=1}^M$ (distributed according to $\tilde{q}$) such that

$$P((\tilde{X}_i)_{i=1}^M = (X_i)_{i=1}^M) \geq 1 - \frac{1}{2}M\|q-\tilde{q}\|_1.$$

Proof. If we set $[q(x)-\tilde{q}(x)]_+$ as $q(x)-\tilde{q}(x)$ when $q(x)\geq\tilde{q}(x)$ and 0 otherwise, then it holds that

$$\begin{aligned} a = \frac{1}{2}\int_{\Re^n}|q(x)-\tilde{q}(x)|dx &= \int_{\Re^n}[q(x)-\tilde{q}(x)]_+dx \\ &= 1-\int_{\Re^n}\min\{q(x),\tilde{q}(x)\}dx. \end{aligned}$$

On $\Re^n\times\Re^n$, we define the two probability measures $\mathcal{Q}_1$ and $\mathcal{Q}_2$ by

$$\int_{\Re^n\times\Re^n} f(x,y)d\mathcal{Q}_1 = (1-a)^{-1}\int_{\Re^n} f(x,x)\min\{q(x),\tilde{q}(x)\}dx$$

and

$$\int_{\Re^n\times\Re^n} f(x,y)d\mathcal{Q}_2 = a^{-2}\int_{\Re^n}[q(x)-\tilde{q}(x)]_+[q(y)-\tilde{q}(y)]_+f(x,y)dxdy.$$

We set

$$\mathcal{Q} = (1-a)\mathcal{Q}_1 + a\mathcal{Q}_2.$$

Each pair $(X_i,\tilde{X}_i)$ is sampled w.r.t. $\mathcal{Q}$. By integrating out w.r.t. x and y respectively, it can be verified that the marginal distributions of $\mathcal{Q}$ match with q and $\tilde{q}$. Moreover,

$$\mathcal{Q}(X_i\neq\tilde{X}_i)\leq a = \frac{1}{2}\int_{\Re^n}|q(x)-\tilde{q}(x)|dx$$

as required. Finally, the union bound ensures that

$$\mathcal{Q}((X_i)_{i=1}^M\neq(\tilde{X}_i)_{i=1}^M)\leq\sum_{i=1}^M\mathcal{Q}(X_i\neq\tilde{X}_i)\leq\frac{1}{2}M\int_{\Re^n}|q(x)-\tilde{q}(x)|dx.$$

□

Lemma 3.16 is now applied to generalize the unit cube analysis. $A(x,t_M)$ of Equation (3.16) replaces the line $\{(s,0,\ldots,0):0<s<t_M\}$ for the cube. A double linearization is performed: the theory in Section 3.6 to linearize the boundary and a Taylor expansion to ensure the local linearity of q.

Lemma 3.17. *Assume that (A2)-(A4) hold with $\gamma\geq 1$ in (A4). Moreover, choose $t_M = M^{-1/n}\log^{2/n}M$ and $x\in\partial\mathcal{X}\setminus N_{c_1t_M}$, where $c_1>1$ is such that Lemma 3.8 holds, that is, $A(x,t_M)\subset\mathcal{X}$, $A(x,-t_M)\subset\mathcal{X}^C$ and $A(x,t_M)\cap A(y,t_M)=\emptyset$ when $y\in\partial\mathcal{X}\setminus N_{c_1t_M}$ with $x\neq y$. Then for any $\alpha>0$,*

$$\begin{aligned} &\int_0^{t_M}E[I(d_{1,k}\leq t_M)d_{1,k}^\alpha|X_1=x-sn(x)]q(x-sn(x))ds \\ =&V_n^{-\alpha/n}\frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)}\int_0^{t_M}q(x-sn(x))^{1-\alpha/n}ds \\ &+q(x)^{1-\alpha/n-1/n}(D-V_n^{-\alpha/n-1/n})\frac{\Gamma(k+\alpha/n+1/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n+1/n)}+R \end{aligned}$$

with

$$|R|\leq c_2M^{-\alpha/n-2/n}\log^{2+2\alpha/n+4/n}M$$

for some constant c_2 (depending only on q, $\mathcal{X}$ and k). D was defined in Lemma 3.14.

Proof. By rotation and translation, we may assume without losing generality that

$$x = (0, 0, \dots, 0)$$

and similarly $n(x) = (-1, 0, \dots, 0)$. Given $y = x - sn(x)$, the notation Ξ_1 refers to $B(y, t_M) \cap \mathcal{U}_x$ and Ξ_2 denotes $B(y, t_M) \cap \mathcal{X}$ as in Lemma 3.3. Define a new density $\tilde{q}$ by setting

$$\tilde{q}(z) = q(x)I(z \in \Xi_1)$$

for $z \in B(y, t_M)$ and

$$\tilde{q}(z) = \frac{(1 - q(x)\lambda(\Xi_1))q(z)}{1 - P(X_1 \in \Xi_2)}$$

otherwise. Lemma 3.3 and Assumption (A4) ensure the existence of constants $c_1, c_2 > 0$ such that

$$\begin{aligned} \int_{B(y,t_M)} |\tilde{q}(z) - q(z)|dz &\leq c_1 V_n t_M^{n+1} + q(x)\lambda(\Xi_1 \setminus \Xi_2) + q(x)\lambda(\Xi_2 \setminus \Xi_1) \\ &\leq c_2 t_M^{n+1}. \end{aligned} \tag{3.61}$$

Moreover, for t_M small enough to ensure $P(X_1 \in \Xi_2) \leq 1/2$, we have

$$|1 - \frac{1 - q(x)\lambda(\Xi_1)}{1 - P(X_1 \in \Xi_2)}| \leq c_3 t_M^{n+1}.$$

This implies the inequality

$$\int_{\Re^n} |\tilde{q}(z) - q(z)|dz \leq c_4 t_M^{n+1}$$

for some constant $c_4 > 0$. By the coupling argument (Lemma 3.16), there exists an i.i.d. sample $(\tilde{X}_i^{(1)})_{i=1}^M$ distributed according to the density $\tilde{q}$ such that for each $i > 1$,

$$P(X_i \neq \tilde{X}_i^{(1)}) \leq c_4 t_M^{n+1}$$

and consequently

$$P((X_i)_{i=2}^M \neq (\tilde{X}_i^{(1)})_{i=2}^M) \leq \sum_{i=2}^M P(X_i \neq \tilde{X}_i^{(1)}) \leq c_4 M t_M^{n+1}.$$

The new sample has a convenient uniformity property in the neighborhood of y. Taking $X_1 = \tilde{X}_1^{(1)}$ independent of $(X_i, \tilde{X}_i^{(1)})_{i=2}^M$ leads to the formula

$$|E[I(d_{1,k} \leq t_M)d_{1,k}^\alpha | X_1 = y] - E[I(\tilde{d}_{1,k,1} \leq t_M)\tilde{d}_{1,k,1}^\alpha | \tilde{X}_1^{(1)} = y]| \leq 2c_4 M t_M^{n+\alpha+1}, \tag{3.62}$$

because on the event $(X_i)_{i=2}^M = (\tilde{X}_i^{(1)})_{i=2}^M$ the nearest neighbor distances are the same for both samples. The notation $\tilde{d}_{1,k,1}$ refers to the k-nearest neighbor in the sample $(\tilde{X}_i^{(1)})_{i=1}^M$.

Using the notation $\Delta = [0, q(x)^{-1/n}] \times [-q(x)^{-1/n}/2, q(x)^{-1/n}/2]^{n-1}$, we introduce a third sample, $(\tilde{X}_i^{(2)})_{i=1}^M$:

> When $\tilde{X}_i^{(1)} \in B(y, t_M)$, set $\tilde{X}_i^{(2)} = \tilde{X}_i^{(1)}$.
>
> Otherwise sample $\tilde{X}_i^{(2)}$ from the uniform distribution on $\Delta \setminus B(y, t_M)$.

Conditionally on $\tilde{X}_1^{(1)} = y$, the variable $\tilde{d}_{1,k,1}^{\alpha} g_M(\tilde{d}_{1,k,1})$ (see Equation (3.41) for the definition of g_M) depends only on points in the ball $B(y, t_M)$. Thus

$$E[I(\tilde{d}_{1,k,1} \le t_M)\tilde{d}_{1,k,1}^{\alpha} | \tilde{X}_1^{(1)} = y] = E[I(\tilde{d}_{1,k,2} \le t_M)\tilde{d}_{1,k,2}^{\alpha} | \tilde{X}_1^{(2)} = (s,0)], \quad (3.63)$$

the term in right side was estimated in Lemma 3.14, even though it is important to observe that the cube here is not of unit length (but the side length is bounded from below and above because of assumption (A5)). In fact,

$$\begin{aligned} q(x)^{\alpha/n} \int_0^{t_M} E[I(\tilde{d}_{1,k,2} \le t_M)\tilde{d}_{1,k,2}^{\alpha} | \tilde{X}_1^{(2)} = (s,0)]ds &= V_n^{-\alpha/n} \frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)} t_M \\ + q(x)^{-1/n}(D - V_n^{-\alpha/n-1/n}) \frac{\Gamma(k+\alpha/n+1/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n+1/n)} &+ O(t_M^{2+\alpha}). \end{aligned} \quad (3.64)$$

This result is obtained by employing a change of variables to transform the conditional expectation in (3.64) into a conditional expectation w.r.t. the sample $(q(x)^{1/n}\tilde{X}_i^{(2)})_{i=1}^M$, which is uniform on the unit cube $[0,1]^n$. Finally, we observe that

$$E[I(d_{1,k} \le t_M)d_{1,k}^{\alpha} | X_1 = y]|q(y) - q(x)| \le c_5 t_M^{1+\alpha}$$

with c_5 the Lipschitz constant of q on $\mathcal{X}$, which shows that

$$\begin{aligned} &\int_0^{t_M} E[I(d_{1,k} \le t_M)d_{1,k}^{\alpha} | X_1 = x - sn(x)]q(x - sn(x))ds \\ &\quad = q(x) \int_0^{t_M} E[I(d_{1,k} \le t_M)d_{1,k}^{\alpha} | X_1 = x - sn(x)]ds + O(t_M^{2+\alpha}). \end{aligned}$$

Moreover, the term $q(x)^{1-\alpha/n}t_M$ in Equation (3.64) can be replaced by an integral at the expense of an error term of order $O(t_M^{2+\alpha})$. □

The logarithmic case is again very similar, even if the resulting formula is more complicated.

Lemma 3.18. *Assume that (A2)-(A4) hold with $\gamma \ge 1$ in (A4). Moreover, choose $t_M = M^{-1/n}\log^{2/n} M$ and $x \in \partial\mathcal{X} \setminus N_{c_1 t_M}$, where $c_1 > 1$ is chosen similarly as in Lemma 3.17. Then*

$$\begin{aligned} &\int_0^{t_M} E[I(d_{1,k} \le t_M)\log d_{1,k} | X_1 = x - sn(x)]q(x - sn(x))ds \\ &\quad = C_1 q(x)^{1-1/n} \frac{\Gamma(M)}{\Gamma(M+1/n)} + C_2 \int_0^{t_M} q(x - sn(x))ds \\ &\quad - n^{-1} \int_0^{t_M} q(x - sn(x)) \log q(x - sn(x))ds + R \end{aligned}$$

with

$$|R| \le c_2 M^{-2/n} \log^{3+4/n} M$$

for a constant $c_2 > 0$, which depends only on q, $\mathcal{X}$ and k. The definitions of C_1 and C_2 are found in Lemma 3.15.

Proof. We take x as the origin. The proof of Lemma 3.17 holds for the logarithm almost as such. However, the error terms and final conclusion are a bit different. Let us define the samples $(\tilde{X}_i^{(1)})_{i=1}^M$ and $(\tilde{X}_i^{(2)})_{i=1}^M$ similarly as in the previous proof. We make the preliminary observation that by Assumption (A3), there exist a constant c_1 such that for $t > 0$,

$$\begin{aligned} P(-\log d_{1,k} > t) &\leq P(\omega_{X_1}(d_{1,k}) < c_1 e^{-t}) \\ &\leq 1 - \sum_{j=0}^{k-1} \binom{M-1}{j} c_1 e^{-jt}(1 - c_1 e^{-t})^{M-j-1} = O(Me^{-t}) \end{aligned}$$

and using this it can be shown that for a constant $c_2 > 0$,

$$\begin{aligned} &|E[I(d_{1,k} \leq t_M) \log d_{1,k} | X_1 = y] - E[I(\tilde{d}_{1,k,1} \leq t_M) \log \tilde{d}_{1,k,1} | \tilde{X}_1^{(1)} = y]| \\ &\qquad \leq c_2 M t_M^{n+1} \log M \end{aligned}$$

the only difference to Equation (3.62) being the term $\log M$. Another difference comes from the fact that

$$E[I(d_{1,k} \leq t_M) \log d_{1,k} | X_1 = y] |q(y) - q(x)| = O(t_M \log M)$$

and by Lemma 3.15, analogously to Equation (3.64)

$$\begin{aligned} &\int_0^{t_M} E[I(\tilde{d}_{1,k,2} \leq t_M) \log(q(x)^{1/n} \tilde{d}_{1,k,2}) | \tilde{X}_1^{(2)} = (s,0)] ds \\ &\qquad = C_1 q(x)^{-1/n} \frac{\Gamma(M)}{\Gamma(M + 1/n)} + t_M C_2 + O(t_M^2). \end{aligned}$$

We also need

$$\begin{aligned} &\int_0^{t_M} E[I(\tilde{d}_{1,k,2} \leq t_M) \log \tilde{d}_{1,k,2} | \tilde{X}_1^{(2)} = (s,0)] ds = -t_M n^{-1} \log q(x) \\ &\qquad + \int_0^{t_M} E[I(\tilde{d}_{1,k,2} \leq t_M) \log(q(x)^{1/n} \tilde{d}_{1,k,2}) | \tilde{X}_1^{(2)} = (s,0)] ds. \end{aligned}$$

The terms $q(x) \log q(x)^{1/n} t_M$ and $q(x) t_M$ can be written as line-integrals at the expense of an additional error term of order $O(t_M^2)$. □

3.9 Proofs of the Main Theorems

Proof. (Proof of Theorem 3.3)
Lemma 3.1 implies the bound

$$\begin{aligned} P(d_{1,k} > t_M) &= P(\omega_{X_1}(d_{1,k}) > \omega_{X_1}(t_M)) \\ &\leq M^k (1 - \omega_{X_1}(t_M))^{M-k-1} \leq M^k e^{-c t_M^n M} \end{aligned}$$

for a constant $c > 0$ independent of M and

$$\begin{aligned} E[d_{1,k}^\alpha] &= E[d_{1,k}^\alpha I(d_{1,k} \leq t_M)] + O(M^k e^{-c t_M^n M}) \\ &= E[d_{1,k}^\alpha I(d_{1,k} \leq t_M)] + O(e^{-k \log^2 M}). \end{aligned}$$

The remainder goes to zero faster than any polynomial of M. Furthermore, recalling the notation of Equation (3.41),

$$E[d_{1,k}^{\alpha}I(d_{1,k}\le t_M)] = E[d_{1,k}^{\alpha}g_M] + E[d_{1,k}^{\alpha}I(d_{1,k}\le t_M, X_1\in\partial_{t_M}\mathcal{X})].$$

We apply Lemma 3.2 and Lemma 3.17 to the function

$$f(x) = E[d_{1,k}^{\alpha}I(d_{1,k}\le t_M)|X_1 = x]:$$

$$\begin{aligned}
E[d_{1,k}^{\alpha}I(d_{1,k}\le t_M, X_1\in\partial_{t_M}\mathcal{X})] &= \int_{\partial_{t_M}\mathcal{X}} f(x)dx \\
&= \int_{\partial\mathcal{X}}\int_0^{t_M} E[I(d_{1,k}\le t_M)d_{1,k}^{\alpha}|X_1 = x-sn(x)]q(x-sn(x))dsdS + O(t_M^{\alpha+2}) \\
&= V_n^{-\alpha/n}\frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)}\int_{\partial\mathcal{X}}\int_0^{t_M} q(x-sn(x))^{1-\alpha/n}dsdS \\
&\quad + (D - V_n^{-\alpha/n-1/n})\frac{\Gamma(k+\alpha/n+1/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n+1/n)}\int_{\partial\mathcal{X}} q(x)^{1-\alpha/n-1/n}dS \\
&\quad + O(M^{-2/n-\alpha/n}\log^{2+2\alpha/n+4/n}M) \\
&= V_n^{-\alpha/n}\frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)}\int_{\partial_{t_M}\mathcal{X}} q(x)^{1-\alpha/n}dx \\
&\quad + (D - V_n^{-\alpha/n-1/n})\frac{\Gamma(k+\alpha/n+1/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n+1/n)}\int_{\partial\mathcal{X}} q(x)^{1-\alpha/n-1/n}dS \\
&\quad + O(M^{-2/n-\alpha/n}\log^{2+2\alpha/n+4/n}M).
\end{aligned}$$

We apply Lemma 3.11 as well:

$$\begin{aligned}
E[d_{1,k}^{\alpha}g_M] =& V_n^{-\alpha/n}\frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)}\int_{\mathcal{X}\setminus\partial_{t_M}\mathcal{X}} q(x)^{1-\alpha/n}dx \\
&+ O(M^{-2/n-\alpha/n}\log^{2\alpha/n+4/n}M).
\end{aligned}$$

The proof is finished by summing the two formulas. □

Proof. (Proof of Theorem 3.4)
The proof is analogous to the previous one. Observe that for some constant $c>0$,

$$|E[I(d_{1,k}>t_M)\log d_{1,k}]| \le M^k e^{-ct_M^nM}(c^{-1}+|\log t_M|).$$

We decompose

$$E[I(d_{1,k}\le t_M)\log d_{1,k}] = E[g_M\log d_{1,k}] + E[I(d_{1,k}\le t_M, X_1\in\partial_{t_M}\mathcal{X})\log d_{1,k}]$$

and set

$$f(x) = E[I(d_{1,k}\le t_M)\log d_{1,k}|X_1 = x].$$

By (A3)-(A4), the term $d_{1,k}^n/\omega_{X_1}(d_{1,k})$ is bounded from below and above; consequently $f/\log M$ is a bounded function by Lemma 3.12 independently of M ($\psi(M)$

is of order $\log M$). Thus Lemmas 3.2 and 3.18 can be applied:

$$
\begin{aligned}
&E[I(d_{1,k} \le t_M, X_1 \in \partial_{t_M}\mathcal{X}) \log d_{1,k}] \\
&\quad = \int_{\partial\mathcal{X}} \int_0^{t_M} E[I(d_{1,k} \le t_M) \log d_{1,k} | X_1 = x - sn(x)] q(x - sn(x)) ds dS \\
&\quad + O(t_M^2 \log M) = C_1 \frac{\Gamma(M)}{\Gamma(M+1/n)} \int_{\partial\mathcal{X}} q(x)^{1-1/n} dS + C_2 \int_{\partial_{t_M}\mathcal{X}} q(x) dx \\
&\quad - n^{-1} \int_{\partial_{t_M}\mathcal{X}} q(x) \log q(x) dx + O(M^{-2/n} \log^{3+4/n} M).
\end{aligned}
$$

For the interior term, we refer to Lemma 3.13:

$$
\begin{aligned}
E[g_M \log d_{1,k}] &= n^{-1}\psi(k) - n^{-1}\psi(M) - n^{-1} \log V_n \\
&\quad - n^{-1} \int_{\mathcal{X} \setminus \partial_{t_M}\mathcal{X}} q(x) \log q(x) dx + O(M^{-\gamma/n - \alpha/n} \log^{3+4/n} M).
\end{aligned}
$$

□

Chapter 4

Variance Bounds

In Chapter 3 an asymptotic analysis of the moments $E[d_{1,k}^{\alpha}]$ was performed. Usually the theoretical expectation is not available, and the expectation is estimated by

$$\frac{1}{M}\sum_{i=1}^{M} d_{i,k}^{\alpha}; \tag{4.1}$$

for the logarithmic distance the same role is taken by

$$\frac{1}{M}\sum_{i=1}^{M} \log d_{i,k}. \tag{4.2}$$

For completeness of the analysis, it is of interest to ask, whether the average is close to the expected value. Let us define the sample

$$(Z_i)_{i=1}^{M} = (X_i, Y_i)_{i=1}^{M}$$

with each $Z_i \in \Re^{n+1}$ (and $X_i \in \Re^n$, so $Y_i \in \Re$ is taken as a scalar) and the variables

$$h(Z_i, Z_{N[i,1]}, \ldots, Z_{N[i,k]}),$$

where the nearest neighbors are calculated in the sample $(X_i)_{i=1}^{M}$ as in Chapter 3 with the function h bounded. Then the average

$$\frac{1}{M}\sum_{i=1}^{M} h(Z_i, Z_{N[i,1]}, \ldots, Z_{N[i,k]}) \tag{4.3}$$

appropriately generalizes Equation (4.1). Taking (4.3) as the quantity of interest, we derive two theorems that apply to (4.1) and (4.2) bounding the variance of the two averages.

To understand what is to be expected from a variance bound, take $(W_i)_{i=1}^{M}$ as independent random variables and choose a measurable function g. Then it is simple to compute

$$\mathrm{Var}[\sum_{i=1}^{M} g(W_i)] = \sum_{i=1}^{M} \mathrm{Var}[g(W_i)], \tag{4.4}$$

which is linear w.r.t. M under the i.i.d. assumption. This easily provable additivity of variance is a fundamental result in probability theory with far reaching impact on our understanding of science in general. However, the terms in the sum (4.3) are not independent with each other due to the dependency on nearest neighbors. Nevertheless Equation (4.4) generalizes as the inequality

$$\mathrm{Var}[\sum_{i=1}^{M} h(Z_i, Z_{N[i,1]}, \ldots, Z_{N[i,k]})] \leq cM \tag{4.5}$$

for some constant c independent of M ([2, 71] and [49]). While the reader can follow the references to find out that many results from random geometry apply straightforwardly with nearest neighbors as a special case, the results are of asymptotic nature and they give the constant c only for large M. The more concrete approach taken in [16] has the advantage of being optimized for nearest neighbor graphs with finite M instead of going to the limit $M \to \infty$. Moreover, the proofs require only a bounded fourth moment whereas the moment conditions encountered in random geometry are much more restrictive.

To avoid ties, the following assumption is used:

A5) $(Z_i)_{i=1}^M = (X_i, Y_i)_{i=1}^M$ is a sample of independent vectors with the variables $(X_i)_{i=1}^M$ having their realizations in a set $\mathcal{X}$ for which Assumption (A1) in Chapter 2 holds. Moreover, for all $i, j, l > 0$ with $j \neq l$ and $j \neq i$,

$$P(\|X_i - X_j\| \neq \|X_i - X_l\|) = 1.$$

The distance on $\mathcal{X}$ (i.e. ρ) is set as the Euclidean metric.

Motivated by the work in [16], it is shown that

Theorem 4.1. *(Law of Large Numbers for Nearest Neighbor Statistics) Suppose that (A5) holds. Let $h(z_1, z_{1,1}, \ldots, z_{1,k})$ be a measurable function (taking values in $\Re$) and define the random variables*

$$h_i = h(Z_i, Z_{N[i,1]}, \ldots, Z_{N[i,k]}),$$

where the nearest neighbors are calculated in the sample $(X_i)_{i=1}^M$. Assume that each h_i is bounded by some constant $\|h\|_\infty$. Then for $M > k$,

$$\mathrm{Var}[\sum_{i=1}^{M} h_i] \leq \|h\|_\infty^2 (1 + 2kL(n))(1 + 2k)M.$$

The constant $L(n)$ is defined in Section 4.2.

In comparison to [16], the bound is tighter w.r.t. k, but depends on the bound on h_i. Due to the latter deficit, the application to (4.1) and (4.2) is not trivial, but if the probability of large nearest neighbor distances is small, it is possible to extend the proof of Theorem 4.1 straightforwardly into

Corollary 4.1. *Suppose that (A3)-(A5) hold and $|h_i| \leq d_{i,k}^{\alpha}$ for some $\alpha > 0$. Then we have the bound*

$$\mathrm{Var}[\sum_{i=1}^{M} h_i] \leq (1+2kL(n))(1+2k)M^{1-2\alpha/n}\log^{4\alpha/n} M + O(M^{-\beta}),$$

where $\beta > 0$ is an arbitrary positive number fixed a priori. As another special case, we also have

$$\mathrm{Var}[\sum_{i=1}^{M} \log d_{i,k}] \leq 4(1+2kL(n))(1+2k)M\log^{4} M + O(M^{-\beta}).$$

In words, because each term $E[d_{i,k}^{\alpha}]$ tends to be of order $M^{-\alpha/n}$, the average (4.1) is indeed close to the expectation for large M the fluctuations being of order $M^{-1/2}\log^{2/n} M$ times the magnitude of the terms (measured by the standard deviation).

If $0 < 2\alpha \leq n$, a more practical inequality than that in Theorem 4.1 is possible by using the theory in Chapter 2 as in the following bound, the logarithmic factor does not appear. This result is of some independent interest, because it does not require the existence of a fourth moment as in [16], but exploits the fact that a sum of the distances $d_{i,k}^{\alpha}$ is bounded deterministically.

Theorem 4.2. *Suppose that (A5) holds and $\mathcal{X} \subset [0,1]^n$. Furthermore, we assume that for any distinct indices $(j, j_1, \ldots, j_k)$,*

$$\sqrt{E[h(Z_j, Z_{j_1}, \ldots, Z_{j_k})^2 | (X_i)_{i=1}^{M}]} \leq \|X_j - X_{j_k}\|^{\alpha} \quad a.s.$$

for some constant $\alpha > 0$. Then if $0 \leq 2\alpha \leq n$, we have the bound

$$\mathrm{Var}[\sum_{i=1}^{M} h_i] \leq 2^{3+2\alpha}(kL(n)+1)(k+1)^{1+2\alpha/n} n^{\alpha} M^{1-2\alpha/n}.$$

4.1 The Efron-Stein Inequality

The Efron-Stein inequality is a special case of a concentration inequality (see [57]) the purpose of which are to generalize the variance formula for the empirical mean of independent random variables to more general functions. The result can be stated for any independent sample $(Z_i)_{i=1}^{M}$ as

Theorem 4.3. *Let $(Z_i)_{i=1}^{M}$ be a set of independent random variables and*

$$f(Z_1, \ldots, Z_M)$$

a function of this sample. Moreover, for $1 \leq i \leq M$, let

$$f_i(Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_M)$$

be a measurable function of $(Z_i)_{i=1}^M$ *excluding* Z_i. *Then*

$$\mathrm{Var}[f] \leq \frac{1}{M}\sum_{i=1}^{M} E[(f-f_i)^2].$$

Proof. The proof can be found e.g. in [57]. □

Intuitively, Theorem 4.3 applies when f is stable in the sense of approximate invariance with respect to small perturbations.

4.2 How Many Points Can Share the Same Nearest Neighbor?

Related to Theorem 4.3, it is important to verify the stability of nearest neighbor graphs with respect to removal of samples. Fortunately, such an analysis is well-known in the literature on computational geometry and the proof in [47] is repeated here.

The set of indices corresponding to points in $(X_i)_{i=1}^M$ that have X_i among their k nearest neighbors is defined by

$$K_{i,k} = \{1 \leq j \leq M : \ j \neq i, N[j,l] = i \text{ for some } 1 \leq l \leq k\}.$$

In this section, our main goal is to bound the cardinality $|K_{i,k}|$.

(A5) imposes the condition $p = 2$ on the norm, which is not necessary ($p > 1$ works just as well, see [16]), but sufficient for the purposes of the thesis.

For a unit vector e, the cone of degree 30° is defined by

$$C(e) = \{x \in \Re^n : \ x^T e \geq \|x\| \cos 30^\circ\}. \tag{4.6}$$

The constant $L(n)$ is defined as the smallest positive integer with

$$\Re^n = \cup_{i=1}^{L(n)} C(e_i)$$

for some unit vectors $e_1, \ldots, e_{L(n)}$. In words, $L(n)$ bounds the number of cones of degree 30° needed to cover the space $\Re^n$. For example, $L(1) = 2$ and $L(2) = 6$ (see Figures 4.1(a) and 4.1(b)).

Theorem 4.4. *Suppose that (A5) holds. Then for any* $k > 0$, $1 \leq i \leq M$ *and* $1 \leq k < M$, *the cardinality* $|K_{i,k}|$ *is bounded by* $kL(n)$.

Proof. Fix the point X_j and a vector e with $\|e\| = 1$. The non-centered cone is defined by

$$C_{X_j}(e) = \{x \in \Re^n : \ (x - X_j)^T e \geq \cos 30^\circ \|x - X_j\|\}.$$

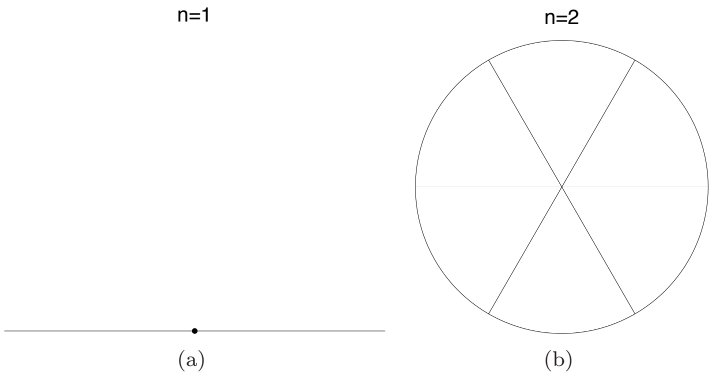


Figure 4.1: The plane can be divided into six cones (4.6), and the real axis consists of two cones. The angle between the lines on the plane is 60°.

Notice that for $z, y \in C_{X_j}(e)$, we have the geometrically intuitive bound

$$(z - X_j)^T(y - X_j) \geq \frac{1}{2}\|z - X_j\|\|y - X_j\|. \tag{4.7}$$

Let us now make the counterassumption that there exists $k+1$ points $(X_{j_i})_{i=1}^{k+1} \subset C_{X_j}(e)$ with X_j among their k nearest neighbors. By (A2), it holds that $\|X_{j_{k+1}} - X_j\| > \|X_{j_k} - X_j\| > \ldots > \|X_{j_1} - X_j\|$. Then inequality (4.7) implies that for any $1 \leq i \leq k$,

$$\begin{aligned}\|X_{j_{k+1}} - X_{j_i}\|^2 &\leq \|X_{j_{k+1}} - X_j\|^2 + \|X_{j_i} - X_j\|^2 - \|X_{j_{k+1}} - X_j\|\|X_{j_i} - X_j\| \\ &< \|X_{j_{k+1}} - X_j\|^2\end{aligned}$$

the last inequality being strict. Thus we may conclude that X_j cannot be among the k nearest neighbors of the point $X_{j_{k+1}}$ leading to a contradiction.

To finish the proof one should notice that the space $\Re^n$ can be covered with $L(n)$ cones of degree 60°. Each point in the sample falls into one of these cones and it follows that X_j can be among the k nearest neighbors of at most $kL(n)$ points. □

Theorem 4.4 implies a concrete bound on $|K_{i,k}|$ with an exponential growth w.r.t. n. To see this, we note that each of the 2^n quadrants of $\Re^n$ is a cone of degree 90° verifying that $L(n) \geq 2^n$. Let $e_1, \ldots, e_{L(n)}$ be unit vectors such that the sets $C(e_i)$ cover $\Re^n$. For e_1 and e_2 we have $e_1^T e_2 \leq \cos 60°$ and

$$\|e_1 - e_2\|^2 \geq 2 - e_1^T e_2 \geq 2 - 2\cos 60° = 1.$$

It follows that the balls $B(e_i, 1/2)$ are disjoint and also subsets of the ball $B(0, 2)$. A volume argument then implies that

$$L(n) \leq 4^n.$$

A much sharper bound is in fact possible, because the balls $B(e_i, 1/2)$ are disjoint and touch $B(0, 1/2)$ rendering the analysis into an examination of kissing numbers [7]. Without going deeply into the matter, $L(n)$ is bounded by the kissing number of the space $\Re^n$ for which upper bounds for dimensions $n = 1, \ldots, 24$ have been tabulated. This also shows that a closed-form formula for $L(n)$ is probably not possible. It should be mentioned that kissing numbers were encountered also in [13].

On the other hand, it is interesting that on expectation $|K_{i,k}|$ tends to be close to k as indicated by the following theorem.

Theorem 4.5. *Let $(a_i)_{i=1}^M$ be a sequence of numbers. Then for any $0 < k \leq M$, the equality*

$$\sum_{i=1}^{M} \sum_{j \in K_{i,k}} a_j = k \sum_{i=1}^{M} a_i$$

holds.

Proof. Define the sets

$$C_{i,l} = \{1 \leq j \leq M : \ j \neq i, N[j,l] = i\}.$$

Then

$$\sum_{i=1}^{M} \sum_{j \in K_{i,l}} a_j = \sum_{l=1}^{k} \sum_{i=1}^{M} \sum_{j \in C_{i,l}} a_j.$$

Moreover,

$$C_{i_1,l} \cap C_{i_2,l} = \emptyset$$

when $i_1 \neq i_2$ (a point has only one l-th nearest neighbor) and

$$\cup_{1 \leq i \leq M} C_{i,l} = \{1, \ldots, M\} \qquad \text{(each point has a } l\text{-th nearest neighbor).}$$

Thus

$$\sum_{i=1}^{M} \sum_{j \in C_{i,l}} a_j = \sum_{i=1}^{M} a_i$$

and the claim follows. □

4.3 Proofs of the Variance Bounds

Proof. (Proof of Theorem 4.1) The argument is based on the use of the Efron-Stein inequality. Let us construct a new sample $(\tilde{Z}_i)_{1 \leq i \leq M, i \neq l}$ by removing a variable Z_l and define $\tilde{h}_i^{(l)}$ by calculating $h(\tilde{Z}_i, \tilde{Z}_{\tilde{N}[i,1]}, \ldots, \tilde{Z}_{\tilde{N}[i,k]})$ on this modified set of variables. Then in terms of indicator functions

$$|h_i - \tilde{h}_i^{(l)}| \leq 2\|h\|_\infty I(i \in K_{l,k})$$

when $i \neq l$ and by Theorem 4.4,

$$\begin{aligned} (\sum_{i=1}^{M} h_i - \sum_{i=1, i \neq l}^{M} \tilde{h}_i^{(l)})^2 &\leq \|h\|_\infty^2 (1 + 2|K_{l,k}|)^2 \\ &\leq \|h\|_\infty^2 (1 + 2kL(n))(1 + 2|K_{l,k}|). \end{aligned}$$

Using Theorem 4.5 we obtain the final result

$$\sum_{l=1}^{M}(\sum_{i=1}^{M} h_i - \sum_{i=1, i\neq l}^{M} \tilde{h}_i^{(l)})^2 \leq \|h\|_\infty^2 (1+2kL(n))(1+2k)M.$$

□

Proof. (Proof of Corollary 4.1) By Lemma 3.1,

$$P(d_{i,k} > M^{-1/n}\log^{2/n} M) \leq e^{\log M - c_1 \log^2 M}$$

for some constant $c_1 > 0$ and (A3)-(A4) imply that

$$\text{Var}[\sum_{i=1}^{M} h_i] \leq \text{Var}[\sum_{i=1}^{M} h_i I(d_{i,k} \leq M^{-1/n}\log^{2/n} M)] + c_2 M^2 e^{c_3 \log M - c_3 \log^2 M}$$

for some c_2 and c_3 independent of M. Theorem 4.1 applies directly to the first term in the right side.

For the logarithmic distance, it holds for some constant $c_4 > 0$ that

$$P(|\log d_{i,k}| > \log^2 M) = P(d_{i,k} < e^{-\log^2 M}) \leq c_4 M^{k - \log M}$$

by Theorem 3.5 and Assumption (A5). Again, this expression approaches zero faster than any polynomial w.r.t. M and the proof is completed similarly as in the previous case. □

Proof. (Proof of Theorem 4.2) Let us again construct a new sample $(\tilde{Z}_i)_{1\leq i\leq M, i\neq l}$ by removing a variable Z_l. Then by our assumptions we have $E[h_i^2|(X_i)_{i=1}^M] \leq d_{i,k}^{2\alpha}$ and $E[(\tilde{h}_i^{(l)})^2|(X_i)_{i=1}^M] \leq d_{i,k+1}^{2\alpha}$. An application of the formula $(\sum_{i=1}^M a_i)^2 \leq M\sum_{i=1}^n a_i^2$ and the fact that $(c+d)^2 \leq 3c^2 + 3d^2$ gives

$$\begin{aligned}
&E[(\sum_{i=1}^{M} h_i - \sum_{1\leq i\leq M, i\neq l} \tilde{h}_i^{(l)})^2|(X_i)_{i=1}^M] \\
&\leq (\sum_{j\in K_{l,k}\cup\{l\}} d_{j,k}^{\alpha} + \sum_{j\in K_{l,k}} d_{j,k+1}^{\alpha})^2 \\
&\leq 3(|K_{l,k}|+1)\sum_{j\in K_{l,k}\cup\{l\}} d_{j,k}^{2\alpha} + 3(|K_{l,k}|+1)\sum_{j\in K_{l,k}} d_{j,k+1}^{2\alpha} \\
&\leq 3(kL(n)+1)\sum_{j\in K_{l,k}\cup\{l\}} d_{j,k}^{2\alpha} + 3(kL(n)+1)\sum_{j\in K_{l,k}} d_{j,k+1}^{2\alpha}.
\end{aligned}$$

Thus by Corollary 2.1 and Theorem 4.5,

$$\begin{aligned}
&\sum_{l=1}^{M} E[(\sum_{i=1}^{M} h_i - \sum_{1\leq i\leq M, i\neq l} \tilde{h}_i^{(l)})^2|(X_i)_{i=1}^M] \\
&\leq 3(kL(n)+1)(k+1)(\sum_{i=1}^{M} d_{i,k}^{2\alpha} + \sum_{i=1}^{M} d_{i,k+1}^{2\alpha}) \\
&\leq 2^{3+2\alpha}(kL(n)+1)(k+1)^{1+2\alpha/n} n^{\alpha} M^{1-2\alpha/n}
\end{aligned}$$

and the Efron-Stein inequality finishes the proof. □

Chapter 5

Entropy Estimation

5.1 Introduction

In Chapters 2-4 general theoretical principles related to random local functions and especially nearest neighbor distances were discussed. In the remaining two chapters, two important applications in statistical estimation are discussed. In fact, the idea for boundary corrected expansions in Chapter 3 arose from the need for improved entropy estimators that would better address the curse of dimensionality.

Given a random variable X, it is natural to ask 'How random is X?'. One of the milestones in statistics of the previous century was the development of information theory ([58]), which defines a theoretically well-founded measure of randomness. In general, based on intuitive axioms, the correct way to measure randomness of X is the quantity

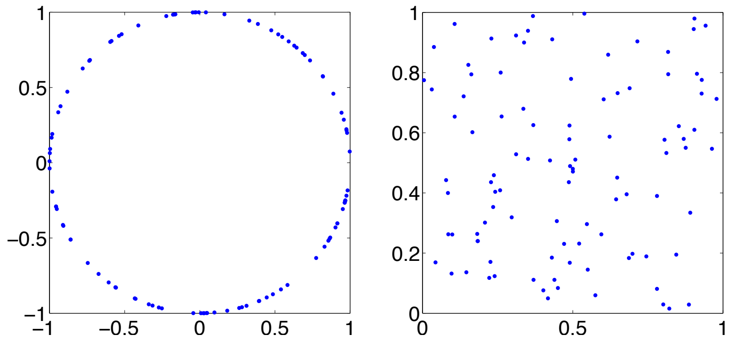
$$-\int_{\mathcal{X}} q(x) \log q(x) dx,$$

where q is the density of X. This is not surprising as it has been known for a long time that the logarithmic entropy measures disorder in equilibrium statistical mechanics. On the other hand, it was shown in [54], that under slightly weaker assumptions, one ends up with a whole family of entropies:

$$\frac{1}{1-\beta} \log \int_{\mathcal{X}} q(x)^{\beta} dx,$$

where $\beta > 0$ (we assume $0 < \beta < 1$). The logarithmic entropy then arises in the limit $\beta \to 1$. While the other entropies are theoretically less satisfying, they are still useful to measure randomness in many applications even if not necessarily optimal. Especially the case $\beta = 0.5$ is useful due to its relation to the Hellinger distance (see e.g. [70]) between q and the uniform distribution.

In this chapter, the task of computing the entropy of X using an i.i.d. sample $(X_i)_{i=1}^M$ is discussed. This is an important problem as in practice q is usually not known in closed form whereas a sample of realizations is available.



(a) (b)

Figure 5.1: The randomness of the random variables taking values in the unit circle is small compared to a uniform case, which also shows in the nearest neighbor distances.

In general, a random variable with a low entropy tends to concentrate in a small portion of the space implying that the nearest neighbor distances should also be small. This is demonstrated in Figure 5.1, where clearly the random variables in Figure 5.1(a) are much less random being located in a one dimensional submanifold and the corresponding nearest neighbor distances are small as well. In terms of information, the points on the circle can be explained with one variable and thus contain less information. Somewhat surprisingly, the connection between randomness and nearest neighbor distances is exact in the asymptotic limit of infinite sample size; to be precise, we recall from Chapter 3 that

$$M^{\alpha/n}E[d_{1,k}^{\alpha}] \to V_n^{-\alpha/n}\frac{\Gamma(k+\alpha/n)}{\Gamma(k)}\int_{\mathcal{X}} q(x)^{1-\alpha/n}dx$$

in the limit $M \to \infty$.

In this chapter, it is shown that taking the boundary effect into account helps to understand the validity of the relation for finite sample sizes. Moreover, it is shown how estimation accuracy can be improved by a weighted estimator. The simulations at the end of the chapter demonstrate the improved accuracy for the bias corrected estimators. However, it also turns out that the logarithmic differential entropy estimator has problems with estimation variance leaving room for further research.

5.2 The Estimation Problem

Suppose that Assumption (A4) holds and the sample $(X_i)_{i=1}^M$ is i.i.d. with a common density q. Then as stated previously, Rényi entropies are defined by the formula

$$H_\beta[q] = \frac{1}{1-\beta}\log(\int_{\mathcal{X}} q(x)^\beta dx) \tag{5.1}$$

for $\beta \geq 0$; the celebrated differential entropy is the special case $\beta \to 1$:

$$-\int_{\mathcal{X}} q(x) \log q(x) dx. \tag{5.2}$$

At first sight, the task of estimating these quantities seems challenging, as q is not known and should somehow be inferred from data. To this end, one can of course consider building a density estimate $\hat{q}$ of q and using the approximation in place of q to estimate the integral (5.1) or (5.2) arriving at a widely investigated branch of research (references include [11, 23, 27]).

While entropy estimation by performing an intermediate step of density estimation is an approach that comes easily to mind, one quickly gets the impression that the curse of dimensionality and free parameters tend to be problems to these methods [3]. To address this issue, one can adopt a sophisticated density estimation scheme (e.g. local linear estimators as in [33]) or look for other techniques. From a theoretical point of view introducing local linearity is attractive; on the other hand, it is likely that practical problems with robustness arise.

Interestingly, if one is interested only in entropy and not in the density q itself, the intermediate density estimation step is not necessary. In fact, from the results in random geometry (e.g. [47, 48, 50]), one finds quicky that the total length of many random graphs is related to the Rényi entropy (5.1). This deep connection has been exploited e.g. in [8] and [26] to derive robust graph theoretic entropy estimators. The goal of this chapter is to analyze and improve the estimators in the special case of nearest neighbor distances.

5.3 Rényi Entropy and Nearest Neighbors

We start from the limit in Chapter 3:

$$V_n^{\alpha/n} \frac{\Gamma(k)}{\Gamma(k+\alpha/n)} M^{\alpha/n} E[d_{1,k}^{\alpha}] \to \int_{\mathcal{X}} q(x)^{1-\alpha/n} dx \quad \text{(when } M \to \infty) \tag{5.3}$$

for a fixed $\alpha > 0$. The connection to the Rényi entropy (5.1) is obvious. As mentioned earlier, such an asymptotic relation between the two things is not just a property of nearest neighbor distances, but a similar limit arises for many other functionals in random geometry even if the constant in front of $\int_{\mathcal{X}} q(x)^{1-\alpha/n} dx$ varies. The practical advantages of using nearest neighbor distances include most importantly simplicity and understandability.

The simplest (even though not the only) way to use Equation (5.3) is formally

$$H_{1-\alpha/n}[q] \approx \hat{H}_{1,1-\alpha/n}[q] = \frac{n}{\alpha} \log\left(\frac{V_n^{\alpha/n}\Gamma(k)}{\Gamma(k+\alpha/n)} M^{\alpha/n-1} \sum_{i=1}^{M} d_{i,k}^{\alpha}\right). \tag{5.4}$$

In this estimator, one may choose $k = 1$, which usually works well. One observes that if $\beta > 1$ in (5.1), then it must be that $\alpha < 0$. While this choice is fully possible (see [34]), the restriction $0 < \beta < 1$ is imposed here.

Instead of this estimator, to stay consistent with Theorem 3.3, we use

$$\frac{n}{\alpha}\log(V_n^{\alpha/n}\frac{\Gamma(M+\alpha/n)\Gamma(k)}{\Gamma(M)\Gamma(k+\alpha/n)}M^{-1}\sum_{i=1}^{M}d_{i,k}^{\alpha}) \tag{5.5}$$

as using $\Gamma(M+\alpha/n)/\Gamma(M)$ instead of $M^{\alpha/n}$ is a minor choice due to the approximation

$$\Gamma(M+\alpha/n)/\Gamma(M) = M^{\alpha/n} + M^{\alpha/n-1}$$

of Lemma 2.4. In [34], the asymptotic consistency of the estimator was proven in a general setting. Here, our goal is not to analyze such proofs, but rather to show an inherent problem of the method: it suffers rather badly from the curse of dimensionality due to its reliance on distances. Theorem 3.3 characterizes the bias in the following way:

Theorem 5.1. *Suppose that (A2)-(A4) hold with $1 \leq \gamma \leq 2$ in (A4). Then for a fixed k,*

$$\begin{aligned} &V_n^{\alpha/n}\frac{\Gamma(k)\Gamma(M+\alpha/n)}{\Gamma(k+\alpha/n)\Gamma(M)}M^{-1+1/n}\sum_{i=1}^{M}d_{i,k}^{\alpha} - M^{1/n}\int_{\mathcal{X}} q(x)^{1-\alpha/n}dx \\ &\quad\to \frac{\Gamma(k+\alpha/n+1/n)}{\Gamma(k+\alpha/n)}(D - V_n^{-\alpha/n-1/n})\int_{\partial\mathcal{X}} q(x)^{1-\alpha/n-1/n}dS \end{aligned}$$

when $M \to \infty$. The constant D was defined in Theorem 3.3. As a consequence,

$$\liminf_{M>0} M^{1/n}|E[\hat{H}_{1,1-\alpha/n}[q]] - \frac{n}{\alpha}\log\int_{\mathcal{X}} q(x)^{1-\alpha/n}dx| > 0.$$

Proof. The first claim follows from Theorem 3.3. For the second one, we can use the approximation

$$\log a - \log b = \log(1+\frac{a-b}{b}) = \frac{a-b}{b} + O(\frac{(a-b)^2}{b^2}).$$

□

The error expansion is unique compared to earlier work in the sense that it complete characterizes the error up to the first order. The result is in agreement with the rate of convergence in [18]. The bias depends on the surface integral

$$\int_{\partial\mathcal{X}} q(x)^{1-\alpha/n-1/n}dS,$$

which is non-zero in the presence of boundaries (see Example 3.2). This motivates the conclusion that convergence is slow already for $n = 3$. If we could derive an estimator that has a bias that goes to zero faster than $M^{-1/n}$ with respect to M, then at least for large M such an estimator would improve accuracy and hopefully significantly alleviate the curse of dimensionality. This is done in next section.

5.3.1 Boundary Corrected Estimators

There seem to be various slightly different possibilities for addressing the problem at the boundary. The proposal here, introduced by us in [39], is to estimate

$$\int_{\mathcal{X}} q(x)^{1-\alpha/n} dx$$

by using a linear combination of the quantities

$$\delta_{M,l,\alpha} = \frac{1}{M} \sum_{i=1}^{M} d_{i,l}^{\alpha}$$

for different values of l. Fixing $\alpha > 0$ and $k \geq 2$, the weights $(w_l)_{l=1}^{k}$ are chosen in such a way that

$$\sum_{l=1}^{k} w_l \frac{\Gamma(l+\alpha/n)}{\Gamma(l)} = V_n^{\alpha/n} \frac{\Gamma(M+\alpha/n)}{\Gamma(M)} \tag{5.6}$$

and

$$\sum_{l=1}^{k} w_l \frac{\Gamma(l+\alpha/n+1/n)}{\Gamma(l)} = 0. \tag{5.7}$$

Such a choice is always possible when $k \geq 2$ and it depends only on α, k, n and M. If $k > 2$, the solution is not unique and the convention of choosing the sequence with the smallest possible l^2-norm is adopted. With these choices, the boundary corrected estimator of $H_{1-\alpha/n}[q]$ is

$$\hat{H}_{2,1-\alpha/n}[q] = \frac{n}{\alpha} \log(\sum_{l=1}^{k} w_l \delta_{M,l,\alpha}). \tag{5.8}$$

Example 5.1. *As an example, if* $k = 5$,$n = 3$ *and* $\alpha = 1$, *then inside the logarithm in* (5.8) *we would have*

$$1.65\delta_{M,1,1} + 1.12\delta_{M,2,1} + 0.41\delta_{M,3,1} - 0.34\delta_{M,4,1} - 1.10\delta_{M,5,1}.$$

The weights are the l^2-norm minimizing solution of Equations (5.6) and (5.7), which in this case read as

$$\sum_{l=1}^{5} w_l \frac{\Gamma(l+1/3)}{\Gamma(l)} = \left(\frac{4}{3\pi}\right)^{1/3} \frac{\Gamma(M+1/3)}{\Gamma(M)}$$

$$\sum_{l=1}^{5} w_l \frac{\Gamma(l+2/3)}{\Gamma(l)} = 0.$$

The benefit of weighting is best clarified by

Theorem 5.2. *If (A2)-(A4) hold with* $1 \leq \gamma \leq 2$ *in (A4), then for a fixed* $k \geq 2$

$$\sum_{l=1}^{k} w_l E[\delta_{M,l,\alpha}] - \int_{\mathcal{X}} q(x)^{1-\alpha/n} dx = O(M^{-\gamma/n} \log^{2+2\alpha/n+4/n} M).$$

Proof. In the order of magnitude notation, $|w_l| = O(M^{\alpha/n})$. Theorem 3.3 implies that

$$\begin{aligned} w_l E[d_{1,l}^{\alpha}] &= w_l V_n^{-\alpha/n} \frac{\Gamma(l+\alpha/n)\Gamma(M)}{\Gamma(l)\Gamma(M+\alpha/n)} \int_{\mathcal{X}} q(x)^{1-\alpha/n} dx \\ &+ w_l(D - V_n^{-\alpha/n-1/n}) \frac{\Gamma(l+\alpha/n+1/n)\Gamma(M)}{\Gamma(l)\Gamma(M+\alpha/n+1/n)} \int_{\partial\mathcal{X}} q(x)^{1-\alpha/n-1/n} dS \\ &+ O(M^{-\gamma/n-\alpha/n} \log^{2+2\alpha/n+4/n} M). \end{aligned}$$

When the sum over l is taken, the second term in the approximation vanishes, whereas the first one becomes

$$\int_{\mathcal{X}} q(x)^{1-\alpha/n} dx.$$

□

The idea behind the boundary corrected estimator is now visible. The weights $(w_l)_{l=1}^{k}$ ensure the disappearance of the error due to the boundaries, which improves the rate of convergence by a factor of $M^{-1/n}$. From the general point of view, it is remarkable that a weighting scheme obtains a cancellation of error; such a principle could possibly be applied in other estimation problems as well.

In general, the improvement in terms of bias comes with a more or less increased variance. The following theorem shows that if $n > 4$, then we can say at least that standard deviation tends to approach zero faster than bias when $M \to \infty$. On the other hand, for dimensions smaller than 5, standard deviation is in principle the dominant part for large M after the boundary correction, even though in practice probabilistic deviations tend not to cause problems in that regime. Recall that the constant $L(n)$ was defined in Section 4.2 and grows exponentially with respect to n.

Theorem 5.3. *If the assumptions of Theorem 4.2 hold with $n \geq 2$, then for a fixed $k \geq 2$ and $0 \leq \alpha \leq n/2$,*

$$\mathrm{Var}[\sum_{l=1}^{k} w_l \delta_{M,l,\alpha}] \leq 2^{3+2\alpha} (\sum_{l=1}^{k} w_l^2)(kL(n)+1)(k+1)^{1+2\alpha/n} n^{\alpha} M^{-1-2\alpha/n}.$$

Proof. By the Cauchy-Schwarz inequality [56],

$$|\sum_{l=1}^{k} w_l d_{i,l}^{\alpha}| \leq \sqrt{\sum_{l=1}^{k} w_l^2} \sqrt{\sum_{l=1}^{k} d_{i,l}^{2\alpha}} \leq d_{i,k}^{\alpha} \sqrt{k \sum_{l=1}^{k} w_l^2}.$$

Theorem 4.2 can be applied straightforwardly:

$$\mathrm{Var}[\sum_{l=1}^{k} w_l \delta_{M,l,\alpha}] \leq 2^{3+2\alpha} (\sum_{l=1}^{k} w_l^2)(kL(n)+1)(k+1)^{2+2\alpha/n} n^{\alpha} M^{-1-2\alpha/n}.$$

□

The restriction $0 < \alpha \leq n/2$ was avoided in Theorem 4.2 under (A3) and (A4); that result could have been applied as well.

The variance bound in Theorem 5.3 is not tight in terms of constants, but the dependency on $\sum_{l=1}^{k} w_l^2$ suggests the intuitive fact, that large weights should be avoided. Finding a tight asymptotic expansion for variance is a topic of future research that could lead to a more optimal choice for the weights $(w_l)_{l=1}^{k}$.

5.4 Differential Entropy

The differential entropy is more widely used than other Rényi entropies as it often arises naturally from information theoretic considerations. For this reason, nonparametric estimation of differential entropy has been widely investigated. The related derivations differ from those in the previous two sections only in technical details.

The state-of-the-art nearest neighbor estimator is analogous to $\hat{H}_{1,1-\alpha/n}$:

$$-\int_{\mathcal{X}} q(x)\log q(x)dx \approx \log V_n - \psi(1) + \psi(M) + \frac{n}{M}\sum_{i=1}^{M}\log d_{i,1}, \tag{5.9}$$

where ψ is the digamma function (extension to $k > 1$ is trivial). The main reference on this topic is [30]; the multivariate estimator has gained popularity ever since probably due to its simplicity (no free parameters) and robustness. The method appears often in the literature on the estimation of mutual information (see e.g. [31] and [15]).

Analogously to previous section, we can say the following about the bias.

Theorem 5.4. *Suppose that Assumptions (A2)-(A4) hold with $1 \leq \gamma \leq 2$ in (A4). Then for a fixed k,*

$$M^{1/n}|\frac{n}{M}E[\sum_{i=1}^{M}\log d_{i,k}] - nC_2(M,k) + \int q(x)\log q(x)dx| \to \infty,$$

when $M \to \infty$.

Proof. This is a direct consequence of Theorem 3.4, because $\psi(M)$ grows to infinity in the limit $M \to \infty$. □

Even though the bias goes infinity in the limit $M \to \infty$, such convergence is slow due to the fact that ψ grows approximately logarithmically. In any case, the most important thing is that the rate of convergence of the estimator (5.9) cannot exceed $M^{-1/n}$ in the presence of a boundary cut-off. Analogously to the Rényi entropy case, we choose the weights $(w_l)_{l=1}^k$ in such a way that

$$\sum_{l=1}^{k} w_l C_1(M,l) = 0 \tag{5.10}$$

$$\sum_{l=1}^{k} w_l = n. \tag{5.11}$$

Again the weights can be found if $k > 2$.

The following straightforward application of Theorem 3.4 demonstrates the boundary correction.

Theorem 5.5. *Suppose that (A2)-(A4) hold with $1 \leq \gamma \leq 2$ in (A4). If the weights $(w_l)_{l=1}^k$ are chosen according to Equations (5.10)-(5.11),*

$$\begin{aligned}\sum_{l=1}^{k} w_l E[\log d_{1,l}] - \sum_{l=1}^{k} w_l C_2(M,l) = &- \int_{\mathcal{X}} q(x)\log q(x)dx \\ &+ O(M^{-\gamma/n} \log^{3+4/n} M).\end{aligned}$$

Consequently, the proposed estimator takes the form

$$-\int_{\mathcal{X}} q(x)\log q(x)dx \approx \frac{1}{M}\sum_{l=1}^{k}\sum_{i=1}^{M} w_l \log d_{i,l} - \sum_{l=1}^{k} w_l C_2(M,l). \tag{5.12}$$

To the best of our knowledge, no corresponding work exists in the literature on nearest neighbor entropy estimation with the exception [39]. A similar variance bound as in Theorem 5.3 could be proven here; one simply replaces $d_{i,k}^{\alpha}$ with $\log d_{i,k}$.

From the practical point of view, instead of (5.10) it is easiest to choose $(w_l)_{l=1}^k$ to satisfy

$$\sum_{l=1}^{k} w_l \psi(l + 1/n) = 0 \tag{5.13}$$

$$\sum_{l=1}^{k} w_l \frac{\Gamma(l+1/n)}{\Gamma(l)} = 0, \tag{5.14}$$

which automatically implies that $\sum_{l=1}^{k} w_l C_1(M,l) = 0$. This way, the computation of D_1 and D_2 in Theorem 3.4 is avoided, even though finding them is mainly a technical matter at least in lower dimensions.

Example 5.2. *As an example, if $n = 3$ and $k = 13$, then in Equation (5.12) we would use approximately*

$$\begin{aligned}(w_1, \ldots, w_{13}) = (&-2.28, 0.63, 1.53, 1.73, 1.61, 1.33, \\ &0.94, 0.50, 0.025, -0.48, -0.99, -1.51, -2.05).\end{aligned}$$

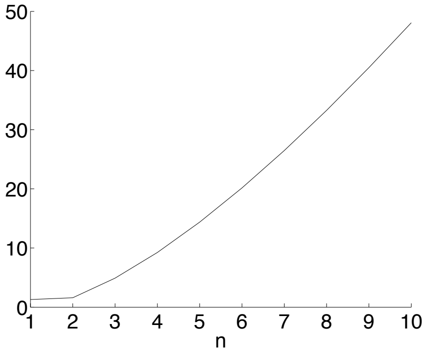


Figure 5.2: The norm $\sqrt{\sum_{l=1}^{n+10} w_l^2}$, where the weights are calculated according to Equations (5.13)-(5.14).

The weights are the solution of Equations (5.13) and (5.14), which now take the form

$$\sum_{l=1}^{13} w_l \psi(l+1/3) = 0 \tag{5.15}$$

$$\sum_{l=1}^{13} w_l \frac{\Gamma(l+1/3)}{\Gamma(l)} = 0 \tag{5.16}$$

under the constraint $\sum_{l=1}^{13} w_l = 3$. *Again, taking the solution with the smallest* l^2*-norm provides an easy way to choose among the possible solutions.*

Unfortunately, the boundary correction has a drawback in this case. Namely, if one computes the weights according to Equations (5.11) and (5.13)-(5.14), it becomes quickly evident that the norm $\sqrt{\sum_{l=1}^{k} w_l^2}$ grows fast. This is demonstrated in Figure 5.2 for the choice $k = n + 10$. On the other hand, the experimental section shows that things are better than expected in practice, especially for a large sample size. The same problem arises for the other Rényi entropy estimators as well even though the experimental results seem more favorable to them. On a deeper level, the growth of the weights is a manifestation of the curse of dimensionality.

5.5 Simulations

5.5.1 Rényi Entropy

The simulations concern the estimation of

$$n \log \int_{\mathcal{X}} q(x)^{1-1/n} dx$$

corresponding to $\beta = 1 - 1/n$ in Equation (5.1). The dimensionality n is varied from 3 to 6 and we choose the weights $(w_l)_{l=1}^{k}$ to fulfil Equations (5.6) and (5.7).

The choice $k = n + 1$ is made even if it is likely a suboptimal one. The estimator of Equation (5.5) is compared to the boundary corrected statistic (5.8).

In the first set of simulations, the methods are tested with uniformly distributed points in the cube $[0, 1]^n$ letting M increase up to 9000 in steps of 300. As can be easily verified, the real value is zero:

$$H_{1-1/n}(q) = 0.$$

The uniform distribution can be viewed as the distribution with maximal randomness in the unit cube. The results of the simulations are drawn in Figure 5.3. As a measure of accuracy, the absolute deviation

$$E[|H_{1-1/n}(q) - \hat{H}_{i,1-1/n}(q)|]$$

is computed by averaging over 1000 realizations ($i = 1, 2$).

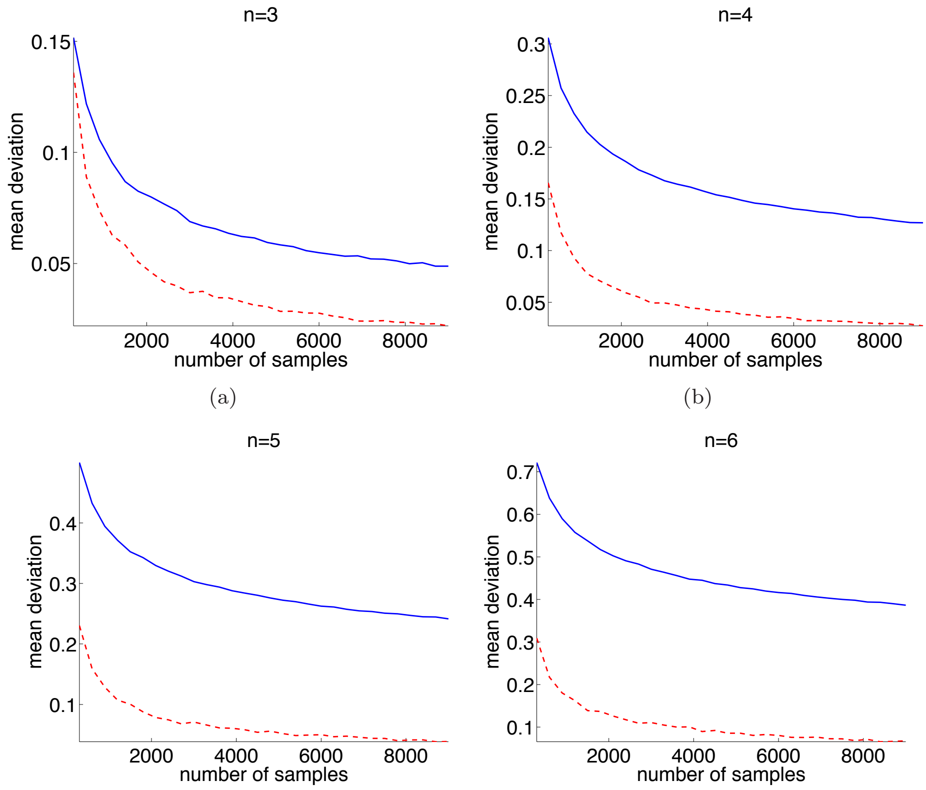
Figure 5.3 shows that here the proposed method is more accurate for all sample sizes, especially in terms of percentages. From Figure 5.4 we see that while standard deviation is larger, the increase is not too high and the trade-off is favourable to the bias corrected estimator.

As a second experiment, to assess performance with a more complicated distribution, the simulations were repeated with the truncated Gaussian distribution on the unit ball. In this case, the variables $(X_i)_{i=1}^M$ are samples from the multivariate normal distribution restricted to the unit ball. The experiment involves both boundaries and correlation between components. The results are drawn in Figure 5.5. Again, the real value was computed in closed form and the estimates were compared to it. The results are rather similar to the uniform case in the sense that the standard deviation in Figure 5.6 did increase, but not too much.

5.5.2 Differential Entropy

The experiments for the differential entropy were done in the same way as for the Rényi entropy. The results with uniform data are drawn in Figures 5.7 and 5.8, whereas those for truncated Gaussian data are shown in Figures 5.9 and 5.10. Instead of letting k vary with n, the choice $k = 30$ was used all the time.

From Figures 5.7 and 5.9, it is seen that the boundary correction works well for dimensions $n = 3$ and $n = 4$, whereas after that problems appear due to the weight increase problem. It can be concluded that the boundary correction for the differential entropy is mainly useful in applications with a large number of samples available (e.g. information compression and related topics). From Figures 5.8 and 5.10 it is seen that the problem is due to the increased variance brought by the weighting. In Chapter 7 we suggest a possible improvement.



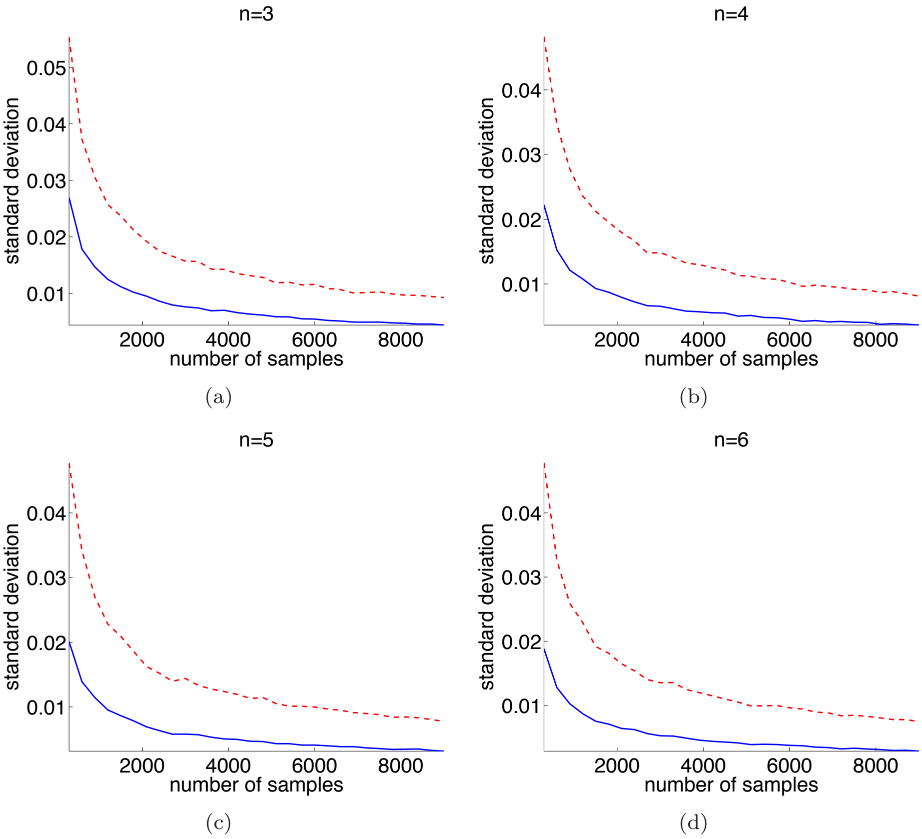
(a)

(b)

(c)

(d)

Figure 5.3: Rényi entropy: Mean absolute deviation for the experiment with uniform data. The dashed line corresponds to the boundary corrected estimator.



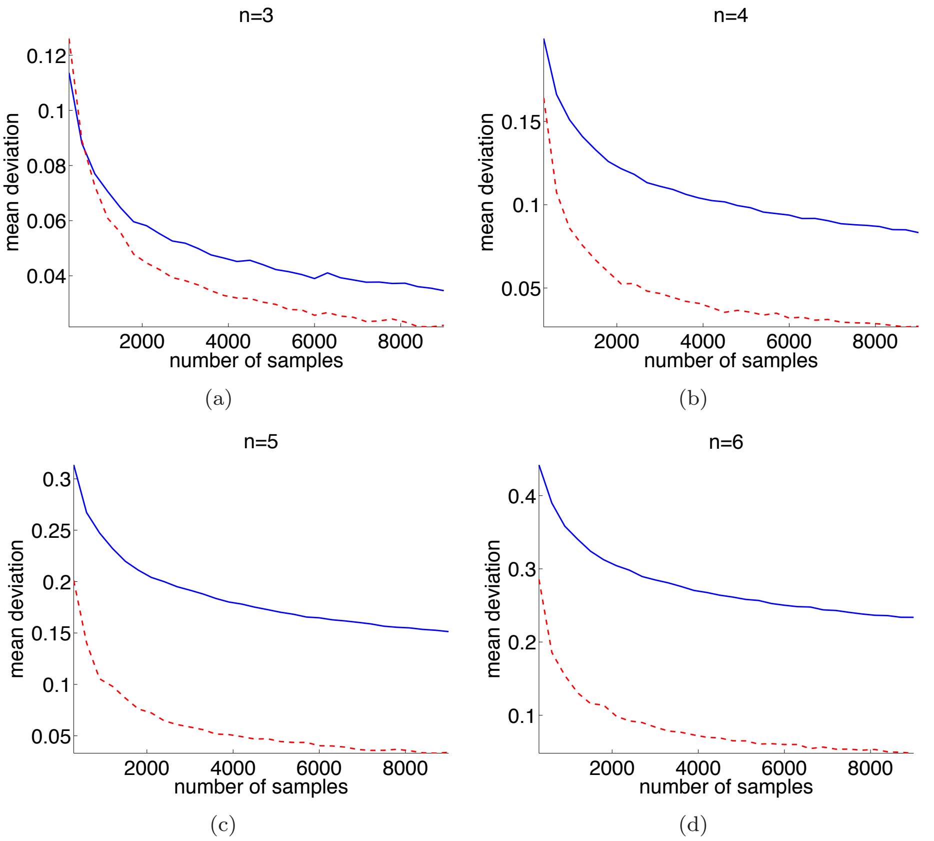
(a)

(b)

(c)

(d)

Figure 5.4: Rényi entropy: Standard deviation for the experiment with uniform data. The dashed line corresponds to the boundary corrected estimator.



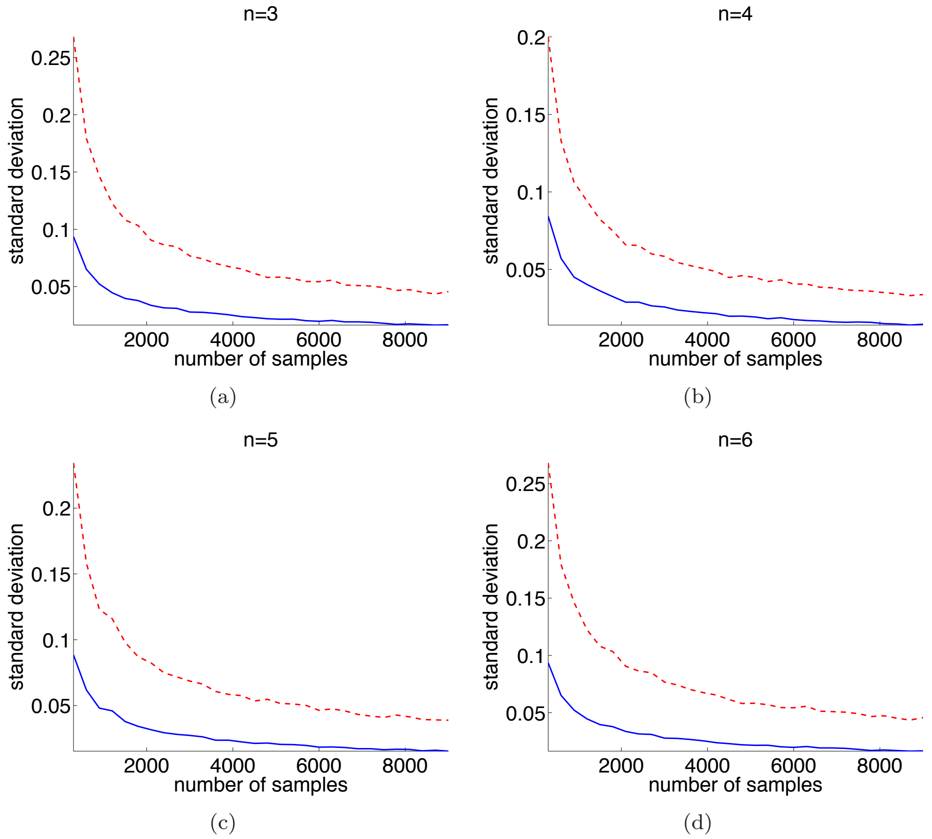
(a)

(b)

(c)

(d)

Figure 5.5: Rényi entropy: Mean absolute deviation for the experiment with truncated Gaussian data. The dashed line corresponds to the boundary corrected estimator.



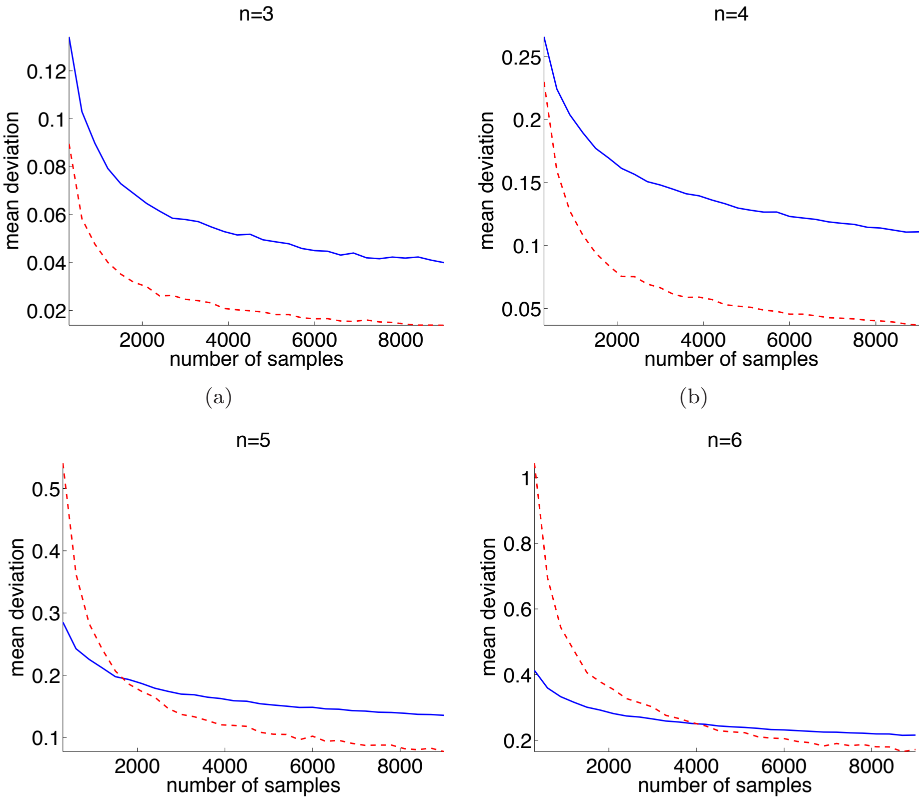
(a)

(b)

(c)

(d)

Figure 5.6: Rényi entropy: Standard deviation for the experiment with truncated Gaussian data. The dashed line corresponds to the boundary corrected estimator.



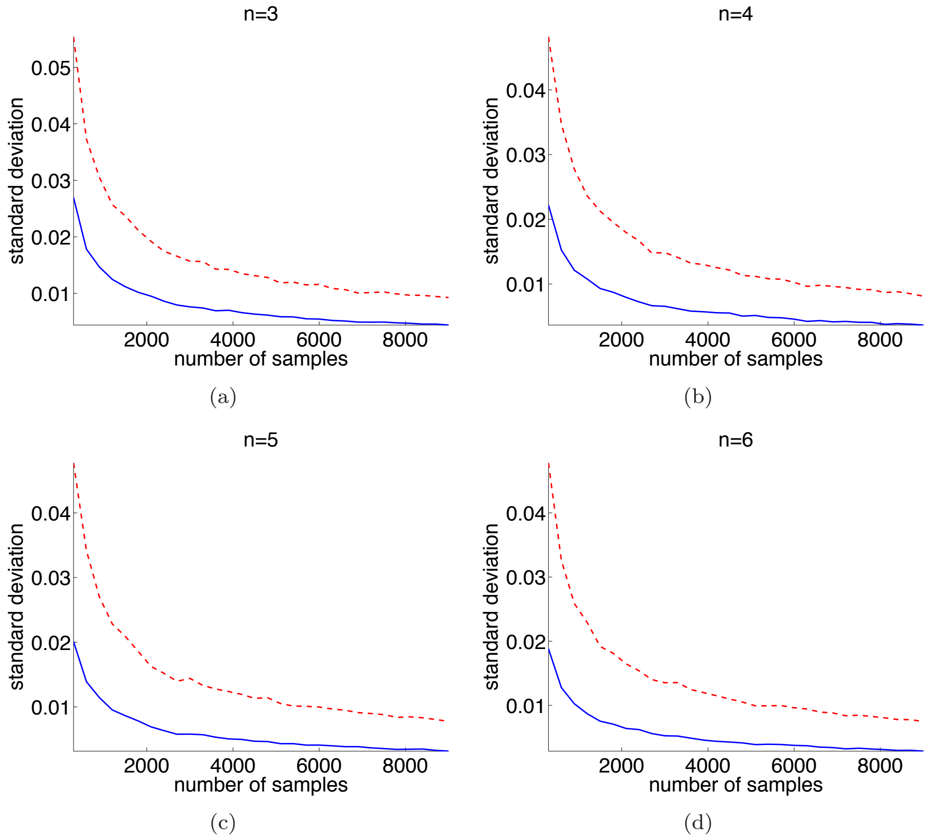
(a)

(b)

(c)

(d)

Figure 5.7: Differential entropy: Mean absolute deviation for the experiment with uniform data. The dashed line corresponds to the boundary corrected estimator.



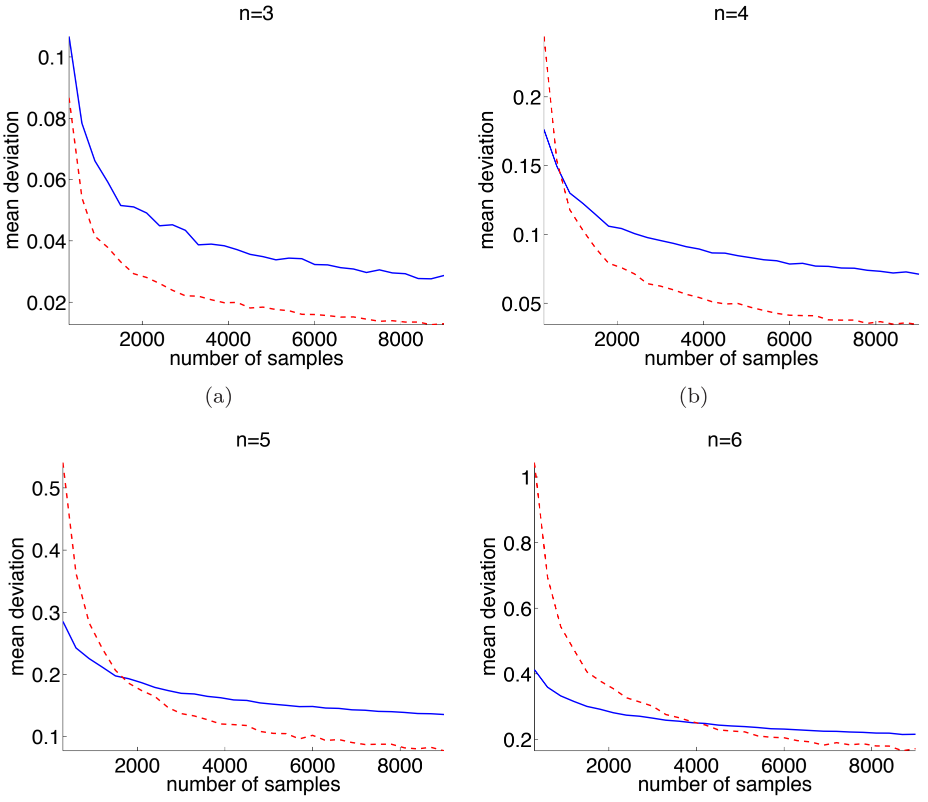
(a)

(b)

(c)

(d)

Figure 5.8: Differential entropy: Standard deviation for the experiment with uniform data. The dashed line corresponds to the boundary corrected estimator.



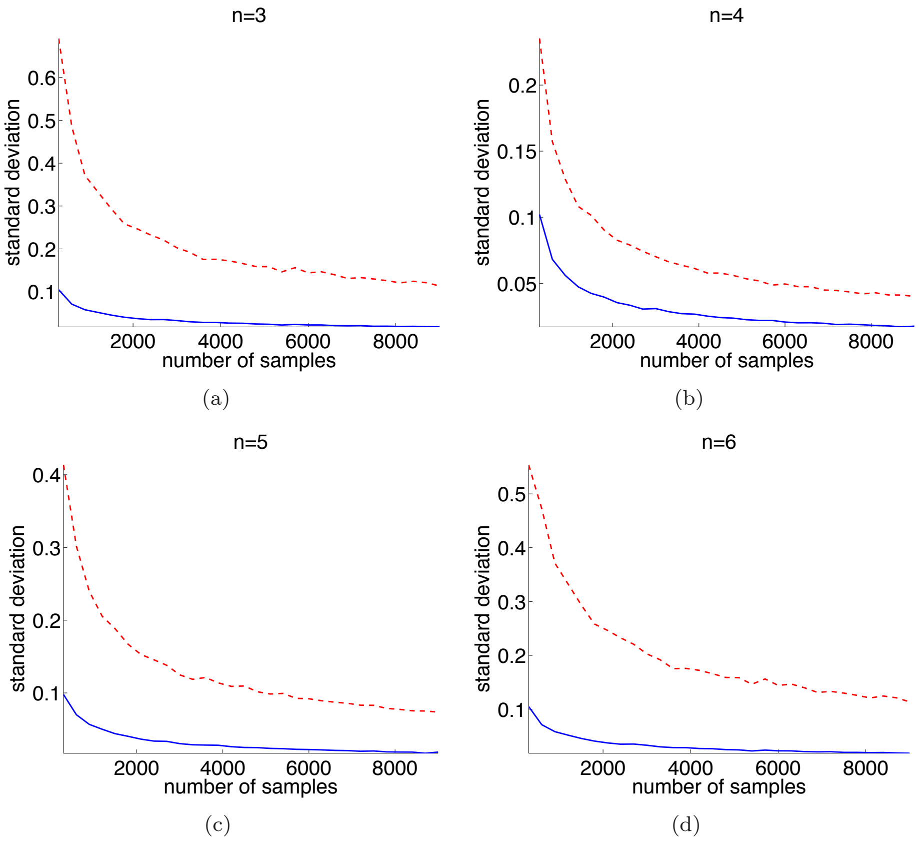
(a)

(b)

(c)

(d)

Figure 5.9: Differential entropy: Mean absolute deviation for the experiment with truncated Gaussian data. The dashed line corresponds to the boundary corrected estimator.



(a)

(b)

(c)

(d)

Figure 5.10: Differential entropy: Standard deviation for the experiment with truncated Gaussian data. The dashed line corresponds to the boundary corrected estimator.

Chapter 6

Residual Variance Estimation

6.1 A Review of the Problem

While literature on the topic of residual variance estimation is rather scarce, in the statistics community the topic is better known as noise variance estimation and quite a lot of research has been done under that name. However, noise variance estimation usually involves the assumption that the data $(X_i, Y_i)_{i=1}^M$ consists of i.i.d. random variables generated by

$$Y = m(X) + r, \tag{6.1}$$

where r is zero-mean noise independent of X. Then the residual (noise) variance is $V = E[r^2]$. When speaking of residual variance estimation, the independence assumption on r is not made and instead the definitions $m(X) = E[Y|X]$ and $r = Y - E[Y|X]$ are imposed. Formally Equation (6.1) still holds, but in general r may not be independent of X.

The most straightforward approach to residual variance estimation is to build an approximation $\hat{m}$ to the optimal function m and then estimate

$$V \approx \hat{V} = E[(Y - \hat{m}(X))^2] = E[(m(X) - \hat{m}(X))^2] + V.$$

The drawback of this type of an approach is clear: the first step involves building a regression estimate to m, which is to be avoided. In fact, the classical results in [62] imply that for any regression estimator, there exists a sequence (m_M) such that each function in the sequence has the same Lipschitz constant and

$$\liminf_{M\to\infty} M^{2/(2+n)} E[(m_M(X) - \hat{m}_M(X))^2] > 0,$$

where $\hat{m}_M$ is the approximation of m_M with M samples. In words: the optimal rate of convergence achievable for Lipschitz continuous functions is $M^{-2/(2+n)}$ unless prior information is available. Such a rate is not satisfying as in [10] it was shown that in residual variance estimation, error of order $M^{-1/2}$ is achievable for $n \leq 4$.

A better idea is to estimate V directly without the intermediate step of approximating m. As an example, for $n = 1$ difference based methods are known to obtain low biases ([68, 55, 20]). Other (excluding nearest neighbor based methods) direct approaches include the use of U-statistics ([46]) and in general various kernel estimators ([24]).

Unfortunately, most of the direct methods have not been shown to stay consistent if the independence of r from X is dropped. Once the homoscedasticity assumption is removed, possibly the most succesful branch of research has been estimators based on the use of nearest neighbors, which is the focus of this chapter.

The residual variance estimation problem should not be confused with estimating the whole variance function (see e.g. [5, 45] and [6])

$$g(x) = E[(Y - E[Y|X])^2|X = x],$$

which is a considerably more difficult task and not necessary in many applications. Of course, methods for the estimation of g yield V as a special case, but this is an overly complicated approach. Moreover, multivariate variance estimation with random design is a relatively unexplored topic as most work concentrates on univariate models or a fixed design limiting the scope of the methods.

6.2 Nearest Neighbor Approaches

Possibly the simplest nearest neighbor residual variance estimator is given by the 1-NN estimator

$$V \approx \frac{1}{2M} \sum_{i=1}^{M} (Y_i - Y_{N[i,1]})^2, \tag{6.2}$$

where $N[i, 1]$ is computed in the sample $(X_i)_{i=1}^M$. An early reference on the method is [61]. Later on, the 1-NN estimator has been found to be useful especially in machine learning and data-analysis ([1, 12, 37]) for tasks such as input and model structure selection.

To clarify the logic behind the estimator, let us assume that the sample $(X_i, Y_i)_{i=1}^M$ is generated by the model (6.1) with independent noise r. Now it is reasonable to assume that the points X_i and $X_{N[i,1]}$ are close to each other when the number of samples is large enough and we may approximate heuristically

$$V \approx \frac{1}{2M} \sum_{i=1}^{M} (r_i - r_{N[i,1]})^2.$$

Using the assumption that the variables $(r_i)_{i=1}^M$ are independent of the input variables $(X_i)_{i=1}^M$ and each other, it holds that $E[r_i r_{N[i,1]}] = 0$ and one may furthermore write

$$E[V] \approx \frac{1}{2M} \sum_{i=1}^{M} E[r_i^2] + \frac{1}{2M} \sum_{i=1}^{M} E[r_{N[i,1]}^2] = E[r^2],$$

which is the residual variance.

Based on these considerations, clearly it is possible to prove that the estimator (6.2) is consistent when the output noise r is independent of X. Of course, the rate of convergence may be slow and various generalizations have been proposed. Of these, possibly the most important is the Gamma test ([13, 61]) and the local linear estimator in [60].

A natural question is, if consistency of (6.2) and its extensions hold also in a more general setting. While currently no proof exists, it seems likely that convergence indeed holds for heteroscedastic noise, because the distribution of the random variable $X_{N[i,1]}$ approaches that of X_i in the limit $M \to \infty$. Nevertheless it is not guaranteed that such convergence is fast, which motivated us to look for alternatives to the 1-NN method.

One extension is to allow $k_M > 1$ by

$$V \approx \frac{1}{(1+k_M^{-1})M} \sum_{i=1}^{M} (Y_i - \frac{1}{k_M} \sum_{j=1}^{k_M} Y_{N[i,j]})^2, \tag{6.3}$$

where the assumption $k_M/M \to 0$ as $M \to \infty$ is essential for consistency. However, even though it is possible to show that the approximation (6.3) is able to give consistent estimates under general assumptions (a large body of research exists on k-NN estimators, see e.g. [9, 28]), it is not without problems. One practical difficulty is the choice of k_M, which is not obvious, as for example cross-validation inevitably increases variance.

It is of course possible to approximate m with a local polynomial or a neural network model instead of a simple locally constant approximator. But that would probably not be a particularly elegant solution and would not utilize the possibilities brought by increased complexity in a particularly good way. An alternative solution based on modified nearest neighbor graphs was introduced in [10]. The method there has rather similar properties to those of (6.2), but is not affected by heteroscedasticity.

In this chapter we analyze a slightly simpler method based on modifying the approximator (6.2) as

$$\hat{V}_M = \frac{1}{M} \sum_{i=1}^{M} (Y_i - Y_{N[i,1]})(Y_i - Y_{N[i,2]}). \tag{6.4}$$

The idea appeared first time in [14] and was later analyzed in [17, 41, 36]. Moreover, a generalization to higher order moments was provided in [17]. To understand the logic behind the method, assume that the function m is continuous. Then the heuristic approximation $m(X_{N[i,k]}) - m(X_i) \approx 0$ and conditional independence yield

$$V \approx \frac{1}{M} \sum_{i=1}^{M} E[(r_i - r_{N[i,1]})(r_i - r_{N[i,2]})] = E[\frac{1}{M} \sum_{i=1}^{M} r_i^2], \tag{6.5}$$

which is the residual variance. Moreover, it can be seen that the quality of the estimate depends only on the smoothness of m and therefore the estimator does not have problems with heteroscedastic noise. This is the first benefit of using a product of differences: the problematic terms $r_{N[i,1]}$ and $r_{N[i,2]}$ vanish from the expectation.

Even though the ability to solve the residual variance estimation problem in a general context was a major motivating factor behind the product estimator, in this regard it does not bring any particular advantage over the solution in [10]. However, as the main result of the chapter it is shown that (6.4) achieves an improvement in the rate of convergence. So in effect, while the method can be motivated as a solution for heteroscedastic noise, it actually has unexpected benefits over the 1-NN method without significant disadvantages.

6.3 Analysis of The Product Estimator

Our analysis is divided into two parts: first worst-case bounds are demonstrated under relatively weak assumptions leading to a similar rate of convergence as in [17]. The proof technique is a typical application of the theory in Chapter 3 and similar derivations work in other fields of non-parametric statistics as well; the geometric nearest neighbor bounds lead to rather simple proofs. Secondly, we take a look on what happens when sufficient regularity is present.

6.3.1 General Bounds on the Bias

Often in nonparametric statistics, estimation bias tends to be large, whereas variance is less of a problem. The most straightforward way to bound the systematic error of the estimator (6.4) is to invoke the Hölder continuity of m and the bounds derived in Chapter 3. The argument based on the continuity of m is common in non-parametric statistics and a good starting point here.

Theorem 6.1. *Suppose that $(X_i, Y_i)_{i=1}^M$ is i.i.d. with each variable X_i taking values in a space $\mathcal{X}$; moreover, it is assumed that the assumptions of Theorem 2.1 hold. Moreover, we assume that $E[Y_1^2] < \infty$ and that the conditional expectation*

$$m(x) = E[Y_1|X_1 = x]$$

belongs to $H(c_m, \gamma_m)$ with $0 < \gamma_m \leq 1$ (the Hölder continuity in Definition 3.1 generalizes trivially to any metric space). Then for $n = 2\gamma_m$ and $M \geq kC_n$, the inequality

$$|E[\hat{V}_M] - V| \leq \frac{2C_n c_m^2}{M}(2 + \log(\frac{M}{2})) \tag{6.6}$$

holds and for $n > 2\gamma_m$,

$$|E[\hat{V}_M] - V| \leq \frac{c_m^2 n}{n - 2\gamma_m}(\frac{2C_n}{M})^{2\gamma_m/n} - \frac{4\gamma_m c_m^2 C_n^{2\gamma_m/n}}{(n - 2\gamma_m)M} + \frac{2c_m^2 C_n^{2\gamma_m/n}}{M}. \tag{6.7}$$

The constant C_n was defined in Theorem 2.1.

Proof. Notice that

$$\begin{aligned} E[(Y_i - Y_{N[i,1]})(Y_i - Y_{N[i,2]})] &= E[(m(X_i) - m(X_{N[i,1]}))(m(X_i) - m(X_{N[i,2]}))] \\ &\quad + E[(r_i - r_{N[i,1]})(m(X_i) - m(X_{N[i,2]}))] + E[(r_i - r_{N[i,1]})(r_i - r_{N[i,2]})] \\ &\quad + E[(m(X_i) - m(X_{N[i,1]}))(r_i - r_{N[i,2]})]. \end{aligned} \tag{6.8}$$

Because $r_i = Y_i - E[Y_i|X_i]$, it holds by the basic properties of conditional expectations that

$$E[r_i r_{N[i,2]}] = E[r_{N[i,1]} r_{N[i,2]}] = E[r_{N[i,1]} r_i] = 0.$$

For example,

$$\begin{aligned} &E[r_{N[i,1]} r_{N[i,2]}] = E[E[r_{N[i,1]} r_{N[i,2]} | (X_i)_{i=1}^M]] \\ &= E[\sum_{j_1=1}^M \sum_{j_2=1}^M E[r_{j_1} r_{j_2} | (X_i)_{i=1}^M] I(N[i,1] = j_1) I(N[i,2] = j_2)] \\ &= E[\sum_{j_1=1}^M \sum_{j_2=1}^M E[r_{j_1} | (X_i)_{i=1}^M] E[r_{j_2} | (X_i)_{i=1}^M] I(N[i,1] = j_1) I(N[i,2] = j_2)] = 0. \end{aligned}$$

In the two last equalities we used the fact that the variables $(r_i)_{i=1}^M$ are conditionally independent and mean zero given $(X_i)_{i=1}^M$ due to the assumption that $(X_i, Y_i)_{i=1}^M$ is i.i.d. Observe also that the term with $j_1 = j_2$ is neglected as $I(N[i,1] = j_1) I(N[i,2] = j_1)$ is always zero.

Similarly

$$E[(r_i - r_{N[i,1]})(m(X_i) - m(X_{N[i,2]})) + (m(X_i) - m(X_{N[i,1]}))(r_i - r_{N[i,2]})]$$

is equal to zero. On the other hand,

$$|\sum_{i=1}^M (m(X_i) - m(X_{N[i,1]}))(m(X_i) - m(X_{N[i,2]}))| \leq \sum_{i=1}^M c_m^2 d_{i,2}^{2\gamma_m},$$

which can be bounded by Theorem 2.1. What remains of Equation (6.8) after averaging over i is the residual variance. □

The rate of convergence depends on the instrinsic dimensionality of the set $\mathcal{X}$. Avoiding the multiplier $\log M$ when $n = 2\gamma_m$ seems difficult under the assumptions that were made. A slightly tighter and more concrete bound on the bias is possible if the packing dimension of $\mathcal{X}$ is n with $\mathcal{X} \subset \Re^n$:

Theorem 6.2. *Suppose that $(X_i, Y_i)_{i=1}^M$ is i.i.d. and $\mathcal{X} \subset [0,1]^n$. Moreover, we assume that $m \in H(c_m, \gamma_m)$ with $0 < \gamma_m \leq 1$. Then for $n \geq 2$, the inequality*

$$|E[\hat{V}_M] - V| \leq c_m^2 (\frac{2^{n+1} n^{n/2}}{M})^{2\gamma_m/n} \tag{6.9}$$

holds. If $n \geq 3$ and the assumptions of Theorem 2.2 hold, then

$$|E[\hat{V}_M] - V| \leq (\frac{2^{n+1} \lambda(\mathcal{X})}{V_n M})^{2\gamma_m/n} + o(M^{-2\gamma_m/n}), \tag{6.10}$$

where the remainder $o(M^{-2\gamma_m/n})$ approaches zero faster than $M^{-2\gamma_m/n}$.

Proof. The proof is similar to that of Theorem 6.1 with the difference that Corollary 2.1 and Theorem 2.2 are invoked instead of Theorem 2.1. □

Interestingly, the rate $M^{-2\gamma_m/n}$ is essentially the same as that obtained in [10] and it also agrees with [17]. It is known to be the best possible for the 1-NN estimator of Equation (6.2) with homoscedastic noise as has been demonstrated for $\gamma_m = 1$:

Remark 6.1. *Assume that the sample $(X_i)_{i=1}^M$ is i.i.d. and uniformly distributed on the unit cube $[0,1]^n$. Moreover, the variables $(Y_i)_{i=1}^M$ are assumed to be linearly related to the inputs:*

$$Y_i = w^T X_i$$

for some vector $w \in \Re^n$. Then using the lower bound of Section 2.4, it was shown in [35] that

$$E[(Y_i - Y_{N[i,1]})^2] \geq \frac{1}{n}\|w\|^2 V_n^{-2/n} \frac{\Gamma(M)}{\Gamma(M+2/n)}$$

showing that by using the squared difference, $M^{-2/n}$ is the best rate of convergence that can be achieved.

6.3.2 The Bias Under Sufficient Regularity

As the main contribution of this chapter, it is shown next that worst-case considerations give a wrong view of the practical rate of convergence.

Lemma 6.1. *Suppose that (A2'), (A3) and (A4) hold with $0 \leq \gamma \leq 1$ in (A4). Let $H : \Re^n \to \Re^{n\times n}$ be a matrix valued function with $\|H(x)\| \leq 1$ for all $x \in \mathcal{X}$. Then for $M > 2k$ and fixed $j_2 > j_1$,*

$$|E[(X_{N[1,j_1]} - X_1)^T H(X_1)(X_{N[1,j_2]} - X_1)]| \leq cM^{-2/n-\gamma/n} \log^{2+6/n} M,$$

where c is a constant independent of M.

Proof. Define $h = (X_{N[1,j_1]} - X_1)^T H(X_1)(X_{N[1,j_2]} - X_1)$. We set

$$t_M = M^{-1/n} \log^{2/n} M.$$

The expectation decomposes into three parts (recall g_M from Equation (3.41)):

$$\begin{aligned} E[h] = E[g_M(X_1, d_{1,k})h] + E[I(X_1 \in \partial_{t_M}\mathcal{X}, d_{1,k} \leq t_M)h] \\ + E[I(d_{1,k} > t_M)h]. \end{aligned} \tag{6.11}$$

By Lemma 3.1,

$$|E[I(d_{1,k} > t_M)h]| \leq c^{-1} M^k e^{-ct_M^n M}$$

for a constant $c > 0$ independent of M. On the other hand, by Assumption (A2')

$$\lambda(\partial_{t_M}\mathcal{X}) = O(t_M)$$

and

$$|E[I(X_1 \in \partial_{t_M}\mathcal{X}, d_{1,k} \leq t_M)h]| = O(t_M^3).$$

The first term in the right side of (6.11) is the most important factor in the decomposition.

Choose $x \in \mathcal{X} \setminus \partial_{t_M}\mathcal{X}$; by definition $B(x, t_M) \subset \mathcal{X}$. Let us define the density $\tilde{q}_x(y)$ by setting

$$\tilde{q}_x(y) = q(x)$$

when $y \in B(x, t_M)$ and

$$\tilde{q}_x(y) = \frac{1 - q(x)\lambda(B(x, t_M))}{1 - P(X_1 \in B(x, t_M))} q(y)$$

when $y \notin B(x, t_M)$. Then by assumption (A4),

$$\int_{\mathcal{X}} |q(y) - \tilde{q}_x(y)| dy \leq c t_M^{n+\gamma}$$

for an appropriate constant c. By Lemma 3.16 we find an i.i.d. sample $(\tilde{X}_i)_{i=2}^M$ such that $P((\tilde{X}_i)_{i=2}^M \neq (X_i)_{i=2}^M) \leq cMt_M^{n+\gamma}$. Moreover, we choose $\tilde{X}_1 = X_1$ (independent of $(\tilde{X}_i)_{i=2}^M$) and denote by C the event $(\tilde{X}_i)_{i=2}^M = (X_i)_{i=2}^M$.

The new sample $(\tilde{X}_i)_{i=2}^M$ is uniform around x, whereas it differs from the original one only with a small probability. Let us now denote by $\tilde{d}_{1,k}$ the k-th nearest neighbor distance in the new sample; $\tilde{h}$ is defined in a corresponding way. Then,

$$E[I(C) g_M(X_1, d_{1,k}) h | X_1 = x] = E[I(C) g_M(\tilde{X}_1, \tilde{d}_{1,k}) \tilde{h} | \tilde{X}_1 = x]. \tag{6.12}$$

Moreover,

$$|E[I(C^C) g_M(X_1, d_{1,k}) h | X_1 = x]| \leq t_M^2 P(C^C) \leq cMt_M^{n+\gamma+2}, \tag{6.13}$$

and the same holds of course also with respect to $(\tilde{X}_i)_{i=1}^M$.

Thus, the sample $(\tilde{X})_{i=1}^M$ gives nearly the same expectation as $(X_i)_{i=1}^M$ and an application of Theorem 3.5 yields

$$\begin{aligned} E[g_M(\tilde{X}_1, \tilde{d}_{1,k})\tilde{h} | \tilde{X}_1 = x] &= k! \binom{M-1}{k} \int_{S_{x,k}(t_M)} (1 - \tilde{\omega}_x(\|x_{1,k} - x_1\|))^{M-k-1} \times \\ &\times (x_{1,1} - x)^T H(x)(x_{1,2} - x) \prod_{i=1}^k \tilde{q}_x(x_{1,i}) dx_{1,1}, \ldots, dx_{1,k}. \end{aligned} \tag{6.14}$$

But Equation (6.14) is zero because $\tilde{q}_x$ is locally constant, whereas h is antisymmetric w.r.t. the replacement of $X_{N[1,1]} - X_1$ with $-X_{N[1,1]} + X_1$. □

Theorem 6.3. *In addition to $m \in H(c_m, \gamma_m)$ $(1 < \gamma_m \leq 2)$ and $E[Y_1^2] < \infty$, assume that the inputs $(X_i)_{i=1}^M$ are i.i.d. with a common density q satisfying (A2'), (A3) and (A4) with $0 \leq \gamma \leq 1$ in (A4). Then for some constant c (depending on $\mathcal{X}$, m and the density q),*

$$|E[\hat{V}_M] - V| \leq cM^{-1/n - \gamma_m/n} + cM^{-2/n-\gamma/n} \log^{2+6/n} M.$$

Proof. Recall that of the terms in Equation (6.8), only the first one in the right side poses difficulties. Firstly we notice that

$$\begin{aligned}&E[(m(X_{N[1,2]})-m(X_1))(m(X_{N[1,1]})-m(X_1))]\\&\quad=E[(X_{N[1,2]}-X_1)^T\nabla_{X_1}m(\nabla_{X_1}m)^T(X_{N[1,1]}-X_1)]+O(M^{-1/n-\gamma_m/n}),\end{aligned}$$

because for example

$$|(X_{N[1,2]}-X_1)^T(\nabla_{X_1}m-\nabla_{X_{N[1,1]}}m)(\nabla_{X_1}m)^T(X_{N[1,1]}-X_1)|\leq c_1d_{1,2}^{2+\gamma_m},$$

where c_1 is some constant depending on m. In fact,

$$\begin{aligned}E[d_{1,2}^{\gamma_m}]=E[\omega_{X_1}(d_{1,2})^{2/n+\gamma_m/n}(\frac{d_{1,2}}{\omega_{X_1}(d_{1,2})^{1/n}})^{2+\gamma_m}]&\leq c_2E[\omega_{X_1}(d_{1,2})^{2/n+\gamma_m/n}]\\&\leq c_3M^{-2/n-\gamma_m/n}\end{aligned}$$

by (A2) and (A3). Now it is possible to apply Lemma 6.1:

$$|E[(X_{N[1,2]}-X_1)^T\nabla_{X_1}m(\nabla_{X_1}m)^T(X_{N[1,1]}-X_1)]|\leq c_4M^{-2/n-\gamma/n}\log^{2+6/n}M$$

with c_4 some constant. □

6.3.3 Bounding the Variance

The variance of the residual variance estimate $\hat{V}_M$ can be bounded straightforwardly using Theorem 4.1:

Theorem 6.4. *(A General Variance Bound). Assume that $(X_i,Y_i)_{i=1}^M$ is i.i.d. with $|Y_1|\leq 1$. The variance of $\hat{V}_M$ is bounded by*

$$\mathrm{Var}[\hat{V}_M]\leq\frac{80(1+4L(n))}{M}.$$

Proof. Because the output variables $(Y_i)_{i=1}^M$ are assumed to be bounded by 1, it holds that that $|(Y_i-Y_{N[i,1]})(Y_i-Y_{N[i,2]})|\leq 4$. Consequently Theorem 4.1 implies that

$$\mathrm{Var}[\sum_{i=1}^M(Y_i-Y_{N[i,1]})(Y_i-Y_{N[i,2]})]\leq 80(1+4L(n))M.$$

□

In general, by a comparison to the bounds on estimation bias in Theorem 6.3, it is seen that usually variance is negligible in the limit $M\to\infty$ if $n>6$. For example, Theorem 6.3 shows that in the presence of sufficient regularity, bias of order $M^{-3/n}$ can be achieved, which is more than the order $M^{-1/2}$ for statistical fluctuations in the case $n>6$.

The following bound is not optimal with respect to the constants, but it demonstrates the fact that when the residual is on average small, the variance of the estimator may be smaller than simpler proof techniques would indicate.

Theorem 6.5. *Suppose that $(X_i, Y_i)_{i=1}^M$ is i.i.d. with $|Y_1| \leq 1$; moreover $\mathcal{X} \subset [0,1]^n$ with $n \geq 2$. We assume that $m \in H(c_m, \gamma_m)$, where $\gamma_m \geq 1$. If the definition*

$$\sigma_4^4 = \sup_{x\in\mathcal{X}} E[r_1^4 | X_1 = x]$$

is made, then the variance of $\hat{V}_M$ is bounded by

$$\mathrm{Var}[\hat{V}_M] \leq \frac{32(1+2L(n))\sigma_4^4}{M} + O(\sigma_4^2 M^{-1-2/n}) + O(M^{-1-\min\{4/n,1\}}),$$

where the remainder terms are viewed with σ_4 and M as the free parameters and everything else fixed.

Proof. We divide $\sum_{i=1}^M (Y_i - Y_{N[i,1]})(Y_i - Y_{N[i,2]})$ into three parts $I_1 + I_2 + I_3$ with

$$\begin{aligned}
I_1 &= \sum_{i=1}^M (r_i - r_{N[i,1]})(r_i - r_{N[i,2]}) \\
I_2 &= \sum_{i=1}^M (m(X_i) - m(X_{N[i,1]}))(r_i - r_{N[i,2]}) \\
&\quad + \sum_{i=1}^M (r_i - r_{N[i,1]})(m(X_i) - m(X_{N[i,2]})) \\
I_3 &= \sum_{i=1}^M (m(X_i) - m(X_{N[i,1]}))(m(X_i) - m(X_{N[i,2]})).
\end{aligned}$$

By the smoothness of m and the bound on Y_1, it holds that

$$|(m(X_i) - m(X_{N[i,1]}))(m(X_i) - m(X_{N[i,2]}))| \leq (4 + c_m^2)\min\{1, \|X_i - X_{N[i,2]}\|^2\}.$$

Because of this inequality, Theorem 4.2 implies that

$$\mathrm{Var}[I_3] = O(M^{-1-\min\{4/n,1\}}).$$

Similarly

$$\begin{aligned}
&E[(m(X_i) - m(X_{N[i,1]}))^2 (r_i - r_{N[i,2]})^2 | (X_i)_{i=1}^M] \\
&\quad \leq (8\sigma_4^2 + 2c_m^2\sigma_4^2)\min\{1, \|X_i - X_{N[i,2]}\|^2\}.
\end{aligned}$$

Consequently, it holds that

$$\mathrm{Var}[I_2] = O(\sigma_4^2 M^{-1-2/n}).$$

Finally, Theorem 4.2 can applied to I_1, because

$$E[(r_i - r_{N[i,1]})^2 (r_i - r_{N[i,2]})^2 | (X_i)_{i=1}^M] \leq 4\sigma_4^4$$

and consequently

$$\mathrm{Var}[I_1] \leq \frac{32(2L(n)+1)\sigma_4^4}{M}.$$

□

The previous result is rather rough and it is more interesting to ask, whether the convergence

$$ME[(\hat{V}_M - \frac{1}{M}\sum_{i=1}^{M} r_i^2)^2] \to 0$$

happens in the limit $M \to \infty$. It is easy to be convinced that this would indicate asymptotic optimality of the estimator if the residual variance is above zero. The convergence question was considered by us in [36], but in that paper we were not able to show that the basic product method is asymptotically optimal. The exact asymptotic variance remains an open question at this point, but it can probably be derived from the limit theory of random geometric graphs (e.g. [49]).

6.4 Extending to $k > 1$

To define the estimator $\hat{V}_M$ we used only the first and second nearest neighbors. However, extending the method so as to employ $2k$ nearest neighbors is rather obvious and it is stated mathematically as

$$\hat{V}_{M,k} = \frac{1}{M(1+k^{-1})}\sum_{i=1}^{M}(Y_i - \frac{1}{k}\sum_{j=1}^{k} Y_{N[i,2j]})(Y_i - \frac{1}{k}\sum_{j=1}^{k} Y_{N[i,2j-1]}).$$

Theoretically, it is possible to show the following consistency result (see [36], the theorem is stated here in a simplified form):

Theorem 6.6. *If $n \leq 4$, then*

$$\limsup_{M\to\infty} E[M(\hat{V}_{M,k} - \frac{1}{M}\sum_{i=1}^{M} r_i^2)^2] = O(k^{-1}),$$

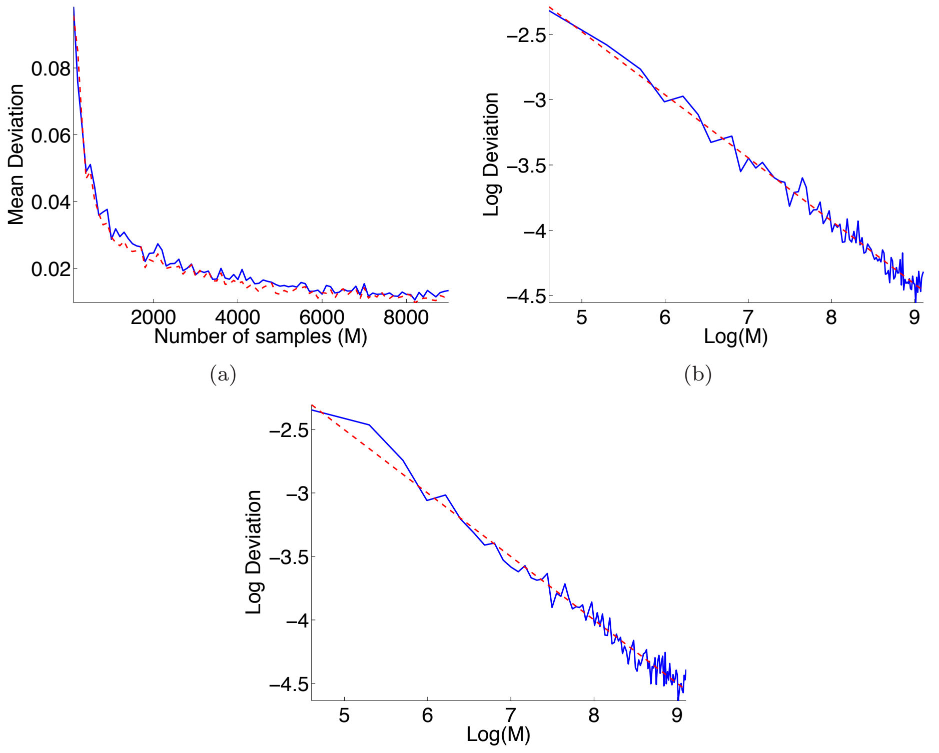
where the remainder term depends only on k and the dimensionality n.

In fact, the bias considerations in Section 6.3.2 indicate that Theorem 6.6 holds also for $n = 5$. The result implies that by increasing k with k/M small, $\hat{V}_{M,k}$ approaches the minimum variance estimator at least in L^2-sense. However, obviously increasing k also increases the systematic error (bias) of the estimator and thus one has to compromise between variance and bias.

Details on the bounds for the case $k > 1$ are found in our paper [36]. Because in most practical problems the benefit of using $k > 1$ is small and brings the additional difficulty of choosing k, fixing $k = 1$ is usually recommended.

6.5 Simulations

The 1-NN estimator (6.2) is compared to the product estimator in the presence of heteroscedastic noise. The task is to assess the practical impact of the faster rate of convergence for the latter method.



(a)

(b)

(c)

Figure 6.1: Results for the model in Equation (6.15) with uniformly distributed X. The dashed line in a) corresponds to the 1-NN estimator. The related logarithmic curve for the product estimator is in b) and that for 1-NN in c). The least squares fits (dashed lines) in b) and c) have slopes -0.48 and -0.5 respectively.

6.5.1 Linear Problems

In the first simulation example, the observations are related to the inputs by

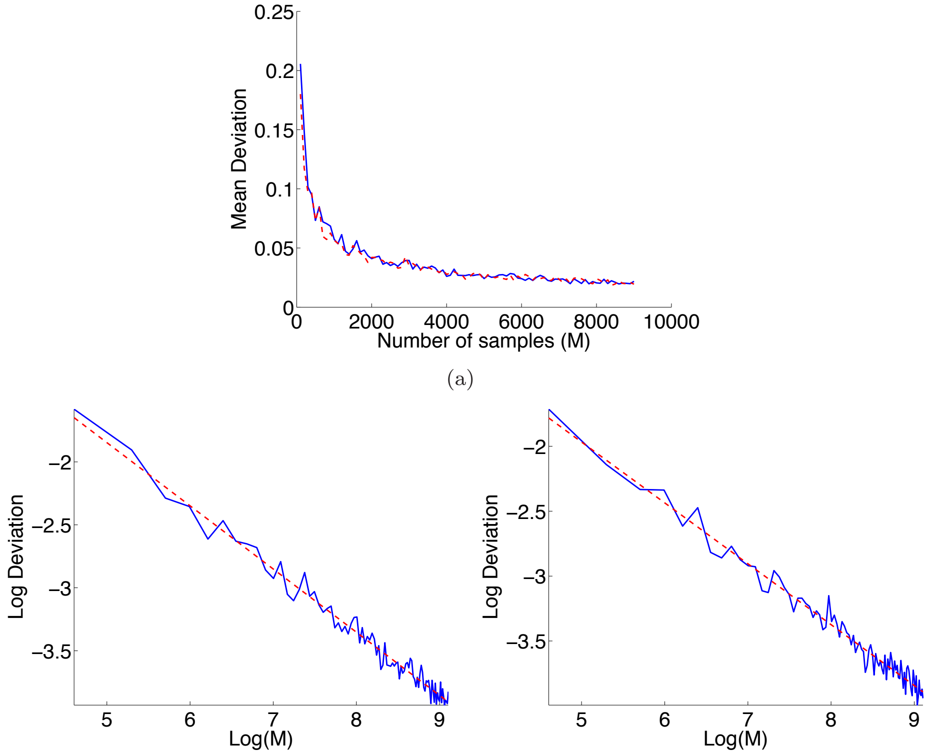
$$Y = X^{(1)} + 3X^{(2)} + \sin(4\pi X^{(1)})\epsilon, \tag{6.15}$$

where $(X^{(1)}, X^{(2)})$ is sampled from the uniform distribution on $[0,1]^2$ and $\epsilon \sim N(0,1)$ is independent Gaussian noise. The variance of the residual is in this case 0.5. The experiment is repeated 100 times with the number of samples ranging from 100 to 9000 and the mean absolute deviation (the absolute value of the difference between the estimate and the real value) from the real noise variance is calculated.

The results are drawn in Figure 6.1. To clarify the rate of convergence, we have drawn the logarithm of the mean deviation with respect to $\log M$.

In the second experiment, the higher dimensional model

$$Y = X^{(1)} + X^{(2)} + X^{(3)} + X^{(4)} + X^{(5)} + X^{(6)} + \epsilon \tag{6.16}$$



(a)

(b)

(c)

Figure 6.2: Results for the model in Equation (6.16). The dashed line in a) corresponds to the 1-NN estimator. The related logarithmic curve for the product estimator is in b) and that for 1-NN in c). The dashed lines in b) and c) have slopes -0.5 and -0.47 respectively.

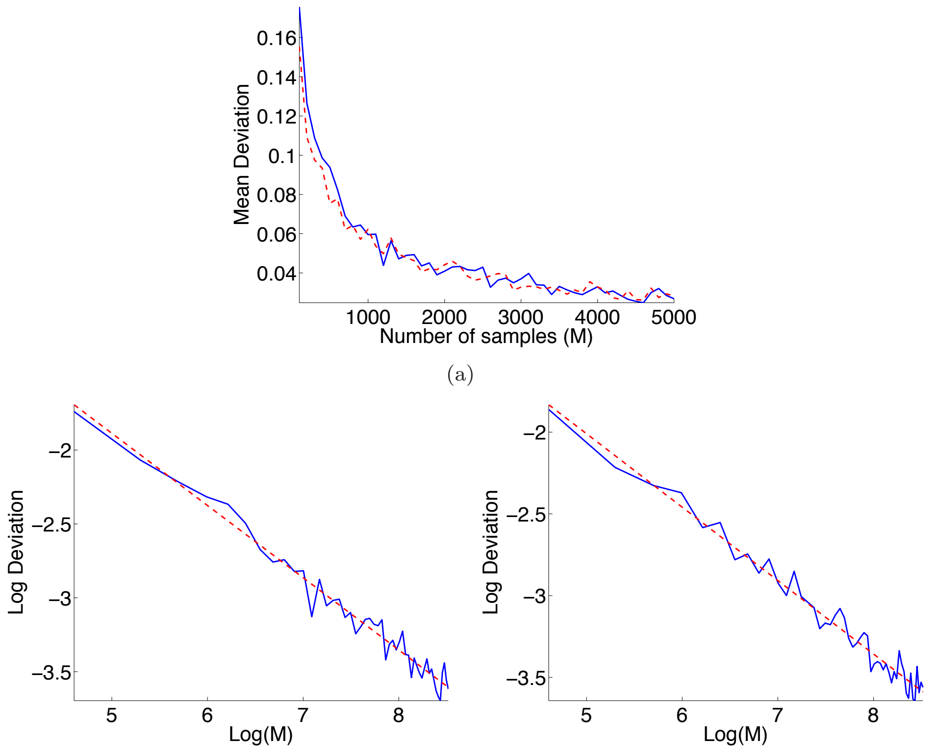
was tested. The data consists of six dimensional vectors and in principle the product estimator should have some advantage over the 1-NN estimator. However, from the results in Figure 6.2 we can see that again this is not the case; apparently here asymptotics show up only for very large values of M.

6.5.2 Non-linearities

While linear models are a good benchmark, the absence of second order terms in the Taylor expansion makes estimation relatively easy. To examine the general non-linear case, the same experiment as in Section 6.5.1 was performed with

$$Y = \sin(2\pi X^{(1)})\sin(2\pi X^{(2)})\sin(2\pi X^{(3)}) + 0.2\sin(4\pi X^{(1)})\epsilon, \tag{6.17}$$

where X is uniformly distributed and ϵ is independent Gaussian noise. From Figure 6.3 it can be seen that again there is no significant difference between the product estimator and the 1-NN estimator.



(a)

(b)

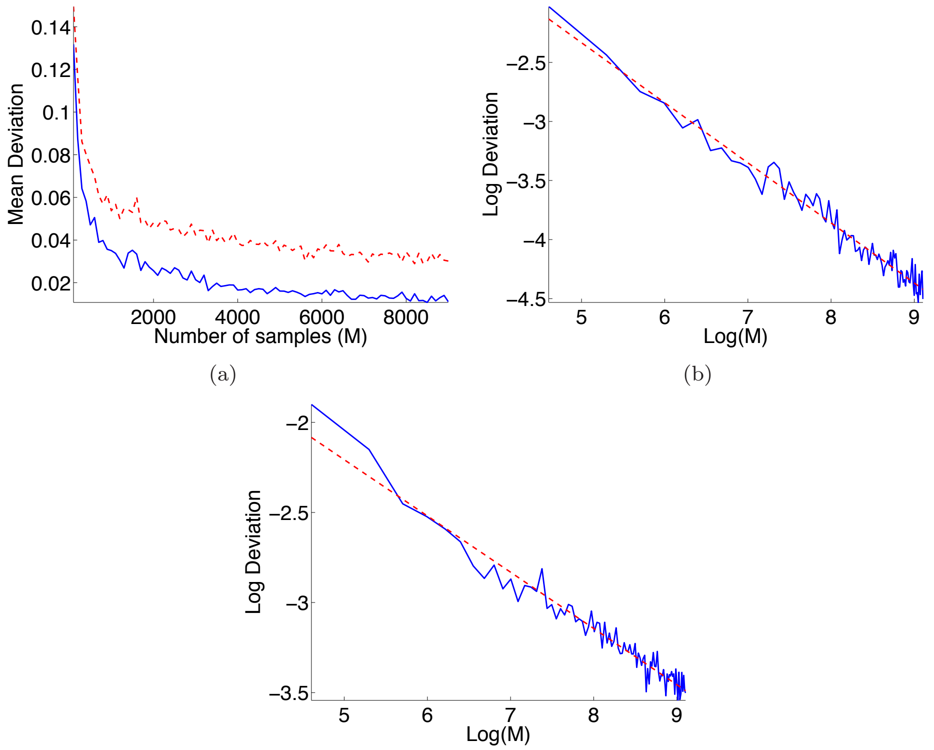
(c)

Figure 6.3: Results for the model in Equation (6.17). The dashed line in a) corresponds to the 1-NN estimator. The related logarithmic curve for the product estimator is in b) and that for 1-NN in c). The dashed lines in b) and c) correspond to lines with slopes −0.47 and −0.45 respectively.

To assess the effect of dimensionality, we took the model

$$Y = (X^{(1)})^2 + 3X^{(2)} + \sin(4\pi X^{(1)})\epsilon \tag{6.18}$$

and repeated the experiment as before. However, now the input was taken as uniform on the cube $[0,1]^6$. Consequently, four noise variables were added. Figure 6.4 shows that the 1-NN estimator is more accurate in this case.



(a)

(b)

(c)

Figure 6.4: Results for the model in Equation (6.18). The dashed line in a) corresponds to the 1-NN estimator. The related logarithmic curve for the product estimator is in b) and that for 1-NN in c). The dashed lines in b) and c) correspond to lines with slopes −0.51 and −0.31 respectively.

Chapter 7

Conclusion and Open Questions

7.1 The Boundary Correction

The Nearest Neighbor Distance Expansion

In Equation (1.5), we find an asymptotic characterization of the behavior of $E[d_{1,k}^{\alpha}]$ in the limit $M \to \infty$: if the i.i.d. sample $(X_i)_{i=1}^{M}$ takes values in a set $\mathcal{X} \subset \Re^n$, then under sufficient regularity,

$$M^{\alpha/n} E[d_{1,k}^{\alpha}] \to V_n^{-\alpha/n} \frac{\Gamma(k+\alpha/n)}{\Gamma(k)} \int_{\mathcal{X}} q(x)^{1-\alpha/n} dx \qquad (M \to \infty) \tag{7.1}$$

and also the rate of convergence is understood [18]. Equation (1.6) on the other hand conjectures the possibility for a higher order expansion, which would better approximate $M^{\alpha/n} E[d_{1,k}^{\alpha}]$ for finite sample sizes. To assess which sources of error are dominant in such an expansion, it was shown in Chapter 3 that the major source of deviation from the limit in (7.1) tends to come from the boundary $\partial\mathcal{X}$.

When analyzing the boundary effect, at first sight it seems that highly non-linear integrals arise rendering simplification attempts challenging. In Chapter 3 it was shown that the problem can be circumvented if x in the conditional expectation

$$E[d_{1,k}^{\alpha} | X_1 = x] \tag{7.2}$$

is allowed to vary in certain sense. It then follows that the boundary effect enters in a natural way:

$$\begin{aligned} E[d_{1,k}^{\alpha}] =& V_n^{-\alpha/n} \frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)} \int_{\mathcal{X}} q(x)^{1-\alpha/n} dx \\ &+ (D - V_n^{-\alpha/n-1/n}) \frac{\Gamma(k+\alpha/n+1/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n+1/n)} \int_{\partial\mathcal{X}} q(x)^{1-\alpha/n-1/n} dS \\ &+ O(M^{-\gamma/n-\alpha/n} \log^{2+2\alpha/n+4/n} M), \end{aligned} \tag{7.3}$$

where D depends on α and n (see Theorem 3.3). Unfortunately Assumptions (A2) and (A4) are restrictive and leave the analysis of unbounded densities as an open problem. It is probable that some new phenomena arise in this case.

Secondly, the possibility of moving to even higher orders in one way or another should be investigated; some attempts to this direction have been made in [59] albeit in the context of classification. It is not necessarily so useful to have the exact form of the constants in the expansion as long as the dependency on k and M is understood.

Implications in Statistical Estimation

While the expansion (7.3) is of theoretical interest in its own right, it also has implications in the theory of statistical estimation. In the context of entropy estimation, an error analysis was established in Chapter 5 characterizing the dominant error term for the standard nearest neighbor estimator of Equation (1.8):

$$H_{1-\alpha/n}(X) \approx \frac{n}{\alpha} \log(\frac{V_n^{\alpha/n}\Gamma(k)}{\Gamma(k+\alpha/n)} M^{\alpha/n-1} \sum_{i=1}^{M} d_{i,k}^{\alpha}). \tag{7.4}$$

Moreover, it turned out that the boundary effect vanishes under an appropriate weighted average of different estimates and the improved accuracy is demonstrated through simulations as well.

While the proposed weighting takes the form

$$M^{-1+\alpha/n} \sum_{i=1}^{M} \sum_{k=1}^{l} w_k d_{i,k}^{\alpha}$$

for some weights depending on l, n and α (see Example 5.1), it might turn out that

$$M^{-1+\alpha/n} \sum_{i=1}^{M} (\sum_{k=1}^{l} w_k d_{i,k})^{\alpha}$$

works out as well and yields a lower variance in some cases. Especially for differential entropy estimation, it would be of interest to consider estimators using the quantity

$$\sum_{i=1}^{M} \log(\sum_{k=1}^{l} w_k d_{i,k})$$

as an attempt to reduce the increase in variance, while preserving the low bias.

From the more general point of view, similar ideas apply not only to variance but also in general for statistical testing and estimation in those cases where nearest neighbor distributions might turn out to be useful. It would be interesting to go through statistical literature and find out where nearest neighbor based estimators are applicable in some natural way and what are their properties.

7.2 Nearest Neighbor Bounds

The nearest neighbor upper and lower bounds of Chapter 2 are a potentially useful tool because of their non-asymptotic and general nature. Among others, it was proven that under the i.i.d. assumption on $(X_i)_{i=1}^M$

$$V_{n,p}^{-\alpha/n} E[\mathcal{M}(X_1)^{-\alpha/n}] \frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)} \leq E[d_{1,k}^\alpha] \leq (\frac{2^n n^{n/2} k}{M})^{\alpha/n},$$

where $\mathcal{M}(x)$ is the maximal function of the common density q. Standard tools from measure theory were then applied in order to relate the lower bound to the L^p-norms of q. Moreover, a geometric boundary analysis provides also the bound

$$E[d_{i,k}^\alpha] \leq (\frac{2^n \lambda(\mathcal{X}) k}{V_{n,p} M})^{\alpha/n} + o(M^{-\alpha/n})$$

as long as $\mathcal{X}$ is not a fractal and stays sufficiently regular. Moreover, it was shown that upper bounds can also be derived using the instrinsic dimensionality of $\mathcal{X}$ even though in that case some tightness seems to be lost. For an unexplored topic of research, observe that the upper bounds were of geometric nature; whether introducing some probabilistic structure would lead to sharper inequalities is an open question.

Some applications were discussed including the analysis of high dimensional spaces and convergence analysis in non-parametric statistics. In addition to these, because nearest neighbor distances relate closely to the geometric properties of the underlying space, one might want to explore, whether unexplored applications in the analysis of metric spaces and discrete geometry could be found, see e.g. [44]. We also mention in [40] that the bounds might be useful in addressing important problems in random sequential adsorption [63].

7.3 Residual Variance Estimation

The theoretical properties of the product estimator

$$\frac{1}{M}\sum_{i=1}^{M}(Y_{N[i,2]} - Y_i)(Y_{N[i,1]} - Y_i) \tag{7.5}$$

were analyzed in Chapter 6 and shown to be attractive, though other good alternatives exist including the Gamma test. It was shown that while in general

$$E[(\frac{1}{2M}\sum_{i=1}^{M}(Y_{N[i,1]} - Y_i)^2 - \mathrm{Var}[r])^2] \geq c_1 M^{-\min\{1,4/n\}}$$

for some constant $c_1 > 0$ independent of M, under sufficient regularity we have

$$E[(\frac{1}{M}\sum_{i=1}^{M}(Y_{N[i,2]} - Y_i)(Y_{N[i,1]} - Y_i) - \mathrm{Var}[r])^2] \leq c_2 M^{-\min\{1,6/n\}} \log^{2+6/n} M$$

for any $n > 4$ and some constant $c_2 > 0$ independent of M, whereas for $n \leq 4$

$$E[(\frac{1}{M}\sum_{i=1}^{M}(Y_{N[i,2]} - Y_i)(Y_{N[i,1]} - Y_i) - \mathrm{Var}[r])^2] \leq c_3 M^{-1}$$

with $c_3 > 0$ independent of M. In other words, the product estimator is less sensitive to the curse of dimensionality than (6.2).

The proof technique exploits a local uniformity property of nearest neighbor distributions, which possibly could find other uses as well e.g. in the design of statistical estimators. It would also be important to examine the effect of non-independent sampling on convergence. It is likely that when the dependency between samples is not too strong, consistency can be established; in fact, the author believes that actually not even stationarity is necessary if an appropriate definition of residual variance is used. Of course, problematic cases such as time series data exist and the practicioners should be aware of potential difficulties.

While other unexplored theoretical topics can be found, such as deriving confidence intervals and a locally linear extension of (7.5), promoting the practical use of the methods is most important as was stated in [13]. In fact, it is surprising that nearest neighbor based methods have not gained more popularity, as they have a relatively low computational complexity and moreover, non-parametric residual variance estimation is a natural generalization of linear correlation.

Bibliography

[1] N. Reyhani. Y. Ji A. Sorjamaa, J. Hao and A. Lendasse. Methodology for long-term prediction of time series. *Neurocomputing*, 70(16-18):2861–2869, 2007.

[2] F. Avram and D. Bertsimas. On central limit theorems in geometrical probability. *Annals of Applied Probability*, 3(4):1033–1046, 1993.

[3] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. van der Meulen. Nonparametric entropy estimation: an overview. *International Journal of Mathematical Statistics Sciences*, 6:17–39, 1997.

[4] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In *CDT '99: Proceedings of the 7th International Conference on Database Theory*, volume 1540 of *Lecture Notes in Computer Science*, pages 217–235. Springer-Verlag, 1999.

[5] L. D. Brown and M. Levine. Variance estimation in nonparametric regression via the difference sequence method. *Annals of Statistics*, 35(5):2219–2232, 2007.

[6] T. Cai, M. Levine, and L. Wang. Variance function estimation in multivariate nonparametric regression. *Journal of Multivariate Analysis*, 100(1):126–136, 2009.

[7] J. H. Conway and N. J. Sloane. *Sphere Packings, Lattices and Groups.* Grundlehren der mathematischen Wissenschaften. Springer, 3rd edition edition, 1998.

[8] J. A. Costa and A. O. Hero. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing*, 52(8):2210–2221, 2004.

[9] L. Devroye, L. Györfi, Krzyżak, and G. Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *Annals of statistics*, 22(3):1371–1385, 1994.

[10] L. Devroye, L. Györfi, and D. Schäfer. The estimation problem of minimum mean squared error. *Statistics and Decisions*, 21:15–28, 2003.

[11] Y. G. Dmitriev and F. P. Tarasenko. On the estimation of functionals of the probability density and its derivatives. *Theory of probability and its applications*, 18(3):628–633, 1974.

[12] E. Eirola, E. Liitiäinen, A. Lendasse, F. Corona, and M. Verleysen. Using the Delta test for variable selection. In *ESANN, European Symposium on Artificial Neural Networks, Bruges (Belgium), April*, pages 25–30, 2008.

[13] D. Evans. *The Gamma Test: Data derived estimates of noise for unknown smooth models using near neighbour asymptotics*. PhD thesis, Cardiff University, 2002.

[14] D. Evans. Estimating multiplicative noise. In *Proceedings of the 18th International Conference on Noise and Fluctuations*, pages 99–102, 2005.

[15] D. Evans. A computationally efficient estimator for mutual information. *Proceedings of the Royal Society A*, 464(2093):1203–1215, 2008.

[16] D. Evans. A law of large numbers for nearest neighbour statistics. *Proceedings of the Royal Society A*, 464(2100):3175–3192, 2008.

[17] D. Evans and A. J. Jones. Non-parametric estimation of residual moments and covariance. *Proceedings of the Royal Society A*, 464(2099):2831–2846, 2008.

[18] D. Evans, A. J. Jones, and W. M. Schmidt. Asymptotic moments of near neighbour distance distributions. *Proceedings of the Royal Society A*, 458(2028):2839–2849, 2002.

[19] D. Francois, V. Wertz, and M. Verleysen. The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, 19(7):873–886, 2007.

[20] T. Gasser, L. Stroka, and C. Jennen-Steinmetz. Residual variance and residual pattern in nonlinear regression. *Biometrika*, 73(3):625–633, 1986.

[21] C. Giannella. New instability results for high-dimensional nearest neighbor search. *Information Processing Letters*, 109(19):1109–1113, 2009.

[22] L. Golshani, E. Pasha, and G. Yari. Some properties of Rényi entropy and Rényi entropy rate. *Information Sciences*, 179(14):2426–2433, 2009.

[23] L. Györfi and E. C. van der Meulen. Density-free convergence properties of various estimators of entropy. *Computational Statistics and Data Analysis*, 5(4):425–436, 1987.

[24] P. Hall and J. Marron. On variance estimation in nonparametric regression. *Biometrika*, 77(2):415–419, 1990.

[25] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 1998.

[26] A. Hero, B. Ma, O. Michel, and J. Gorman. Applications of entropic spanning graphs. *IEEE Signal Processing Magazine*, 19(5):85–95, 2002.

[27] H. Joe. On the estimation of entropy and other functionals of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 41(4):683–697, 1989.

[28] M. Kohler, A. Kržyzak, and H. Walk. Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data. *Journal of Multivariate Analysis*, 97(2):311–323, 2006.

[29] M. Kohler, A. Kržyzak, and H. Walk. Optimal global rates of convergence for nonparametric regression with unbounded data. *Journal of Statistical Planning and Inference*, 139(4):1286–1296, 2008.

[30] L. F. Kozachenko and N. N. Leonenko. Sample estimate of entropy of a random vector. *Problems of Information Transmission*, 23(2):95–101, 1987.

[31] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69(6), 2004.

[32] S. R. Kulkarni and S. E. Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41(4):1028–1039, 1995.

[33] M. Lejuene and P. Sarda. Smooth estimators of distribution and density functions. *Computational Statistics and Data Analysis*, 14(4):457–471, 1992.

[34] N. N. Leonenko, L. Pronzato, and V. Savani. A class of Rényi information estimators for multidimensional densities. *Annals of Statistics*, 36(5):2153–2182, 2008.

[35] E. Liitiäinen, F. Corona, and A. Lendasse. Nearest neighbor distributions and noise variance estimation. In *ESANN 2007, European Symposium on Artificial Neural Networks*, pages 67–72, 2007.

[36] E. Liitiäinen, F. Corona, and A. Lendasse. Residual variance estimation using a nearest neighbor statistic. *Journal of Multivariate Analysis*, 101(4):811–823, 2010.

[37] E. Liitiäinen, A. Lendasse, and F. Corona. Non-parametric residual variance estimation in supervised learning. In *Computational and Ambient Intelligence*, volume 450 of *Lecture Notes in Computer Science*, pages 63–71. Springer Berlin/Heidelberg, 2007.

[38] E. Liitiäinen, A. Lendasse, and F. Corona. A boundary corrected expansion of the moments of nearest neighbor distribution. *Random Structures and Algorithms*, 37(2):223–247, 2010.

[39] E. Liitiäinen, Amaury Lendasse, and Francesco Corona. On the statistical estimation of Rényi entropies. In *IEEE Conference on Machine Learning for Signal Processing*, 2009.

[40] Elia Liitiäinen, Amaury Lendasse, and Francesco Corona. Bounds on the power-weighted mean nearest neighbor distance. *Proceedings of the Royal Society, Series A*, 464(2097):2293–2301, 2008.

[41] Elia Liitiäinen, Amaury Lendasse, and Francesco Corona. On non-parametric residual variance estimation. *Neural Processing Letters*, 28(3):155–167, 2008.

[42] Elia Liitiäinen, Michel Verleysen, Francesco Corona, and Amaury Lendasse. Residual variance estimation in machine learning. *Neurocomputing*, 72(16-18):3692–3703, October 2009.

[43] P. Massart. *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003.* Lecture Notes in Mathematics / Ecole d'Eté Probabilit.Saint-Flour. Springer, 1 edition, 2007.

[44] J. Matousek. Note on bi-lipschitz embeddings into normed spaces. *Commentationes Mathematicae Universitatis Carolinae*, 33(1):51–55, 1992.

[45] H. G. Muller and U. Stadtmuller. On variance function estimation with quadratic forms. *Journal of Statistical Planning and Inference*, 35(2):213–231, 1993.

[46] H. G. Muller and U. Stadtüller. Estimation of heteroscedasticity in regression analysis. *Annals of Statistics*, 15(2):610–625, 1987.

[47] M. Penrose. *Random Geometric Graphs.* Number 5 in Oxford Studies in Probability. Oxford University Press, 2003.

[48] M. Penrose and A. R. Wade. Multivariate normal approximation in geometric probability. *Journal of Statistical Theory and Practice*, 2:293–326, 2008.

[49] M. D. Penrose. Gaussian limits for random geometric measures. *Electronic Journal of Probability*, 12(35):989–1035, 2007.

[50] M. D. Penrose. Laws of large numbers in stochastic geometry with statistical applications. *Bernoulli*, 13(4):1125–1150, 2007.

[51] M. D. Penrose and A. R. Wade. On the total length of the random minimal directed spanning tree. *Advances in Applied Probability*, 38(2):336–372, 2006.

[52] V. Pestov. On the geometry of similarity search: Dimensionality curse and concentration of measure. *Information Processing Letters*, 73(1-2):47–51, 2000.

[53] C. E. Rasmussen. *Gaussian Process for Machine Learning.* MIT Press, 2006.

[54] A. Renyi. On measures of information and entropy. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561, 1961.

[55] J. Rice. Bandwidth choice for nonparametric regression. *The Annals of Statistics*, 12(4):1215–1230, 1984.

[56] W. Rudin. *Real and Complex Analysis.* Higher Mathematics Series. McGraw-Hill Science, 1986.

[57] O. Bousquet S. Boucheron and G. Lugosi. *Advanced Lectures on Machine Learning*, chapter Concentration inequalities, pages 208–240. Lecture Notes in Artificial Intelligence. Springer, 2004.

[58] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[59] R. R. Snapp and S. S. Venkatesh. Asymptotic expansions of the k nearest neighbor risk. *Annals of statistics*, 26(3):850–878, 1998.

[60] V. Spokoiny. Variance estimation for high-dimensional regression models. *Journal of Multivariate Analysis*, 82(1):111–133, 2002.

[61] A. Stefánsson, N. Koncar, and A. J. Jones. A note on the Gamma test. *Neural Computing & Applications*, 5(3):131–133, 1997.

[62] C. J. Stone. Optimal rates of convergence for nonparametric estimators. *Annals of Statistics*, 8(6):1348–1360, 1980.

[63] J. Talbot, G. Tarjus, P. R. Van Tassel, and P. Viot. From car parking to protein adsorption: an overview of sequential adsorption processes. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 165(1-3):287–324, 2000.

[64] A. Tewari and A. M. Gokhale. A geometric upper bound on the mean first nearest neighbour distance between particles in three-dimensional microstructures. *Acta Materialia*, 52(17):5165–5168, 2004.

[65] C. Tricot. Two definitions of fractional dimension. *Mathematical Proceedings of the Cambridge Philosophical Society*, 91(1):57–74, 1982.

[66] A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Annals of Statistics*, 36(3):1435–1463, 2008.

[67] M. Verleysen. Machine learning of high-dimensional data: local artificial neural networks and the curse of dimensionality. Agregation in higher education thesis, Université Catholique de Louvain (UCL), February 2001.

[68] J. von Neumann. Distribution of the ratio of the mean square successive difference to the variance. *Annals of Mathematical Statistics*, 12(4):367–395, 1941.

[69] A. R. Wade. Explicit laws of large numbers for random nearest-neighbour type graphs. *Advances in Applied Probability*, 39(2):326–342, 2007.

[70] G. Yang, L. Le Cam, and M. Lucien. *Asymptotics in Statistics: Some Basic Concepts*. Springer Series in Statistics. Springer, 2000.

[71] J. Yukich and Y. Baryshnikov. Gaussian limits for random measures in geometric probability. *Annals of Applied Probability*, 15(1A):213–253, 2005.