RANDOM GRAPHS FOR STRUCTURE DISCOVERY IN HIGH-DIMENSIONAL DATA

by

José António O. Costa

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in The University of Michigan
2005

Doctoral Committee:
Professor Alfred O. Hero III, Chair
Professor Jeffrey A. Fessler
Professor Susan A. Murphy
Professor David L. Neuhoff

To my family.

ACKNOWLEDGEMENTS

Having lived most of my life in a place with (almost) perfect weather, every year there is a day during the peak of the long Michigan winters where I start doubting about my choice of graduate school. However, that question is quickly dismissed by reminding myself of the advisor I was very fortunate to have. I could not have been anywhere else! I am deeply grateful to my advisor, Professor Alfred Hero, for his outstanding support, wise guidance and invaluable advices, both professional and personal. His views of my particular research fields, and science in general, have been the backbone of my research over the past years. His vast knowledge of many different scientific fields, his creativity and insight have been a great source of inspiration and, no doubtfully, will continue to exert great influence in my future work.

I would like to extend my gratitude to Professor Jeffrey Fessler, Professor Susan Murphy and Professor David Neuhoff. Right from the beginning of grad school, they have introduced me to new worlds, taught me how to be a better researcher or pointed out many interesting connections between my work and other different problems that have proved very helpful. I am also very grateful for their willingness to support several of my endeavors throughout the years.

Of course, the path leading to this dissertation would not have started without the original influence of two persons. I would like to thank Professor José Leitão for introducing me to the wonderful world of statistical signal processing, showing me

the value of abstract research in engineering and "helping me to learn" what research is all about. I would also like to thank Professor Mário Figueiredo for his continuing support and many fruitful conversations full of wise scientific and personal insights.

I would like to thank my family – my parents, my sister and my grandmother – for their unconditional support throughout the years, which provided a never ending source of energy to overcome the many obstacles of grad school. This section would not be complete without acknowledging the support of many friends who have honored me with their true friendship. In particular, my deep thanks to Karla Bessa, Doron Blatt, Rui Castro, Pedro Granja, Gabriel Lopes, Pedro Mendes, André and Filipa Neves, Neal Patwari, Eduardo Silva and Paulo Tabuada.

The research presented in this thesis was partially funded by the Fundação para a Ciência e Tecnologia through grant SFRH/BD/2778/2000, the National Institutes of Health through grant NIH 1PO1 CA87634-01, a fellowship of the Department of Electrical Engineering and Computer Science, University of Michigan, and a Horace H. Rackham Pre-doctoral fellowship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure

ABSTRACT

Originally motivated by computational considerations, we demonstrate how computational efficient and scalable graph constructions can be used to encode both statistical and spatial information and address the problems of dimension reduction and structure discovery in high-dimensional data, with provable results.

We discuss the asymptotic behavior of power weighted functionals of minimal Euclidean graphs, proving upper and lower bounds for the respective convergence rates and connecting them to the problem of nonparametric estimation of entropy.

We then extend the convergence results from Euclidean graphs to the setting of data that spans a high-dimensional space but which contain fundamental features that are concentrated on lower-dimensional subsets of this space – curves, surfaces or, more generally, lower-dimensional manifolds. In particular, we have developed a novel geometric probability approach to the problem of estimating intrinsic dimension and entropy of manifold data, based on asymptotic properties of graphs such as Minimal Spanning Trees or k-Nearest Neighbor graphs. Unlike previous solutions to this problem, we are able to prove statistical consistency of the obtained estimators for the wide class of Riemann submanifolds of an Euclidean space. We also propose a graph based dimensionality reduction method aimed at extracting lower dimensional features designed expressly to improve classification tasks, with applications to both supervised and semi-supervised learning problems.

Finally, using neighborhood graphs and the multidimensional scaling principle,

we develop a general tool for dimensionality reduction in sensor networks, where communication constraints exist and distributed optimization is required. This tool is illustrated through an application to localization in sensor networks.

CHAPTER 1

Introduction

1.1 High-dimensional Data Sets and Their Challenges

Increasingly intricate and rich data sets at the heart of many of today's most common applications, from video surveillance to medical information systems, are raising new problems in data storage, access and especially data exploration. Continuing technological advances in both sensing and media storage capabilities are enabling the development of systems that generate massive amounts of new types of data and information. Today's computer and wireless sensor networks, medical and biological systems or imaging and remote sensing applications produce complex high-dimensional signals that need careful interpretation in order to extract useful information. How can one make sense of this enormous quantity of information and use it in a meaningful way, from a simple problem of visualizing a relevant data characteristic to more complex decision making tasks?

The problems inherent to the exploration of high dimensional datasets are primarily consequence of the well known curse of dimensionality phenomenon. On the

one hand, the computational complexity of many algorithms grows exponentially in the number of input dimensions, making it impossible to handle large datasets. On the other hand, the fact that filling a space with data points is increasingly harder as dimension grows can have a drastic impact on the statistical performance of the same algorithms. For example, considering data points with a one-dimensional standard normal distribution, 70% of the probability mass is contained in a sphere/interval of radius one standard deviation (i.e., the $[-1, 1]$ interval). For a ten-dimensional standard normal, the same sphere will contain only 0.02% of the probability mass and a radius of more than three standard deviations is needed to obtain again 70%. As consequence, achieving the same level of accuracy with a particular algorithm will require increasingly larger datasets for higher dimensions. It is this tradeoff between manageable computational resources and good statistical accuracy that apparently makes it seem impossible to ever fully explore such rich and complex datasets.

However, many real life signals that have high dimensional representations, and thus appear complex, can actually be explained by only a few simple variables, as a result of coherent structures in nature that lead to strong correlations between inputs. In fact, it is often the case that the apparent complexity of the data is an artifact of its representation and is not related with the actual complexity of the generating process. For concreteness, consider a set of many images of a person's face observed under different pose and lighting conditions (see Figure 1.1). Mathematically these images can be regarded as a collection of points in a high dimensional vector space, with each input dimension corresponding to the lighting intensity of a particular image pixel. Although the images dimensionality may be quite high (e.g., 4096 for 64×64 pixel images), its meaningful structure can actually be described by only a few variables. In fact, the set of images lies on a three-dimensional manifold, or constrained surface, that can be parameterized by two pose variables, the up-down

Figure 1.1: Sample images from a face database [103].

and the left-right rotations of the face, and by one lighting angle variable.

Understanding the aforementioned high dimensional data sets thus requires greatly reducing the dimensionality of the inputs while preserving perceptual similarities in their structure. This would allow for the efficient processing of the data, revealing structure and providing further insight about the process generating the data set. All these concerns currently play a central role in several scientific fields: from feature extraction in pattern recognition, manifold learning in machine learning, latent variable selection in statistics, compression and coding in information theory, to decentralized detection and estimation in wireless sensor networks.

The aim of this dissertation is the development of robust nonparametric methods to access fundamental quantities characterizing high-dimensional data sets and its consequences for the exploration of complex large scale signals.

1.2 Background and Previous Work

Analyzing high-dimensional data sets can be divided into two complementary tasks.

On the one hand, by looking at a high-dimensional data set, one would like to infer quantities that characterize its complexity. Two important quantities play a central role in this assessment. One is the intrinsic dimension of the data that "roughly" characterizes the number of independent variables needed to explain the

phenomenon originating the data. The other is the differential entropy of the data points that "roughly" measures the statistical information conveyed by the complex signals.

On the other hand, complementary to the task described above, is the problem of transforming the data into a more efficient representation by reducing its dimensionality.

1.2.1 Entropy Estimation

Entropy, relative information and divergence measures play a central role in many applications where they can be used as a discriminant between samples with different characteristics. For example, the α-divergence between two probability densities f and g is:

$$D_\alpha(f||g) = \frac{1}{\alpha - 1} \ln \int f^\alpha(\boldsymbol{x}) g^{1-\alpha}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \ .$$

$D_\alpha(f||g)$ measures the similarity between f and g in the sense that $D_\alpha(f||g) \geq 0$ with equality iff $f = g$ almost everywhere. When $\alpha \to 1$, the well known Kullback-Leibler divergence is obtained, $D_1(f||g) = \int f^\alpha(\boldsymbol{x}) \ln \left(f^\alpha(\boldsymbol{x}) / g^\alpha(\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{x}$.

Applications where such discriminants have a natural application include: texture classification, feature clustering, image indexing or image registration, which are all core problems in areas such as geographical information systems, medical information processing, multi-sensor fusion and image content based retrieval. For example, the mutual information method of image registration (see [63] and references therein) searches through a set of coordinate transformations to find the one that minimizes an the entropy of the joint feature distribution of the two images. In a similar way, a statistical image retrieval algorithm (see [40]) searches trough a database of images to choose the image whose feature distribution is the closest to the query

image in a minimum information divergence sense. Other applications in signal processing include vector quantization [30] and entropy characterization of time-frequency distribution [84].

As the data distribution is usually not known in advance, estimating entropy and divergence from a finite number of data points is central to practical implementations in the above mentioned applications. The problem of entropy estimation has long been of interest to several comunities: e.g., the paper by Beirlant *et al.* [5] presents a thorough survey on the subject of Shannon entropy estimation. When no good parametric model of the features' probability distribution is available, one has to resort to non-parametric methods for entropy estimation. Most of the non-parametric entropy estimation techniques proposed so far are based on estimation of the underlying probability distribution with subsequent substitution (*plug-in*) of these estimates into the entropy functional. These methods, however, suffer from severe drawbacks, specifically: density estimator performance is poor without smoothness conditions; large number of tunable parameters (e.g., kernel bandwidth, type of kernel, etc); no unbiased estimators generally exist; density estimators have high variance and are sensitive to outliers; the high dimensional integration required to evaluate the entropy might be difficult. Motivated by this observation, a method was developed in [39] that directly estimates the entropy functional without having to estimate the probability distribution. The study of the performance of this type of non-parametric direct entropy estimator can thus have a considerable impact in practical applications.

1.2.2 Intrinsic Dimension Estimation

The classical solution to the problem of intrinsic dimensionality estimation is based on the linear projection paradigm [44]. The data set is projected on subspaces

of different dimensions and intrinsic dimensionality is chosen to be the corresponding dimension of the subspace that provides the best fit. This is usually accomplished by applying principal component analysis (PCA), factor analysis, or multidimensional scaling (MDS). Of course, these methods assume that the data set can be well approximated by a linear subspace embedded in the original high-dimensional vector space. As they do not account for non-linearities, linear methods tend to overestimate intrinsic dimension. Both nonlinear PCA [52] methods and the ISOMAP [103] try to circumvent this problem but they still rely on approximate and costly estimates of the fitting residuals.

More sophisticated methods for intrinsic dimension estimation are conceptually related to the estimation of the following functional of the density f of the data points:

$$\log \int_{B(\boldsymbol{y}_0,r)} g(f(\boldsymbol{y}))\, \mu(\mathrm{d}\boldsymbol{y}) \ , \tag{1.1}$$

where g is a strictly increasing function and $B(\boldsymbol{y}_0, r)$ is the ball of radius r centered at $\boldsymbol{y}_0$. Under suitable regularity conditions on f and g, using the mean value theorem results in:

$$\log \int_{B(\boldsymbol{y}_0,r)} g(f(\boldsymbol{y}))\, \mu(\mathrm{d}\boldsymbol{y}) = m\, \log r + c + o(1) \ , \tag{1.2}$$

where m is the data intrinsic dimensionality, c is a constant depending on f, g and the volume of the unit sphere and $o(1) \to 0$ when $r \to 0$. By choosing different functions g and radii r one can develop new estimators for the intrinsic dimension m.

For example, by choosing $g(u) = 1$, then functional (1.1) can be estimated by the number of points falling into $B(\boldsymbol{y}_0, r)$. This is the motivation behind correlation dimension methods [34, 49]. If r is chosen adaptively according to the distance from $\boldsymbol{y}_0$ to its k-nearest neighbor, $T_k(\boldsymbol{y}_0)$, then (1.1) is given by k/n, the proportion of

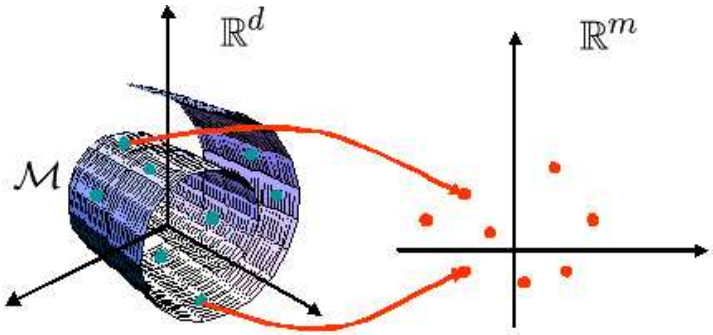


Figure 1.2: Finding a low-dimensional embedding of a high-dimensional set.

samples within a radius $T_k(\boldsymbol{y}_0)$ of $\boldsymbol{y}_0$. This is the starting point for earlier methods for estimating intrinsic dimension based on k-nearest neighbor distances [82].

In [62], a similar approach is followed, but the (binomial) number of points falling in $B(\boldsymbol{y}_0, T_k(\boldsymbol{y}_0))$ is approximated by a Poisson process, for samples uniformly distributed over the manifold. Then, the intrinsic dimension is estimated by maximum likelihood, resulting in the following estimate:

$$\hat{m}_{\boldsymbol{y}_0} = \frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{T_k(\boldsymbol{y}_0)}{T_j(\boldsymbol{y}_0)} \ .$$

1.2.3 Dimensionality Reduction

The general problem of dimensionality reduction consists of finding a mapping φ from a high-dimensional space $\mathcal{X}$ into a low-dimensional space $\mathcal{Y}$, such that the new representation of the data points in $\mathcal{Y}$ is simpler, but their description still preserves the important information about the data for the task intended. See figure 1.2. By itself, the problem of dimensionality reduction is ill-posed and some additional constraints have to be imposed on φ. Depending on the application, different characteristics of the data should be preserved, resulting in a set of possible constraints.

The classical approach to dimensionality reduction is based on projecting the

data on lower dimensional linear subspaces, while trying not to discard too much information. Principal component analysis and multidimensional scaling [44] achieve this by choosing the subspaces that keep most of the variance in the observed data set. Of course this will only provide a good representation of the original data set if, in fact, relationships between different coordinates of the data are constrained to linear functions. Kernel PCA [97] tries to circumvent this problem by adding extra freedom for including linear combinations from a basis of nonlinear functions.

One way of explicitly allowing for nonlinear dependencies among the data, is to assume that the data set lies on a lower dimensional manifold embedded in the original high-dimensional space. The recent papers of Tenenbaum *et al* [103] and Roweis and Saul [94] emphasized the usefulness of this approach, popularizing a new research field within the machine learning community and coining the term *manifold learning*. To keep the problem the most general possible in manifold learning, a fully nonparametric approach is followed. In particular, one wishes to estimate the data manifold or its lower dimensional embedding using only a finite number of points randomly sampled from the manifold. Within the class of manifold learning algorithms, many methods have been proposed in the past five years. They range from Locally Linear Embedding (LLE) [94], Laplacian Eigenmaps [6], Hessian Eigenmaps (HLLE) [26], Local Space Tangent Analysis [111], ISOMAP [103], or Semidefinite Embedding (SDE) [107].

1.3 Contributions of Thesis

This thesis presents a unified framework for the analysis of high-dimensional data sets through the use of random graphs. Originally motivated by computational and implementation issues, we propose the use of graph constructions like minimal span-

ning trees (MST) or k-nearest neighbor (k-NN) graphs to encode both statistical and spatial information and address the problems of dimension reduction and structure discovery in high-dimensional data. This leads to the theoretical study of the properties of such objects, like almost sure convergence of certain functionals of the graph, convergence rates, etc. Closing the loop, we use these results to develop practical estimators of quantities of interest, with provable statistical consistency.

Chapter 2 is concerned with power-weighted weight functionals associated with a minimal graph spanning a random sample of n points from a general multivariate Lebesgue density f over $[0,1]^d$. It is known that under broad conditions, the log of the normalized functional is a strongly consistent estimator of the Rényi α-entropy. We derive $\mathcal{L}_p$-norm (r.m.s. for $p=2$) convergence rates of this functional. In particular, we show that over the space of compacted supported multivariate densities f such that $f \in \Sigma_d(\beta, L)$ (the space of Hölder continuous functions), $0 < \beta \leq 1$, the $\mathcal{L}_p$-norm convergence rate is bounded above by $O\left(n^{-\alpha\beta/(\alpha\beta+1)\ 1/d)}\right)$. We obtain similar rate bounds for minimal graph approximations implemented by a progressive divide-and-conquer partitioning heuristic. We also obtain asymptotic lower bounds for the respective rates of convergence, using minimax techniques from nonparametric function estimation.

In Chapter 3, we study data sets that span a high-dimensional space but which contain fundamental features that are concentrated on lower-dimensional subsets of this space – curves, surfaces or, more generally, lower-dimensional manifolds. We extend the convergence results for minimal graphs from Euclidean spaces to general Riemannian manifolds. In particular, we develop a novel geometric probability approach to the problem of estimating intrinsic dimension and entropy of manifold data, based on asymptotic properties of graphs such as MST or k-NN graphs. Unlike previous solutions to this problem, we are able to prove statistical consistency

of the obtained estimators under weak assumptions of compactness of the manifold and boundedness of the (Lebesgue) sampling density supported on the manifold. The validity of these algorithms is shown by applying them to real data, such as high-dimensional image databases of faces and handwritten digits.

Complementary to Chapter 3, Chapter 4 addresses the problem of finding appropriate "compact" representations for high-dimensional data, particularly suited for classification tasks. Using k-NN graphs to encode the similarity between data points, we formulate the nonlinear dimensionality reduction problem as global quadratic optimization. To regularize the problem, class dependent constraints are added to the standard geometric constraints. This results in a framework that is applicable to both supervised and semi-supervised learning problems.

In Chapter 5, we shift the focus from using random graphs to extract properties of the sample distribution to using random graphs to extract spatial information. In particular, this framework is useful for distributed data gathering networks, like wireless sensor networks, where adjacency graphs can be used to model the spatial dependencies among the data. We introduce a distributed weighted multidimensional scaling algorithm that is applied to node localization in sensor networks. The proposed algorithm naturally accounts for communication constraints in a network scenario and derived bounds on communication costs show its superiority versus a centralized approach. For the localization problem, using received signal-strength (RSS) based range measurements, we demonstrate via simulation that location estimates are nearly unbiased with variance close to the Cramér-Rao lower bound. Further, RSS and time-of-arrival (TOA) channel measurements are used to demonstrate performance as good as the centralized maximum-likelihood estimator (MLE) in a real-world sensor network.

1.4 List of Relevant Publications

The following publications are a product of the research presented in this thesis.

(1) Jose A. Costa and Alfred O. Hero, "Learning intrinsic dimension and entropy of high-dimensional shapes," to appear in *Statistics and analysis of shapes*, Eds. H. Krim and T. Yezzi, Birkhäuser, 2005.

(2) J. A. Costa, N. Patwari and A. O. Hero, "Distributed Localization in Sensor Networks using Adaptive Multidimensional Scaling," invited paper in *Joint Statistical Meeting*, Minneapolis, August, 2005.

(3) J. A. Costa, A. Girotra and A. O. Hero, "Estimating Local Intrinsic Dimension with k-Nearest Neighbor Graphs," in *IEEE Workshop on Statistical Signal Processing*, Bordeaux, France, July, 2005.

(4) J. A. Costa, N. Patwari and A. O. Hero, "Distributed Multidimensional Scaling for Node Localization in Sensor Networks," accepted for publication in *ACM Trans. on Sensor Networks*, June, 2005.

(5) J. A. Costa and A. O. Hero, "Classification Constrained Dimensionality Reduction," in *Proc. of IEEE Int. Conf. on Acoust. Speech and Sig. Proc.*, Philadelphia, March, 2005.

(6) J. A. Costa, N. Patwari and A. O. Hero, "Achieving High-Accuracy Distributed Localization in Sensor Networks," in *Proc. of IEEE Int. Conf. on Acoust. Speech and Sig. Proc.*, Philadelphia, March, 2005. (Nominated for Best Student Paper Award)

(7) J. A. Costa and A. O. Hero, "Learning Intrinsic Dimension and Intrinsic Entropy of High-Dimensional Datasets," in *Proc. of European Sig. Proc. Conference (EUSIPCO)*, Vienna, Austria, September, 2004.

(8) J. A. Costa and A. O. Hero, "Geodesic Entropic Graphs for Dimension and Entropy Estimation in Manifold Learning," *IEEE Trans. on Signal Processing*, vol. 52, no. 8, pp. 2210-2221, August, 2004.

(9) J. A. Costa and A. O. Hero, "Entropic Graphs for Intrinsic Dimension Estimation in Manifold Learning," in *Proc. of IEEE Int. Symposium on Information Theory*, Chicago, July, 2004.

(10) J. A. Costa and A. O. Hero, "Manifold Learning Using k-Nearest Neighbor Graphs," in *Proc. of IEEE Int. Conf. on Acoust. Speech and Sig. Proc.*, Montreal, Canada, May, 2004.

(11) J. A. Costa and A. O. Hero, "Entropic Graphs for Manifold Learning," in *Proc. of IEEE Asilomar Conf. on Signals, Systems, and Computers*, Pacific Groove, CA, November, 2003.

(12) A. O. Hero, J. Costa and B. Ma, "Asymptotic relations between minimal graphs and α-entropy," Technical Report CSPL-334, Communications and Signal Processing Laboratory, The University of Michigan, 48109-2122, March, 2003.

CHAPTER 2

Minimal Euclidean Graphs and Entropy Estimation

2.1 Introduction

Consider a set of n points, $\mathcal{X}_n = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$, obtained by randomly sampling from a probability density f supported on the d-dimensional unit cube, $[0,1]^d$. For example, each point $\boldsymbol{X}_i$ can be a d-dimensional feature vector or an image with a total of d-pixels (organized in the usual lexicographical order).

Often, the solution to several problems of practical interest involves considering the points in $\mathcal{X}_n$ as vertices of a graph that captures the behavior of the system under study. A graph is defined by its set of vertices, a subset of edges, E, of the set of all $O(n^2)$ possible edges connecting every pair of vertices in the graph, and a function w that assigns weight $w(e)$ to edge $e \in E$.

A particular case of graph constructions that will be used throughout this thesis are *minimal graphs*. These graphs are usually constructed by solving an optimization problem aimed at finding a set of edges with total minimum weight, over a class of

allowable graphs. More formally, this optimization problem can be written as

$$L(\mathcal{X}_n) = \min_{E \in \mathcal{E}(\mathcal{X}_n)} \sum_{e \in E} w(e)\,, \tag{2.1}$$

where $\mathcal{E}(\mathcal{X}_n)$ is the class of allowable graphs over $\mathcal{X}_n$, specified by the constraints in the original problem. $L(\mathcal{X}_n)$ is called the total edge weight functional of class $\mathcal{E}(\mathcal{X}_n)$. When $w(e) = |e|^\gamma$, where $|e|$ is the Euclidean distance between the two vertices that edge e connects and $\gamma > 0$ is an *edge exponent* or *power-weighting constant*, the resulting graphs are called *minimal Euclidean graphs*. Examples of such graphs constructions include, among others:

- the Euclidean traveling salesman problem (TSP). In the TSP the objective is to find a graph of minimum weight among the set $\mathcal{C}$ of graphs that have exactly one cycle that visits each point in $\mathcal{X}_n$ once. The resultant graph is called the *minimal TSP tour* and its total edge weight functional is $L_\gamma^{\mathrm{TSP}}(\mathcal{X}_n) = \min_{C \in \mathcal{C}} \sum_{e \in C} |e|^\gamma$. Construction of the TSP graph is NP-hard and arises in many different areas of operations research [59].

- the minimal spanning tree (MST). In the MST problem the objective is to find a graph of minimum weight among the graphs $\mathcal{T}$ which span the sample $\mathcal{X}_n$. This problem admits exact solutions which run in polynomial time and the total edge weight functional of the MST is $L_\gamma^{\mathrm{MST}}(\mathcal{X}_n) = \min_{T \in \mathcal{T}} \sum_{e \in T} |e|^\gamma$. MST's arise in areas including: pattern recognition [104]; clustering [110]; nonparametric regression [3] and testing for randomness [42].

- the k-nearest neighbor graph (k-NNG). The k-NNG problem consists of finding the set $\mathcal{N}_{k,i}$ of k-nearest neighbors of each point X_i in the set $\mathcal{X}_n - \{X_i\}$. This problem has exact solutions which run in linear-log-linear time and the total

edge weight functional is $L_\gamma^{k-\mathrm{NNG}}(\mathcal{X}_n) = \sum_{i=1}^n \sum_{e \in \mathcal{N}_{k,i}} |e|^\gamma$. The k-NNG arises in computational geometry [24], clustering and pattern recognition [93], spatial statistics [22], and adaptive vector quantization [31].

Due to the key role that the total edge weight functional plays in the construction of minimal graphs, it is of great interest to characterize its behavior. It has long been known that, under the assumption of n independent identically distributed (i.i.d.) vertices in $[0,1]^d$, the (suitably normalized) total edge weight functional of certain minimal Euclidean graphs converges almost surely (a.s.), as the number of vertices increases, to a limit which is a monotone function of the Rényi entropy of the multivariate density f of the random vertices. Recall that the Rényi entropy or α-entropy is defined as

$$H_\alpha(f) = \frac{1}{1-\alpha} \log \int f^\alpha(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \ .$$

Graph constructions that satisfy this convergence property include the TSP, MST, k-NNG or the minimal matching graph (MMG), and their power-weighted variants. See the recent books by Steele [100] and Yukich [109] for introduction to this subject. An $O(n^{-1/d})$ bound on the a.s. convergence rate of the normalized weight functional of these and other minimal graphs was obtained by Redmond and Yukich [90, 91] when the vertices are uniformly distributed over $[0,1]^d$.

In the present chapter we obtain bounds on $\mathcal{L}_p$-norm (r.m.s. for $p=2$) convergence rates of power-weighted Euclidean total edge weight functionals for Lebesgue densities f over $[0,1]^d$, for which $f \in \Sigma_d(\beta, L)$, the space of Hölder continuous functions with Lipschitz constant L and $0 < \beta \leq 1$, and $f^{\frac{1}{2}-\frac{\gamma}{d}}$ is integrable. Here the integer dimension d is greater than one and $\gamma \in (1,d)$ is an edge exponent which is incorporated in the weight functional to taper the Euclidean distance between ver-

tices of the graph (see next section for definitions). As a special case of Proposition 6, we obtain a $O\left(n^{-\alpha\beta/(\alpha\beta+1)\,1/d)}\right)$ upper bound on the r.m.s. convergence. This bound implies a slower rate of convergence than the analogous $O(n^{-1/d})$ rate bound proven for uniform f by Redmond and Yukich [90, 91]. Furthermore, the rate constants derived here suggest that slower convergence occurs when either the (Rényi) entropy of the underlying density f or the Lipschitz constant L is large. We also derive lower bounds on the respective convergence rates by recasting the problem as that of estimating the Rényi entropy, or equivalently $\int f^\alpha(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, over the non-parametric class of densities $f \in \Sigma_d(\beta, L)$. For this, we use standard minimax techniques from non-parametric function estimation. Corollary 13 constitutes the main result of this chapter in the form of upper and lower bounds on the rates of convergence of any continuous quasi-additive Euclidean functional.

We also obtain $\mathcal{L}_p$-norm convergence rate bounds for partitioned approximations to minimal graphs implemented by the following fixed partitioning heuristic: 1) dissect $[0,1]^d$ into a set of m^d cells of equal volumes $1/m^d$; 2) compute minimal graphs spanning the points in each non-empty cell; 3) stitch together these small graphs to form an approximation to the minimal graph spanning all of the points in $[0,1]^d$. Such heuristics have been widely adopted, e.g. see Karp [46], Ravi *et al.* [88], and Hero and Michel [39], for examples. The computational advantage of this partitioning heuristic comes from its divide-and-conquer progressive-resolution strategy to an optimization whose complexity is non-linear in n: the partitioned algorithm only requires constructing minimal graphs on small cells, each of which typically contains far fewer than n points. In Proposition 8 we obtain bounds on $\mathcal{L}_p$-norm convergence rate and specify an optimal "progressive-resolution sequence" $m = m(n)$, $n = 1, 2, \ldots$, for achieving these bounds.

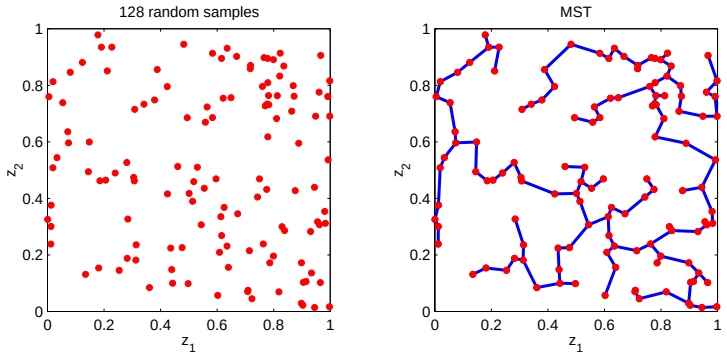
A principal focus of our research on minimal graphs has been on the use of weight

functionals for signal processing applications such as image registration, nonlinear dimensionality reduction, pattern matching and non-parametric entropy estimation, see e.g. [18, 39–41, 70]. Beyond the signal processing applications mentioned above these results may have important practical implications in adaptive vector quantizer design, where the Rényi entropy is more commonly called the Panter-Dite factor and is related to the asymptotically optimal quantization cell density [30, 71]. Furthermore, as empirical versions of vector quantization can be cast as geometric location problems [33], the asymptotics of adaptive VQ may be studied within the present framework of minimal Euclidean graphs. Other applications of the convergence rate results of this chapter include classical problems in Euclidean optimization theory, computational geometry and operations research; for further details see [100] and [109].

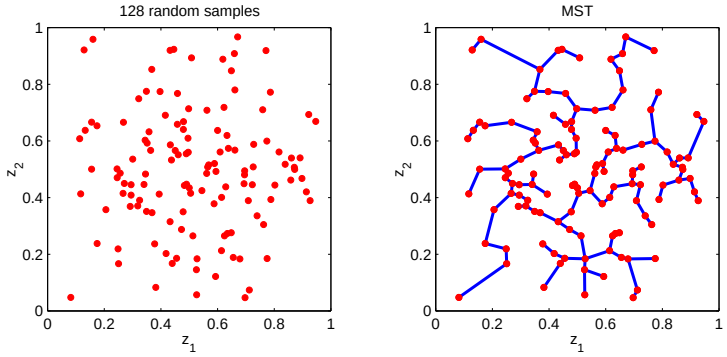
The outline of this chapter is as follows. In Section 2.2 we briefly review Redmond and Yukich's unifying framework of continuous quasi-additive power-weighted edge functionals. In Section 2.3 we give convergence rate upper bounds for such functionals with general Holder continuous density f. In Section 2.4 we extend these results to partitioned approximations. In Section 2.5 we derive lower bounds to the convergence rates. Finally, in Section 2.6 we make a brief comment about nonparametric estimation of the Rényi entropy. We also give an extension of the convergence rate upper bounds to densities in a Sobolev class in Section 2.9.

2.2 Minimal Euclidean Graphs

Since the seminal work of Beardwood, Halton and Hammersley in 1959 [4], the asymptotic behavior of the total edge weight functional of a minimal graph such as the MST and the TSP over i.i.d. random points $\mathcal{X}_n = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ as $n \to \infty$ has



(a) Uniform distribution.



(b) Separable triangular distribution.

Figure 2.1: The MST on a random set of $n = 128$ samples in $[0, 1]^2$.

been of great interest. See Figures 2.1 and 2.2 for an illustration of this behavior.

The monographs by Steele [100] and Yukich [109] provide two engaging presentations of ensuing research in this area. Many of the convergence results have been encapsulated in the general framework of continuous and quasi-additive Euclidean functionals recently introduced by Redmond and Yukich [90]. This framework allows one to relatively simply obtain asymptotic convergence rates once a graph total edge weight functional has been shown to satisfy the required continuity and quasi-additivity properties. We follow this framework in this chapter.

Let F be a finite subset of points in $[0, 1]^d, d \geq 2$. A real-valued function L_γ

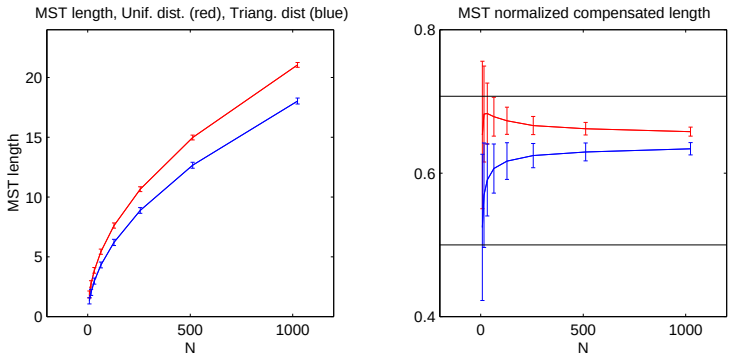


Figure 2.2: Total edge weight functional of MST (left) and weight functional divided by $\sqrt{n}$ (right) as a function of the number of samples n, for the uniform and separable triangular distribution in $[0,1]^d$.

defined on F is called an *Euclidean functional of order* γ if it satisfies:

1. *Translation invariance*: $\forall\, \boldsymbol{y} \in \mathbb{R}^d\,,\ L_\gamma(F) = L_\gamma(F + \boldsymbol{y})$.

2. *Homogeneity of order* γ: $\forall\, \alpha > 0\,,\ L_\gamma(\alpha\, F) = \alpha^\gamma L_\gamma(F)$.

2.2.1 Continuous Quasi-additive Euclidean Functionals

Intuitively, a weight functional $L_\gamma(\mathcal{X}_n)$ of a minimal graph on $[0,1]^d$ is a "continuous quasi-additive" functional if it can be closely approximated by the the sum of the weight functionals of minimal graphs constructed on a dense partition of $[0,1]^d$. The following technical conditions on a Euclidean functional L_γ were defined in [90, 109].

- *Null condition*: $L_\gamma(\phi) = 0$, where ϕ is the null set.

- *Subadditivity*: Let $\mathcal{Q}^m = \{Q_i\}_{i=1}^{m^d}$ be a uniform partition of $[0,1]^d$ into m^d subcubes Q_i with edges parallel to the coordinate axes having edge lengths m^{-1} and volumes m^{-d} and let $\{q_i\}_{i=1}^{m^d}$ be the set of points in $[0,1]^d$ that translate each Q_i back to the origin such that $Q_i - q_i$ has the form $m^{-1}[0,1]^d$. Then there exists a constant C_1 with the following property: for every finite subset

F of $[0,1]^d$

$$L_\gamma(F) \leq m^{-\gamma} \sum_{i=1}^{m^d} L_\gamma\left(m[F \cap Q_i - q_i]\right) + C_1 m^{d-\gamma} \tag{2.2}$$

- *Superadditivity*: For the same conditions as above on Q_i, m, and q_i,

$$L_\gamma(F) \geq m^{-\gamma} \sum_{i=1}^{m^d} L_\gamma\left(m[F \cap Q_i - q_i]\right) \tag{2.3}$$

- *Continuity*: There exists a constant C_2 such that for all finite subsets F and G of $[0,1]^d$,

$$|L_\gamma(F \cup G) - L_\gamma(F)| \leq C_2(\mathrm{card}(G))^{(d-\gamma)/d}, \tag{2.4}$$

where $\mathrm{card}(G)$ is the cardinality of the subset G. Note that continuity implies

$$|L_\gamma(F) - L_\gamma(G)| \leq 2\, C_2(\mathrm{card}(F \triangle G))^{(d-\gamma)/d}, \tag{2.5}$$

where $F \triangle G = (F \cup G) \setminus (F \cap G)$ denotes the symmetric difference of sets F and G.

The functional L_γ is said to be a *continuous subadditive functional* of order γ if it satisfies the null condition, sudadditivity and continuity. L_γ is said to be a *continuous superadditive functional* of order γ if it satisfies the null condition, superadditivity and continuity.

For many continuous subadditive functionals L_γ on $[0,1]^d$ there exists a *dual* or *boundary* superadditive functional L_γ^*. The dual functional satisfies two properties: 1) $L_\gamma(F) + 1 \geq L_\gamma^*(F)$ for every finite subset F of $[0,1]^d$; and, 2) for i.i.d. uniform

random vectors $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n$ over $[0,1]^d$,

$$\left|E[L_\gamma(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n)] - E[L_\gamma^*(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n)]\right| \leq C_3 n^{(d-\gamma-1)/d} \tag{2.6}$$

with C_3 a finite constant. The condition (2.6) is called the *close-in-mean approximation* in [109].

A stronger condition which is useful for showing convergence of partitioned approximations is the *pointwise closeness* condition

$$\left|L_\gamma(F) - L_\gamma^*(F)\right| \leq o\left([\mathrm{card}(F)]^{(d-\gamma)/d}\right), \tag{2.7}$$

for any finite subset F of $[0,1]^d$.

A continuous subadditive functional L_γ is said to be a *continuous quasi-additive functional* if L_γ is continuous subadditive and there exists a continuous superadditive dual functional L_γ^*. We point out that the dual L_γ^* is not uniquely defined. It has been shown by Redmond and Yukich [90, 91] that the boundary-rooted version of L_γ, namely, one where edges may be connected to the boundary of the unit cube over which they accrue zero weight, usually has the requisite property (2.6) of the dual. These authors have displayed duals and shown continuous quasi-additivity and related properties for total edge weight functionals of the power weighted MST, Steiner tree, TSP, k-NNG and others.

Independently of its specific definition, once an Euclidean functional is shown to verify the continuous quasi-additive properties, its asymptotic behavior follows immediately from an *umbrella theorem*:

Theorem ([109, Theorem 7.1]). *Let $d \geq 2$ and $1 \leq \gamma \leq d-1$. Assume $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. sample points over $[0,1]^d$ with Lebesgue density f. Then, for*

any continuous quasi-additive Euclidean functional L_γ of order γ,

$$\lim_{n\to\infty} L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]/n^{(d-\gamma)/d} = \beta_{L_\gamma,d}\int_{[0,1]^d} f^{(d-\gamma)/d}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

almost surely, where $\beta_{L_\gamma,d}$ is a constant independent of the distribution of $\{\boldsymbol{X}_i\}$. Furthermore, the mean length $E[L_\gamma(\mathcal{X}_n)]/n^{(d-\gamma)/d}$ converges to the same limit.

In [90, 109] almost sure limits with a convergence rate upper bound of $O\left(n^{-1/d}\right)$ were obtained for continuous quasi-additive Euclidean functionals $L_\gamma(\boldsymbol{U}_1,\ldots,\boldsymbol{U}_n)$ under the assumption of uniformly distributed points $\boldsymbol{U}_1,\ldots,\boldsymbol{U}_n$ and an additional assumption that L_γ satisfies the "add-one bound"

- *Add-one bound*:

$$\mid E[L_\gamma(\boldsymbol{U}_1,\ldots,\boldsymbol{U}_{n+1})] - E[L_\gamma(\boldsymbol{U}_1,\ldots,\boldsymbol{U}_n)] \mid \leq C_5 n^{-\gamma/d}. \tag{2.8}$$

The MST length functional of order γ satisfies the add-one bound. A slightly weaker bound on a.s. convergence rate also holds when L_γ is merely continuous quasi-additive [109, Ch.5]. The $n^{-1/d}$ convergence rate bound is exact for $d = 2$.

2.3 Convergence Rate Upper Bounds for General Density

In this section we obtain upper bounds on the rate of convergence of $E[L_\gamma(\mathcal{X}_n)]/n^{(d-\gamma)/d}$ to its asymptotic limit, for points sampled from a probability distribution with Hölder continuous Lebesgue density. For convenience we will focus on the case that L_γ is continuous quasi-additive and satisfies the add-one bound, although some of the following results can be established under weaker assumptions. Our method of ex-

tension follows common practice [99, 100, 109]: we first establish convergence rates of the mean $E[L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)]/n^{(d-\gamma)/d}$ for piecewise constant densities and then extend to arbitrary densities. Then we use a concentration inequality to obtain $\mathcal{L}_p$-norm convergence rates of $L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)/n^{(d-\gamma)/d}$.

2.3.1 Mean Convergence Rate for Block Densities

We will need the following elementary result for the sequel, whose proof is given in Section 2.8.

Lemma 1. *Let $g(u)$ be a continuously differentiable function of $u \in \mathbb{R}$ which is concave and monotone increasing over $u \geq 0$. Then for any $u_o > 0$*

$$g(u_o) - \frac{g(u_o)}{u_o}|\Delta| \leq g(u) \leq g(u_o) + g^{'}(u_o)|\Delta| \ ,$$

where $\Delta = u - u_o$ and $g^{'}(u) = dg(u)/du$.

A density $f(\boldsymbol{x})$ over $[0,1]^d$ is said to be a block density with m^d levels if for some set of non-negative constants $\{\phi_i\}_{i=1}^{m^d}$ satisfying $\sum_{i=1}^{m^d} \phi_i m^{-d} = 1$,

$$f(\boldsymbol{x}) = \sum_{i=1}^{m^d} \phi_i 1_{Q_i}(\boldsymbol{x})$$

where $1_Q(\boldsymbol{x})$ is the set indicator function of $Q \subset [0,1]^d$ and $\{Q_i\}_{i=1}^{m^d}$ is the uniform partition of the unit cube $[0,1]^d$ defined above.

Proposition 2. *Let $d \geq 2$ and $1 \leq \gamma \leq d-1$. Assume $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. sample points over $[0,1]^d$ whose marginal is a block density f with m^d levels and support $\mathcal{S} \subset [0,1]^d$. Then for any continuous quasi-additive Euclidean functional L_γ of order*

γ that satisfies the add-one bound (2.8)

$$\left| E[L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d} \int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \right| \leq O\left((nm^{-d})^{-1/d}\right) \ ,$$

where $\beta_{L_\gamma,d}$ is a constant independent of f. A more explicit form for the bound on the right hand side is

$$O\left((nm^{-d})^{-1/d}\right) = \begin{cases} \frac{K_1+C_4}{(nm^{-d})^{1/d}} \int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}\ (1+o(1))\,, & d > 2 \\ \\ \frac{K_1+C_4+\beta_{L_\gamma,d}}{(nm^{-d})^{1/d}} \int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}\ (1+o(1))\,, & d = 2 \end{cases} .$$

Proof. Let n_i denote the number of samples $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ falling into the partition cell Q_i and let $\{\boldsymbol{U}_i\}_i$ denote an i.i.d. sequence of uniform points on $[0,1]^d$. By subadditivity, we have

$$\begin{aligned} L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) &\leq m^{-\gamma} \sum_{i=1}^{m^d} L_\gamma\left(m[\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} \cap Q_i - q_i]\right) + C_1 m^{d-\gamma} \\ &= m^{-\gamma} \sum_{i=1}^{m^d} L_\gamma(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_{n_i}) + C_1 m^{d-\gamma} \end{aligned}$$

since the samples in each partition cell Q_i are drawn independently from a conditionally uniform distribution given n_i. Note that n_i has a Binomial $B(n, \phi_i m^{-d})$ distribution.

Taking expectations on both sides of the above inequality,

$$E[L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)] \leq m^{-\gamma} \sum_{i=1}^{m^d} E\left[E\left[L_\gamma(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_{n_i}) | \, n_i\right]\right] + C_1 m^{d-\gamma}. \quad (2.9)$$

The following rate of convergence for quasi-additive edge functionals L_γ satisfying

the add-one bound (2.8) has been established for $1 \leq \gamma < d$ [109, Thm. 5.2],

$$|E[L_\gamma(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n)] - \beta_{L_\gamma,d} n^{\frac{d-\gamma}{d}}| \leq K_1 n^{\frac{d-1-\gamma}{d}}, \tag{2.10}$$

where K_1 is a function of C_1, C_3 and C_5.

Using the result (2.10) and subadditivity (2.9) on L_γ, for $1 \leq \gamma < d$ we have

$$\begin{aligned} E[L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)] &\leq m^{-\gamma} \sum_{i=1}^{m^d} E\left[\beta_{L_\gamma,d}\, n_i^{\frac{d-\gamma}{d}} + K_1\, n_i^{\frac{d-\gamma-1}{d}}\right] + C_1 m^{d-\gamma} \\ &= m^{-\gamma}\beta_{L_\gamma,d}\, n^{\frac{d-\gamma}{d}} \sum_{i=1}^{m^d} E\left[\left(\frac{n_i}{n}\right)^{\frac{d-\gamma}{d}}\right] + m^{-\gamma} K_1\, n^{\frac{d-\gamma-1}{d}} \sum_{i=1}^{m^d} E\left[\left(\frac{n_i}{n}\right)^{\frac{d-\gamma-1}{d}}\right] \\ &\quad + C_1 m^{d-\gamma}. \end{aligned} \tag{2.11}$$

Similarly for the dual L_γ^* it follows by superadditivity (2.3) and the close-in-mean condition (2.6)

$$\begin{aligned} &E[L_\gamma^*(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)] \\ &\geq m^{-\gamma}\beta_{L_\gamma,d}\, n^{\frac{d-\gamma}{d}} \sum_{i=1}^{m^d} E\left[\left(\frac{n_i}{n}\right)^{\frac{d-\gamma}{d}}\right] - m^{-\gamma}(K_1 + C_3)\, n^{\frac{d-\gamma-1}{d}} \sum_{i=1}^{m^d} E\left[\left(\frac{n_i}{n}\right)^{\frac{d-\gamma-1}{d}}\right], \end{aligned} \tag{2.12}$$

for $1 \leq \gamma < d$.

We next develop lower and upper bounds on the expected values in (2.11) and (2.12). As the function $g(u) = u^\nu$ is monotone and concave over the range $u \geq 0$ for $0 < \nu < 1$, from Lemma 1

$$\left(\frac{n_i}{n}\right)^\nu \geq p_i^\nu - p_i^{\nu-1}\left|\frac{n_i}{n} - p_i\right|, \tag{2.13}$$

where $p_i = \phi_i m^{-d}$. In order to bound the expectation of the above inequality we use

the following bound

$$E\left[\left|\frac{n_i}{n}-p_i\right|\right] \leq \sqrt{E\left[\left|\frac{n_i}{n}-p_i\right|^2\right]} \leq \frac{\sqrt{p_i}}{\sqrt{n}}.$$

Therefore, from (2.13),

$$E\left[\left(\frac{n_i}{n}\right)^\nu\right] \;\geq\; p_i^\nu - p_i^{\nu-\frac{1}{2}}/\sqrt{n}. \tag{2.14}$$

By concavity, Jensen's inequality yields the upper bound

$$E\left[\left(\frac{n_i}{n}\right)^\nu\right] \leq \left[E\left(\frac{n_i}{n}\right)\right]^\nu = p_i^\nu \tag{2.15}$$

Under the hypothesis $1 \leq \gamma \leq d-1$ this upper bound can be substituted into expression (2.11) to obtain

$$\begin{aligned}
&E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)/n^{(d-\gamma)/d}] \\
&\quad\leq \beta_{L_\gamma,d}\sum_{i=1}^{m^d}\phi_i^{\frac{d-\gamma}{d}}m^{-d} + \frac{K_1}{(nm^{-d})^{1/d}}\sum_{i=1}^{m^d}\phi_i^{\frac{d-\gamma-1}{d}}m^{-d} + \frac{C_1}{(nm^{-d})^{(d-\gamma)/d}} \\
&\quad= \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \frac{K_1}{(nm^{-d})^{1/d}}\int_{\mathcal{S}} f^{(d-\gamma-1)/d}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \frac{C_1}{(nm^{-d})^{(d-\gamma)/d}}\,.
\end{aligned} \tag{2.16}$$

Applying the bounds (2.15) and (2.14) to (2.12) we obtain an analogous lower bound for the mean of the dual functional L_γ^*

$$\begin{aligned}
&E[L_\gamma^*(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]/n^{(d-\gamma)/d} \\
&\quad\geq \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{\frac{d-\gamma}{d}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \frac{\beta_{L_\gamma,d}}{(nm^{-d})^{1/2}}\int_{\mathcal{S}} f^{\frac{1}{2}-\frac{\gamma}{d}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \frac{K_1+C_3}{(nm^{-d})^{1/d}}\int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,.
\end{aligned} \tag{2.17}$$

By definition of the dual,

$$E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]/n^{\frac{d-\gamma}{d}} \geq E[L^*_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]/n^{\frac{d-\gamma}{d}} - n^{-\frac{d-\gamma}{d}} \tag{2.18}$$

which when combined with (2.17) and (2.16) yields the result

$$\begin{aligned}
&\left|\frac{E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]}{n^{\frac{d-\gamma}{d}}} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{\frac{d-\gamma}{d}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right| \\
&\quad\leq \frac{K_1+C_3}{(nm^{-d})^{1/d}}\int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \frac{\beta_{L_\gamma,d}}{(nm^{-d})^{1/2}}\int_{\mathcal{S}} f^{\frac{1}{2}-\frac{\gamma}{d}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\
&\qquad + \frac{C_1}{(nm^{-d})^{(d-\gamma)/d}} + n^{-\frac{d-\gamma}{d}}\,.
\end{aligned} \tag{2.19}$$

This establishes Proposition 2. □

2.3.2 Mean Convergence Rate for Hölder Continuous Density Functions

To establish upper and lower bounds we adopt the setting of Hölder continuous density functions.

Recall that the Hölder continuous class $\Sigma_d(\beta, L)$ is defined by [53]

$$\Sigma_d(\beta,L) = \left\{g : |g(\boldsymbol{z}) - p^{\lfloor\beta\rfloor}_{\boldsymbol{x}}(\boldsymbol{z})| \leq L\,|\boldsymbol{x}-\boldsymbol{z}|^\beta, \boldsymbol{x},\boldsymbol{z}\in\mathbb{R}^d\right\}$$

where $p^k_{\boldsymbol{x}}(\boldsymbol{z})$ is the Taylor polynomial (multinomial) of g of order k expanded about the point $\boldsymbol{x}$, $|.|$ denotes a norm in $\mathbb{R}^d$ and $\lfloor\beta\rfloor$ is defined as the greatest integer strictly less than β. $\Sigma_d(1,L)$ is the set of Lipschitz functions with Lipschitz constant L and $\Sigma_d(\beta,L)$ contains increasingly smooth functions as β increases.

Before extending Proposition 2 to this setting we will need to establish an approximation lemma for Hölder continuous functions.

For $\mathcal{Q}^m = \{Q_i\}_{i=1}^{m^d}$ a uniform resolution-m partition as defined in Sub-section 2.2.1, define the resolution-m block density approximation $\phi(\boldsymbol{x}) = \sum_{i=1}^{m^d} \phi_i 1_{Q_i}(\boldsymbol{x})$ of f, where $\phi_i = m^d \int_{Q_i} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. The following lemma establishes how close (in $L_1([0,1]^d)$ sense) these resolution-m block densities approximate functions in $\Sigma_d(\beta, L)$.

Lemma 3. *For $0 < \beta \leq 1$, let $f \in \Sigma_d(\beta, L)$ have support $\mathcal{S} \subset [0,1]^d$. Then there exists a constant $C_6 > 0$, independent of m, such that*

$$\int_{\mathcal{S}} |\phi(\boldsymbol{x}) - f(\boldsymbol{x})|\,\mathrm{d}\boldsymbol{x} \leq C_4\, L\, m^{-\beta}. \tag{2.20}$$

A proof of this lemma is given in Section 2.8.

Remark 1. Lemma 3 shows how close, in $L_1(\mathbb{R}^d)$ sense, a function $f \in \Sigma_d(\beta, L)$ can be approximated by its resolution-m block density approximation. To extend the results in this and the following sections to other classes of functions, all that it is needed is an upper bound to the L_1 approximation error similar to the one in (2.20). In Section 2.9, we show how to do this for densities in the Sobolev space $W^{1,p}(\mathbb{R}^d)$, $1 \leq p < \infty$. The importance of Sobolev spaces derives from the fact that they include functions that are not differentiable in the usual (strong) sense.

We can now return to the problem of finding convergence rate bounds on quasi-additive Euclidean functionals for non-uniform density f. Let $\{\tilde{\boldsymbol{X}}_i\}_{i=1}^n$ be i.i.d. random vectors having marginal Lebesgue density equal to the block density approxi-

mation ϕ. By the triangle inequality,

$$\begin{aligned}
&\left| E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]/n^{\frac{d-\gamma}{d}} - \beta_{L_\gamma,d}\int_S f^{\frac{d-\gamma}{d}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right| \qquad (2.21)\\
&\leq \left| E[L_\gamma(\tilde{\boldsymbol{X}}_1,\ldots,\tilde{\boldsymbol{X}}_n)]/n^{\frac{d-\gamma}{d}} - \beta_{L_\gamma,d}\int_S \phi^{\frac{d-\gamma}{d}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right|\\
&\quad + \beta_{L_\gamma,d}\left|\int_S \phi^{\frac{d-\gamma}{d}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \int_S f^{\frac{d-\gamma}{d}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right|\\
&\quad + \left|E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)] - E[L_\gamma(\tilde{\boldsymbol{X}}_1,\ldots,\tilde{\boldsymbol{X}}_n)]\right| /n^{\frac{d-\gamma}{d}}\\
&= I + II + III
\end{aligned}$$

Term I can be bounded by Proposition 2. To bound II, consider the following elementary inequality, which holds for $a, b \geq 0$, $0 \leq \gamma \leq d$,

$$\left|a^{(d-\gamma)/d} - b^{(d-\gamma)/d}\right| \leq |a-b|^{(d-\gamma)/d},$$

and therefore, by Lemma 3 and Jensen's inequality,

$$II \leq \beta_{L_\gamma,d}\int_{\mathcal{S}} |\phi(\boldsymbol{x}) - f(\boldsymbol{x})|^{\frac{d-\gamma}{d}}\,\mathrm{d}\boldsymbol{x} \leq \beta_{L_\gamma,d}\, C_4'\, L^{(d-\gamma)/d}\, m^{-\beta(d-\gamma)/d}, \qquad (2.22)$$

where $C_4' = C_4^{(d-\gamma)/d}$.

The following Lemma establishes an upper bound on term III in (2.21):

Lemma 4. *Let $d \geq 2$ and $1 \leq \gamma \leq d$. Assume $\{\boldsymbol{X}_i\}_{i=1}^n$ are i.i.d. random vectors over $[0,1]^d$ with density $f \in \Sigma_d(\beta, L)$, $0 < \beta \leq 1$, having support $\mathcal{S} \subset [0,1]^d$. Let $\{\tilde{\boldsymbol{X}}_i\}_{i=1}^n$ be i.i.d. random vectors with marginal Lebesgue density ϕ, the resolution-m block density approximation of f. Then, for any continuous quasi-additive Euclidean*

functional L_γ of order γ

$$\left| E[L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)] - E[L_\gamma(\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_n)] \right| / n^{\frac{d-\gamma}{d}} \leq C_2' \, C_4' \, L^{(d-\gamma)/d} \, m^{-\beta(d-\gamma)/d}, \tag{2.23}$$

where $C_2' = 2^{(2d-\gamma)/d} C_2$.

Proof. As in equation (2.21), we denote the left hand side of (2.23) by III. First invoke continuity (2.5) of L_γ

$$n^{(d-\gamma)/d} III \leq 2 \, C_2 \, E \left[\text{card} \left(\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} \triangle \{\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_n\} \right)^{(d-\gamma)/d} \right].$$

To bound the right hand side of the above inequality we use an argument which is discussed and proved in [99, Theorem 3]. There it is shown that if ϕ approximates f in the $L_1(\mathbb{R}^d)$ sense:

$$\int_{\mathcal{S}} |\phi(\boldsymbol{x}) - f(\boldsymbol{x})| \, \mathrm{d}\boldsymbol{x} \leq \varepsilon,$$

then, by standard coupling arguments, there exists a joint distribution P for the pair of random vectors $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$ such that $P\{\boldsymbol{X} \neq \tilde{\boldsymbol{X}}\} \leq \varepsilon$. It then follows by Lemma 3 and the set inequality $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} \triangle \{\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_n\} \subseteq \cup_{i=1}^n \{\boldsymbol{X}_i\} \triangle \{\tilde{\boldsymbol{X}}_i\}$ that

$$\begin{aligned} III &\leq 2 \, C_2 \, E \left[\text{card} \left(\cup_{i=1}^n \{\boldsymbol{X}_i\} \triangle \{\tilde{\boldsymbol{X}}_i\} \right)^{(d-\gamma)/d} \right] / n^{(d-\gamma)/d} \\ &\leq 2 \, C_2 \, E \left[\left(2 \sum_{i=1}^n 1_{\left\{ \boldsymbol{X}_i \neq \tilde{\boldsymbol{X}}_i \right\}} \right)^{(d-\gamma)/d} \right] / n^{(d-\gamma)/d} \\ &\leq 2 \, C_2 \, (2nP\{\boldsymbol{X}_1 \neq \tilde{\boldsymbol{X}}_1\})^{(d-\gamma)/d} / n^{(d-\gamma)/d} \leq 2^{(2d-\gamma)/d} C_2 \, \varepsilon^{(d-\gamma)/d}, \end{aligned}$$

where the second inequality follows from the fact $\text{card} \left(\{\boldsymbol{X}_i\} \triangle \{\tilde{\boldsymbol{X}}_i\} \right) \in \{0, 2\}$. Finally, by Lemma 3 we can make ε as small as $C_4 \, L \, m^{-\beta}$ and still ensure that ϕ be a block density approximation to f of resolution m. □

We can now substitute bounds (2.19), (2.22) and (2.23) in inequality (2.21) to obtain

$$\begin{aligned}&\left|E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f(\boldsymbol{x})^{(d-\gamma)/d}\,\mathrm{d}\boldsymbol{x}\right| \\ &\quad\le \frac{K_1+C_3}{(nm^{-d})^{1/d}}\left(\int_{\mathcal{S}} f^{\frac{d-1-\gamma}{d}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}+o(1)\right)+\frac{\beta_{L_\gamma,d}}{(nm^{-d})^{1/2}}\left(\int_{\mathcal{S}} f^{\frac{1}{2}-\frac{\gamma}{d}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}+o(1)\right) \\ &\quad+\frac{C_1}{(nm^{-d})^{(d-\gamma)/d}}+\frac{1}{n^{(d-\gamma)/d}}+(\beta_{L_\gamma,d}+C_2')\,C_4'\,L^{(d-\gamma)/d}\,m^{-\beta(d-\gamma)/d}\end{aligned} \tag{2.24}$$

This bound is finite under the assumptions that $f \in \Sigma_d(\beta, L)$ with support in $\mathcal{S} \subset [0,1]^d$ and that $f^{\frac{1}{2}-\frac{\gamma}{d}}$ is integrable over $\mathcal{S}$.

The bound (2.24) is actually a family of bounds for different values of $m = 1, 2, \ldots$. By selecting m as the function of n that minimizes this bound, we obtain the tightest bound among them:

Proposition 5. *Let $d \ge 2$ and $1 \le \gamma \le d-1$. Assume $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. random vectors over $[0,1]^d$ with density $f \in \Sigma_d(\beta, L)$, $0 < \beta \le 1$, having support $\mathcal{S} \subset [0,1]^d$. Assume also that $f^{\frac{1}{2}-\frac{\gamma}{d}}$ is integrable over $\mathcal{S}$. Then, for any continuous quasi-additive Euclidean functional L_γ of order γ that satisfies the add-one bound (2.8)*

$$\left|E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right| \le O\left(n^{-r_1(d,\gamma,\beta)}\right),$$

where

$$r_1(d,\gamma,\beta) = \frac{\alpha\,\beta}{\alpha\,\beta+1}\,\frac{1}{d}$$

where $\alpha = \frac{d-\gamma}{d}$.

Proof. Without loss of generality assume that $nm^{-d} > 1$. In the range $d \ge 2$ and $1 \le \gamma \le d-1$, the slowest of the rates in (2.24) are $(nm^{-d})^{-1/d}$ and $m^{-\beta(d-\gamma)/d}$. We obtain an m-independent bound by selecting $m = m(n)$ to be the sequence increasing

in n which minimizes the maximum of these rates

$$m(n) = \arg\min_m \max\left\{(nm^{-d})^{-1/d}, m^{-\beta(d-\gamma)/d}\right\}.$$

The solution $m = m(n)$ occurs when $(nm^{-d})^{-1/d} = m^{-\beta(d-\gamma)/d}$, or $m = n^{1/[d(\alpha\beta+1)]}$ (integer part) and, correspondingly, $m^{-\beta(d-\gamma)/d} = n^{-\frac{\alpha\beta}{\alpha\beta+1}\frac{1}{d}}$. This establishes Proposition 5. □

To convert the mean convergence bound in Proposition 3 to a $\mathcal{L}_p$ convergence bound requires application of a concentration inequality. Any Euclidean functional L_γ of order γ satisfying the continuity property (2.4) also satisfies the concentration inequality [109, Thm. 6.3] established by Rhee [92]:

$$P\left(|L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n) - E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]| > t\right) \le C\exp\left(\frac{-(t/C_2)^{2d/(d-\gamma)}}{Cn}\right), \tag{2.25}$$

where C is a constant depending only on the functional L_γ and d. The concentration inequality can also be used to bound the $\mathcal{L}_p$ moments

$$E[|L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n) - E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]|^p]^{1/p}\ ,\ p = 1,2,\ldots$$

In particular, as for any r.v. Z: $E[|Z|] = \int_0^\infty P(|Z| > t)\,\mathrm{d}t$, we have by (2.25)

$$\begin{aligned}
&E\left[|L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n) - E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]|^p\right] \\
&\quad= \int_0^\infty P\left(|L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n) - E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]| > t^{1/p}\right)\mathrm{d}t \\
&\quad\le C_2\,C\int_0^\infty \exp\left(\frac{-t^{2d/[p\,(d-\gamma)]}}{Cn}\right)\,\mathrm{d}t \\
&\quad= A_p\,n^{p\,(d-\gamma)/(2d)}\,,
\end{aligned} \tag{2.26}$$

where $A_p = C_2\, C^{p\,(d-\gamma)/(2d)+1} \int_0^\infty \mathrm{e}^{-u^{2d/[p(d-\gamma)]}}\, \mathrm{d}u$.

Combining the above with (2.24), we obtain

Proposition 6. *Let $d \geq 2$ and $1 \leq \gamma \leq d-1$. Assume $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. random vectors over $[0,1]^d$ with density $f \in \Sigma_d(\beta, L)$, $0 < \beta \leq 1$, having support $\mathcal{S} \subset [0,1]^d$. Assume also that $f^{\frac{1}{2}-\frac{\gamma}{d}}$ is integrable over $\mathcal{S}$. Then, for any continuous quasi-additive Euclidean functional L_γ of order γ that satisfies the add-one bound (2.8)*

$$\begin{aligned}
&\left[E\left|L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d} \int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}\right|^p\right]^{1/p} \qquad (2.27)\\
&\quad \leq \frac{K_1 + C_2}{(nm^{-d})^{1/d}} \left(\int_{\mathcal{S}} f^{\frac{d-1-\gamma}{d}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} + o(1)\right) + \frac{\beta_{L_\gamma,d}}{(nm^{-d})^{1/2}} \left(\int_{\mathcal{S}} f^{\frac{1}{2}-\frac{\gamma}{d}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} + o(1)\right)\\
&\quad + \frac{C_1}{(nm^{-d})^{(d-\gamma)/d}} + \frac{1}{n^{(d-\gamma)/d}} + (\beta_{L_\gamma,d} + C_2')\, C_4'\, L^{(d-\gamma)/d}\, m^{-\beta(d-\gamma)/d}\\
&\quad + A_p^{1/p} n^{-(d-\gamma)/(2d)}
\end{aligned}$$

Proof. For any non-random constant μ, using Minkowski inequality, $[E|W+\mu|^p]^{1/p} \leq [E|W|^p]^{1/p} + |\mu|$. Identify

$$\begin{aligned}
\mu &= E[L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d} \int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \qquad (2.28)\\
W &= (L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) - E[L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)])/n^{(d-\gamma)/d} \qquad (2.29)
\end{aligned}$$

and use (2.26) and (2.24) to establish Proposition 6. □

As the m-dependence of the bound of Proposition 6 is identical to that of the bias bound (2.24), minimization of the bound over $m = m(n)$ proceeds analogously to the proof of Proposition 5 and we obtain the following.

Corollary 7. *Let $d \geq 2$ and $1 \leq \gamma \leq d-1$. Assume $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. random vectors over $[0,1]^d$ with density $f \in \Sigma_d(\beta, L)$, $0 < \beta \leq 1$, having support $\mathcal{S} \subset [0,1]^d$. Assume also that $f^{\frac{1}{2}-\frac{\gamma}{d}}$ is integrable over $\mathcal{S}$. Then, for any continuous quasi-additive*

Euclidean functional L_γ of order γ that satisfies the add-one bound (2.8)

$$\left[E\left|L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)/n^{(d-\gamma)/d}-\beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right|^p\right]^{1/p} \leq O\left(n^{-r_1(d,\gamma,\beta)}\right), \tag{2.30}$$

where $r_1(d,\gamma,\beta)$ is defined in Proposition 5.

2.3.3 Discussion

It will be convenient to separate the discussion into the following points.

1. The bounds of Corollary 7 hold uniformly over the class of Lebesgue densities $f \in \Sigma_d(\beta, L)$ and integrable $f^{(d-\gamma)/d-1/2}$. If $\alpha = (d-\gamma)/d \in [1/2, (d-1)/d]$ then, as the support $\mathcal{S} \subset [0,1]^d$ is bounded, this integrability condition is automatically satisfied. To extend Corollary 7 to the range $\alpha \in ((d-1)/d, 1)$ would require extension of the fundamental convergence rate bound of $O\left(n^{-1/d}\right)$ used in (2.10), established by Redmond and Yukich [90], to the case $0 < \gamma < 1$.

2. It can be shown in analogous manner to the proof of the umbrella theorems of [109, Ch. 7] that if f is not a Lebesgue density then the convergence rates in Proposition 6 hold when the region of integration $\mathcal{S}$ is replaced by the support of the Lebesgue continuous component of f.

3. The convergence rate bound satisfies $r_1(d,\gamma,\beta) < 1/d$, which corresponds to Redmond and Yukich's rate bound for the uniform density over $[0,1]^d$ [109, Thm. 5.2]. Thus, the bound predicts slower worst case convergence rates for non-uniform densities.

4. When f is piecewise constant over a known partition of resolution $m = m_o$ faster rate of convergence bounds are available. For example, in Proposition 2 the bound in (2.19) is monotone increasing in m. Therefore the se-

quence $m(n) = m_o$ minimizes the bound as $n \to \infty$ and, proceeding in the same way as in the proof of Proposition 6, the best rate bound is of order $\max\left\{n^{-(d-\gamma)/(2d)},\, n^{-1/d}\right\}$. As the $O(n^{-1/d})$ bound on mean rate of convergence is tight [109, Sec. 5.3] for $d = 2$ and uniform density f, it is concluded that for $\alpha = (d-\gamma)/d \geq 2/d$ the asymptotic rate of convergence of the left hand side of (2.49) is exactly $O(n^{-1/d})$ for piecewise constant f and $d = 2$.

5. For $\alpha = (d-\gamma)) \geq 2/d$, it can be shown that the rate bound of Proposition 2 remains valid even if L_γ does not satisfy the "add-one bound." Thus, with $\alpha \geq 2/d$, Corollary 7 extends to any continuous quasi-additive functional L_γ including, in addition to the MST, the TSP, the minimal matching graph and the k-nearest neighbor graph functionals. As for the case $\alpha < 2/d$, we can use a weaker rate of mean convergence bound [109, Thm. 5.1], which applies to all continuous quasi-additive functionals and uniform f, in place of (2.10) in the proof of Proposition 2 to obtain

$$\left| E[L_\gamma(\boldsymbol{X_1}, \ldots, \boldsymbol{X}_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d} \int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \;\leq\; O\left(n^{-\frac{\alpha}{d/\beta+2}}\right). \tag{2.31}$$

6. A tighter upper bound than Corollary 6 on the $\mathcal{L}_p$-norm convergence rate may be derived if a better m-dependent analog to the concentration inequality (2.25) can be found.

2.4 Convergence Rates for Fixed Partition Approximations

Partitioning approximations to minimal graphs have been proposed by many authors, including Karp [46], Ravi *etal* [89], Mitchell [66], and Arora [2], as ways to reduce computational complexity. These algorithms use a "divide and conquer" strategy: first, partition the data into a collection of disjoint sets; second, compute the minimal graphs on each resulting set; and, finally, use the the total edge weight functionals of the resulting graphs to approximate the desired global total edge weight functional. The fixed partition approximation (i.e., non data-dependent) is a simple example whose convergence rate has been studied by Karp [46, 47], Karp and Steele [48] and Yukich [109] in the context of a uniform density f.

Fixed partition approximations to a minimal graph weight function require specification of an integer resolution parameter m controlling the number of cells in the uniform partition $\mathcal{Q}^m = \{Q_i\}_{i=1}^m$ of $[0,1]^d$ discussed in Section 2.2. When m is defined as an increasing function of n we obtain a progressive-resolution approximation to $L_\gamma(\mathcal{X}_n)$. This approximation involves constructing minimal graphs of order γ on each of the cells Q_i, $i = 1, \ldots, m^d$, and the approximation $L_\gamma^m(\mathcal{X}_n)$ is defined as the sum of their weights plus a constant bias correction $b(m)$

$$L_\gamma^m(\mathcal{X}_n) = \sum_{i=1}^{m^d} L_\gamma(\mathcal{X}_n \cap Q_i) + b(m), \tag{2.32}$$

where $b(m)$ is $O\left(m^{d-\gamma}\right)$. In this section we specify a bound on the $\mathcal{L}_p$-norm convergence rate of the progressive-resolution approximation (2.32) and specify the optimal resolution sequence $\{m(n)\}_{n>0}$ which minimizes this bound. Our derivations are based on the approach of Yukich [109, Sec. 5.4] and rely on the concrete version

of the pointwise closeness bound (2.7)

$$\left|L_\gamma(F) - L_\gamma^*(F)\right| \leq \begin{cases} C[\mathrm{card}(F)]^{(d-\gamma-1)/(d-1)}, & 1 \leq \gamma < d-1 \\ C \log \mathrm{card}(F), & \gamma = d-1 \neq 1 \\ C, & d-1 < \gamma < d \end{cases}, \tag{2.33}$$

for any finite $F \subset [0,1]^d$. This condition is satisfied by the MST, TSP and minimal matching function [109, Lemma 3.7].

We first obtain a fixed-m bound on $\mathcal{L}_1$ deviation of $L_\gamma^m(\mathcal{X}_n)/n^{(d-\gamma)/d}$ from its a.s. limit.

Proposition 8. *Let $d \geq 2$ and $1 \leq \gamma < d-1$. Assume that the Lebesgue density $f \in \Sigma_d(\beta, L)$, $0 < \beta \leq 1$, has support $\mathcal{S} \subset [0,1]^d$. Assume also that $f^{1/2-\gamma/d}$ are integrable over $\mathcal{S}$. Let $L_\gamma^m(\mathcal{X}_n)$ be defined as in (2.32) where L_γ is a continuous quasi-additive functional of order γ which satisfies the pointwise closeness bound (2.33) and the add-one bound (2.8). Then if $b(m) = O(m^{d-\gamma})$*

$$\begin{aligned} E\left[\left|L_\gamma^m(\mathcal{X}_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right|\right] \\ \leq O\left(\max\left\{(nm^{-d})^{-\gamma/[d(d-1)]},\ m^{-\beta(d-\gamma)/d},\ n^{-(d-\gamma)/(2d)}\right\}\right) \end{aligned} \tag{2.34}$$

Proof. Start with

$$\begin{aligned} &E\left[\left|L_\gamma^m(\mathcal{X}_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|\right] \\ &\quad \leq E\left[\left|L_\gamma(\mathcal{X}_n)/n^{\frac{d-\gamma}{d}} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{\frac{d-\gamma}{d}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|\right] + E\left[\left|L_\gamma^m(\mathcal{X}_n) - L_\gamma(\mathcal{X}_n)\right|\right]/n^{\frac{d-\gamma}{d}}. \end{aligned} \tag{2.35}$$

Analogously to the proof of [109, Thm. 5.7], using the pointwise closeness bound (2.33) one obtains a bound on the difference between the partitioned weight function

$L_\gamma^m(F)$ and the minimal weight function $L_\gamma(F)$ for any finite $F \subset [0,1]^d$

$$b(m) - C_1 m^{d-\gamma} \leq L_\gamma^m(F) - L_\gamma(F) \;\; \leq \;\; m^{-\gamma} C \sum_{i=1}^{m^d} (\mathrm{card}(F \cap Q_i))^{(d-\gamma-1)/(d-1)} + 1 + C_2 m^{d-\gamma} + b(m)\,. \tag{2.36}$$

As usual let $\phi(\boldsymbol{x}) = \sum_{i=1}^{m^d} \phi_i m^{-d}$ be a block density approximation to $f(\boldsymbol{x})$. As $\{\mathcal{X}_n \cap Q_i\}_{i=1}^{m^d}$ are independent and $E[|Z|^u] \leq (E[|Z|])^u$ for $0 \leq u \leq 1$

$$\begin{aligned}
&E[\big|L_\gamma^m(\mathcal{X}_n) - L_\gamma(\mathcal{X}_n)\big|] - |b(m) - C_1 m^{d-\gamma}| - 1 - C_2 m^{d-\gamma} - b(m) \\
&\quad \leq m^{-\gamma} C \sum_{i=1}^{m^d} E\left[(\mathrm{card}(\mathcal{X}_n \cap Q_i))^{(d-\gamma-1)/(d-1)}\right] \\
&\quad \leq m^{-\gamma} n^{(d-\gamma-1)/(d-1)} C \sum_{i=1}^{m^d} (\phi_i m^{-d})^{(d-\gamma-1)/(d-1)} \\
&\quad = m^{\gamma/(d-1)} n^{(d-\gamma-1)/(d-1)} C \sum_{i=1}^{m^d} \phi_i^{(d-\gamma-1)/(d-1)} m^{-d} \\
&\quad = m^{\gamma/(d-1)} n^{(d-\gamma-1)/(d-1)} C \int_{\mathcal{S}} \phi^{(d-\gamma-1)/(d-1)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}
\end{aligned}$$

Note that the bias term $|b(m) - C_1 m^{d-\gamma}|$ can be eliminated by selecting $b(m) = C_1 m^{d-\gamma}$. Dividing through by $n^{(d-\gamma)/d}$, noting that

$$\left(|b(m) - C_1 m^{d-\gamma}| + C_2 m^{d-\gamma} + b(m)\right) / n^{(d-\gamma)/d} \leq B(nm^{-d})^{-(d-\gamma)/d}$$

for some constant B,

$$\begin{aligned}
&E\left[\left|\frac{L_\gamma^m(\mathcal{X}_n) - L_\gamma(\mathcal{X}_n)}{n^{(d-\gamma)/d}}\right|\right] \\
&\quad \leq (nm^{-d})^{-\gamma/[d(d-1)]}\, C \int_{\mathcal{S}} \phi^{(d-\gamma-1)/(d-1)}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} + (nm^{-d})^{-(d-\gamma)/d} B + n^{-(d-\gamma)/d}\,.
\end{aligned} \tag{2.37}$$

Combining this with Proposition 6 we can bound the right hand side of (2.36) to obtain

$$
\begin{aligned}
&E\left[\left|L_\gamma^m(\mathcal{X}_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right|\right] \\
&\quad \le \frac{K_1+C_3}{(nm^{-d})^{1/d}}\left(\int_{\mathcal{S}} f^{\frac{d-1-\gamma}{d}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + o(1)\right) + \frac{\beta_{L_\gamma,d}}{(nm^{-d})^{1/2}}\left(\int_{\mathcal{S}} f^{\frac{1}{2}-\frac{\gamma}{d}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + o(1)\right) \\
&\quad + \frac{C_1}{(nm^{-d})^{(d-\gamma)/d}} + \frac{2}{n^{(d-\gamma)/d}} + (\beta_{L_\gamma,d}+C_2')\,C_4'\,L^{(d-\gamma)/d}\,m^{-\beta(d-\gamma)/d} + \ A_1 n^{-(d-\gamma)/(2d)} \\
&\quad + \frac{C}{(nm^{-d})^{\gamma/[d(d-1)]}}\left(\int_{\mathcal{S}} f^{(d-\gamma-1)/(d-1)}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + o(1)\right) + (nm^{-d})^{-(d-\gamma)/d}B. \qquad (2.38)
\end{aligned}
$$

Over the range $1 \le \gamma < d-1$ the dominant terms are as given in the statement of Proposition 8. □

Finally, by choosing $m = m(n)$ to minimize the maximum on the right hand side of the bound of Proposition 8 we have an analog to Corollary 7 for fixed partition approximations:

Corollary 9. *Let $d \ge 2$ and $1 \le \gamma < d-1$. Assume that the Lebesgue density $f \in \Sigma_d(\beta, L)$, $0 < \beta \le 1$, has support $\mathcal{S} \subset [0,1]^d$. Assume also that $f^{1/2-\gamma/d}$ is integrable over $\mathcal{S}$. Let $L_\gamma^m(\mathcal{X}_n)$ be defined as in (2.32) where L_γ is a continuous quasi-additive functional of order γ which satisfies the pointwise closeness bound (2.33) and the add-one bound (2.8). Then if $b(m) = O(m^{d-\gamma})$*

$$
E\left[\left|L_\gamma^{m(n)}(\boldsymbol{X_1},\ldots,\boldsymbol{X}_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right|\right] \le O\left(n^{-r_2(d,\gamma,\beta)}\right), \qquad (2.39)
$$

where

$$
r_2(d,\gamma,\beta) = \frac{\alpha\,\beta}{\frac{d-1}{\gamma}\,\alpha\,\beta + 1}\ \frac{1}{d}\ ,
$$

where $\alpha = \frac{d-\gamma}{d}$. This rate is attained by choosing the progressive-resolution sequence

$m = m(n) = n^{1/[d(\frac{d-1}{\gamma}\alpha\beta+1)]}$.

2.4.1 Discussion

We make the following remarks.

1. Under the assumed condition $\gamma < d-1$ in Corollary 9, $r_2(d,\gamma,\beta) \leq r_1(d,\gamma,\beta)$, where $r_1(d,\gamma,p)$ is defined in Corollary 7. Thus, as might be expected, the partitioned approximation has a $\mathcal{L}_p$-norm convergence rate (2.39) that is always slower than the rate bound (2.49), and the slowdown increases as $(d-1)/\gamma$ increases.

2. In view of (2.38), up to a monotonic transformation, the rate constant multiplying the asymptotic rate $n^{-r_2(d,\gamma,\beta)}$ is an increasing function of $\int_{\mathcal{S}} f^{(d-\gamma-1)/(d-1)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, which is the Rényi entropy of f of order $(d-\gamma-1)/(d-1)$. Thus fastest convergence can be expected for densities with small Rényi entropy.

3. It is more tedious but straightforward to show that the $\mathcal{L}_2$ deviation

$$E\left[\left|L_\gamma^m(\mathcal{X}_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right|^2\right]^{1/2}$$

   obeys the identical asymptotic rate bounds as in Proposition 8 and Corollary 9 with identical bound minimizing progressive-resolution sequence $m = m(n)$.

4. As pointed out in the proof of Proposition 8 the bound minimizing choice of the bias correction $b(m)$ of the progressive-resolution approximation (2.32) is $b(m) = C_1 m^{d-\gamma}$, where C_1 is the constant in the subaddivity condition (2.2). However, Proposition 8 asserts that, for example, using $b(m) = Cm^{d-\gamma}$ with arbitrary scale constant C, or even using $b(m) = 0$, are asymptotically

equivalent to the bound minimizing $b(m)$. This is important since the constant C_1 is frequently difficult to determine and depends on the specific properties of the minimal graph, which are different for the TSP, MST, etc.

5. The partitioned approximation (2.32) is a special case $k = n$ of the greedy approximation to the k-point minimal graph approximation introduced by Ravi *etal* [88] whose a.s. convergence was established by Hero and Michel [39] (Note that the overly strong BV condition assumed in [39] can be considerably weakened by replacing BV space with Hölder space and applying Lemma 3 of this paper). Extension of Proposition 8 to greedy approximations to k-point graphs is an open problem.

2.5 Convergence Rate Lower Bounds

In this section we derive lower bounds for the convergence rates of minimal graphs based on minimax estimation theory. While these bounds are not generally tight lower bounds, they indicate a performance margin between graph estimators and minimax estimators of entropy. Our results can be obtained as an application of the general theory developed by Birgé and Massart in [10] for obtaining lower bounds on the minimax risk of nonparametric estimation of a functional $T(f) = \int \varphi(f(x), f'(x), \ldots, f^{(k)}(x), x)\, \mathrm{d}x$. In fact, Proposition 12, in this section, can be derived as a corollary to Theorem 3 in [10], after some suitable modifications as suggested in Remark 3 of that paper. However, for the benefit of the reader, we provide a more elementary and self contained proof of the lower bound in the sequel, which applies to the specific functional of form (2.40).

Define

$$I_\alpha(f) = \int f^\alpha(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \ . \tag{2.40}$$

From Sections 2.2 and 2.3, $L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)/n^{(d-\gamma)/d}$ is a (strongly) consistent estimator of $I_\alpha(f)$ for $\alpha = \frac{d-\gamma}{d}$. Thus, it is natural to recast our problem as that of estimating $I_\alpha(f)$ over the nonparametric class of densities $f \in \Sigma_d(\beta, L)$.

Let $\hat{I}_\alpha$ be an estimator of $I_\alpha(f)$ $(0 < \alpha < 1)$ based on a sample of n i.i.d. observations from a density f. To access the "quality" of $\hat{I}_\alpha$ we adopt the usual (nonparametric) minimax risk criterion, i.e., we look at $\sup_{f \in \mathcal{F}} E|\hat{I}_\alpha - I_\alpha(f)|^p$, the worst case performance of $\hat{I}_\alpha$ over a known class of densities $\mathcal{F}$, for a choice of $p \geq 1$. Under this criterion it is natural to ask what is the minimum achievable risk for any estimator, i.e., what is

$$\inf_{\hat{I}_\alpha} \sup_{f \in \mathcal{F}} E|\hat{I}_\alpha - I_\alpha(f)|^p \ ,$$

where the infimum is taken over all estimators of $I_\alpha(f)$, as this quantifies the best performance possible for any estimator. Of course, as $L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)/n^\alpha$ is valid estimator of $I_\alpha(f)$, this will also yield a lower bound to the convergence rates of interest. The rest of this section is devoted to deriving these (asymptotic) bounds using standard minimax techniques.

2.5.1 Notation

In the following, we will take the class $\mathcal{F}$ as the set of multivariate Lebesgue densities defined on the unit cube $[0,1]^d$ $(d \geq 1)$, belonging to the Hölder class of functions $\Sigma_d(\beta, L)$.

We will also use the affinity $\|P \wedge Q\|$ between measures P and Q defined by:

$$\|P \wedge Q\| = 1 - \frac{1}{2}\|P - Q\|_1 \tag{2.41}$$

where $\|P\|_1$ is the total variation norm of P defined as

$$\|P\|_1 = \sup_{|f|\leq 1} \left| \int f \, \mathrm{d}P \right|$$

and the supremum is taken over all measurable functions f bounded by 1. If P and Q are absolutely continuous w.r.t. a measure μ, with densities p and q, respectively, then $\|P - Q\|_1 = \int |p - q| \, \mathrm{d}\mu$. In this case, we will write $\|p - q\|_1$ for $\|P - Q\|_1$ and $\|p \wedge q\|$ for $\|P \wedge Q\|$. Also, write p^n as shorthand notation for $\prod_{i=1}^n p(\boldsymbol{x}_i)$, the density of the product measure $\otimes_n P$.

Finally, write $co(\mathcal{F})$ to denote the convex hull of $\mathcal{F}$.

2.5.2 Lower Bounds

In order to get lower bounds for the minimax risk, the usual technique is to build, for every n, a subset $\mathcal{F}_{0,n} \subset \mathcal{F}$ of finite cardinality, such that the problem of estimating $I_\alpha(f)$ over $\mathcal{F}_{0,n}$ is essentially as difficult as the full problem. Assouad's lemma or Fano's lemma are the commonly used tools to address such constructions [43]. However, in the case of entropy estimation (as well as many other functional estimation problems, [108], [61]), these methods only give the trivial lower bound zero. We will thus rely on a result by Le Cam (see for example [108]) that relates the minimax risk to a testing problem between two sets of hypothesis, whose convex hulls are "well" separated in a total variation distance sense. Below is a simplified version of this result suitable for our needs (for a simple proof see [108]):

Lemma 10. *Let $\hat{I}$ be an estimator of $I(f)$[1] based on n i.i.d. observations from a density $f \in \mathcal{F}$. Suppose that there are subsets $\mathcal{G}_1$ and $\mathcal{G}_2$ of $\mathcal{G} = \{f^n : f \in \mathcal{F}\}$ that are 2δ-separated, in the sense that, $|I(f_1) - I(f_2)| \geq 2\delta$ for all $f_1^n \in \mathcal{G}_1$ and $f_2^n \in \mathcal{G}_2$.*

[1]From now on, we will omit the subscript α from $\hat{I}_\alpha$ and $I_\alpha(f)$, unless necessary.

Then

$$\sup_{f\in\mathcal{F}} E|\hat{I} - I(f)| \geq \delta \cdot \sup_{p_i \in co(\mathcal{G}_i)} \|p_1 \wedge p_2\| \ .$$

We will apply Lemma 10 to the usual small perturbations of the uniform density, u, on $[0,1]^d$. Towards this goal, fix $g \in \Sigma_d(\beta, 1)$ with support in $[0,1]^d$ such that $\int g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0$, $\|g\|_2^2 = \int g^2(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} > 0$ and $|g(\boldsymbol{x})| \leq M$. Let $\{Q_j\}_{j=1}^{m^d}$ be the uniform resolution-m partition and $\{\boldsymbol{x}_j\}_{j=1}^{m^d}$ be the set of points in $[0,1]^d$ that translate each Q_j back to the origin, as defined in Sub-section 2.2.1. Let $g_j(\boldsymbol{x}) = g\left(m(\boldsymbol{x} - \boldsymbol{x}_j)\right)$. For $\lambda \in \Lambda = \{-1,1\}^{m^d}$, define the perturbation of u as

$$f_\lambda(\boldsymbol{x}) = 1 + \sum_{j=1}^{m^d} \frac{L}{2}\, m^{-\beta}\, \lambda_j\, g_j(\boldsymbol{x}) \tag{2.42}$$

It is easy to see that $\int f_\lambda(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1$, $f_\lambda \in \Sigma_d(\beta, L)$ and, for m large enough, $f \geq 0$. Hence (for m sufficiently large) $f \in \mathcal{F}$.

We can now apply Lemma 10 to the sets $\mathcal{G}_1 = \{u^n\}$ and $\mathcal{G}_2 = \{f_\lambda^n : \lambda \in \Lambda\}$. We will start by determining the 2δ-separation between $\mathcal{G}_1$ and $\mathcal{G}_2$. Consider the second order Taylor expansion

$$(1+y)^\alpha = 1 + \alpha y + \frac{1}{2}\,\alpha(\alpha-1)\,\xi^{\alpha-2} y^2 \ ,$$

where ξ lies between 1 and $1+y$. This implies that

$$\begin{aligned}
\int f_\lambda^\alpha(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - 1 &= \sum_{j=1}^{m^d} \int_{Q_j} \left(1 + \frac{L}{2}\, m^{-\beta}\, \lambda_j\, g_j(\boldsymbol{x})\kappa\right)^\alpha \mathrm{d}\boldsymbol{x} - 1 \\
&= \frac{1}{2}\left(\frac{L}{2}\right)^2 \alpha(\alpha-1)\, m^{-2\beta} \sum_{j=1}^{m^d} \int_{Q_j} \xi^{\alpha-2}(\boldsymbol{x}) g_j^2(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \ ,
\end{aligned} \tag{2.43}$$

where $1 - M\frac{L}{2}\, m^{-\beta} \leq \xi(\boldsymbol{x}) \leq 1 + M\frac{L}{2}\, m^{-\beta}$. Inserting these bounds in equation

(2.43), we have

$$\frac{1}{2}\left(\frac{L}{2}\right)^2 \alpha(\alpha-1)\,\|g\|_2\, m^{-2\beta}\left(1-M\frac{L}{2}\,m^{-\beta}\right)^{\alpha-2} \le \int f_\lambda^\alpha(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - 1$$
$$\le \frac{1}{2}\left(\frac{L}{2}\right)^2 \alpha(\alpha-1)\,\|g\|_2\, m^{-2\beta}\left(1+M\frac{L}{2}\,m^{-\beta}\right)^{\alpha-2} ,$$

which essentially means that $\int f_\lambda^\alpha(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - 1 \doteq m^{-2\beta}$. We can now use this result to conclude that, for any $\lambda \in \Lambda$ and m sufficiently large,

$$|I(f_\lambda^n) - I(u^n)| = \left|\int f_\lambda^\alpha(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - 1\right| \ge 2\,C\,m^{-2\beta} = 2\delta\ , \tag{2.44}$$

for some constant $C > 0$.

We next derive a lower bound for $\sup_{p_i \in co(\mathcal{G}_i)} \|p_1 \wedge p_2\|$, or equivalently, by (2.41), an upper bound on $\|p_1 - p_2\|_1$. To this end, let $h_n = 2^{-m^d}\sum_{\lambda\in\Lambda} f_\lambda^n \ \in co(\mathcal{G}_2)$. The following Lemma provides the required result:

Lemma 11.

$$\|u^n - h_n\|_1^2 \le \exp\left\{\frac{1}{2}\left(\frac{L}{2}\,\|g\|_2\right)^4 n^2\, m^{-(4\beta+d)}\right\} - 1\ . \tag{2.45}$$

A proof of this Lemma is given in Section 2.8.

Plugging the bounds from equations (2.44) and (2.45), together with (2.41), into Lemma 10 gives us a family of lower bounds, for different values of m:

$$\sup_{f\in\mathcal{F}} E|\hat{I} - I(f)| \ge \frac{1}{2}\,C\,m^{-2\beta}\cdot\left(3 - \exp\left\{\frac{1}{2}\left(\frac{L}{2}\,\|g\|_2\right)^4 n^2\, m^{-(4\beta+d)}\right\}\right)\ . \tag{2.46}$$

We can now choose $m = m(n)$ in order to maximize this bound. This can easily be done by inspection: the first term on the RHS of (2.46) should be as large as possible, i.e., m should be as small as possible; however, such a choice will make

the second term on the RHS of (2.46) negative, rending this bound useless. Hence, under this constraint, a choice for m that maximizes the bound is:

$$m = \left\lceil \left\{ \sqrt{\frac{1}{2}} \left(\frac{L}{2} \|g\|_2 \right)^2 n \right\}^{\frac{2}{4\beta+d}} \right\rceil \ , \tag{2.47}$$

where the constants multiplying n in the previous expression guarantee the positivity of the second term on the RHS of (2.46). Finally, inserting this optimum choice for m into (2.46) and using Jensen's inequality, gives us the desired lower bound:

Proposition 12. *For $\mathcal{F}_{\beta,L} = \{f : f$ is a Lebesgue density on $[0,1]^d$ and $f \in \Sigma_d(\beta, L)\}$, $p \geq 1$ and n sufficiently large, there exists a constant $c = c(\beta, L, d, \alpha) > 0$ such that*

$$\inf_{\hat{I}_\alpha} \sup_{f \in \mathcal{F}_{\beta,L}} \left[E|\hat{I}_\alpha - I_\alpha(f)|^p \right]^{1/p} \geq c \ n^{-\frac{4\beta}{4\beta+d}} \ , \tag{2.48}$$

where the infimum is taken over all estimators $\hat{I}_\alpha$ of $I_\alpha(f)$ based on n i.i.d. observations from density f.

We make the following comments about this proposition.

1. For sufficiently smooth densities, i.e., for $\beta \geq d/4$, we have $4\beta/(4\beta + d) \geq 1/2$. This is the usual $\sqrt{n}$-rate of convergence for regular parametric problems. This suggests that the lower bound in Proposition 12 can be replaced by

$$\inf_{\hat{I}_\alpha} \sup_{f \in \mathcal{F}_{\beta,L}} \left[E|\hat{I}_\alpha - I_\alpha(f)|^p \right]^{1/p} \geq c \ n^{-\left(\frac{4\beta}{4\beta+d} \wedge \frac{1}{2}\right)} \ .$$

2. It was shown in [10], for $\beta \geq d/4$, that there exists an estimator that achieves the $\sqrt{n}$-rate, for densities bounded from above and bounded from below by some positive constant. In [50], Kerkyacharian and Picard closed the problem by showing that the corresponding rates for $\beta < d/4$ are also achievable. Such

estimators are based on corrections, up to second or third order, of a preliminary plug-in estimator $T(\hat{f})$, where $\hat{f}$ is a nonparametric density estimate of f, based on a small part of the sample. However, these type of estimators are of little use in a practical high-dimensional setting, as multivariate integration and density estimation became unmanageable in a high dimensional space.

3. If, instead of the Rényi entropy, we were interested in the Shannon entropy $H_1(f) = -\int f(\boldsymbol{x}) \log f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$, the same rates would be obtained. This can be seen by considering the second order Taylor expansion,

$$(1+y)\log(1+y) = y + \frac{1}{2}\,\xi^{-1}\,y^2$$

and following the same steps as for $I_\alpha(f)$. In [58], Laurent exhibits an efficient estimator of this entropy, for densities defined on a compact set of the real line with smoothness parameter $\beta \geq 1/4$, that achieves the $\sqrt{n}$-rate on densities bounded away from zero on their domain.

Now, combining Proposition 12 with Corollary 7, we obtain upper and lower bounds for the convergence rates of minimal Euclidean graphs:

Corollary 13. *Let $d \geq 2$ and $1 \leq \gamma \leq d-1$. Assume $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. random vectors with density $f \in \mathcal{F}_{\beta,L}$, $\beta \in (0,1]$. Assume also that $f^{\frac{1}{2}-\frac{\gamma}{d}}$ is integrable. Then, for any continuous quasi-additive Euclidean functional L_γ of order γ that satisfies the add-one bound (2.8), there exist positive constants c, C, depending on β, L, d and γ such that for n sufficiently large*

$$c\, n^{-\left(\frac{4\beta}{4\beta+d}\right)} \leq \sup_{f\in\mathcal{F}_{\beta,L}} \left[E \left| L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d} \int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \right|^p \right]^{1/p} \leq C n^{-r_1(d,\gamma,\beta)} \,, \tag{2.49}$$

where $r_1(d, \gamma, \beta)$ is defined in Proposition 5.

There is a big gap between the lower and upper bound in (2.49). For example, for small β or large d, the lower bound rate exponent is at least 4 times faster than the corresponding upper bound rate.

We do not believe that the bounds of Corollary 13 are the tightest possible bounds. On the one hand, the derivation of the upper bound is based on a coupling argument (see proof of Proposition 4) which may over estimate the error. On the other hand, the derivation of the lower bounds was solely based on minimax arguments that do not account for the intrinsic geometric structure of minimal Euclidean graphs.

2.6 Performance of Minimal Graph and Plug-in Entropy Estimators

In this section we derive upper bounds for the maximum risk of plug-in estimators, based on first estimating the density and then plug it in the entropy functional, and compare to minimal-graph based estimators of entropy.

We consider entropy estimates of the form $\hat{H}_\alpha = (1-\alpha)^{-1} \log \hat{I}_\alpha$, where $\hat{I}_\alpha$ is a consistent estimator of $I_\alpha(f) = \int f^\alpha(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$. By a standard perturbation analysis of $\ln x$,

$$|\hat{H}_\alpha - H_\alpha(f)| = \frac{1}{1-\alpha} \frac{|\hat{I}_\alpha - I_\alpha(f)|}{I_\alpha(f)} + o(|\hat{I}_\alpha - I_\alpha(f)|) \ .$$

Thus, as $I_\alpha(f)$ is bounded away from zero uniformly over the class $\mathcal{F}$ (i.e., $\inf_{f \in \mathcal{F}} I_\alpha(f) > 0$), the asymptotic rate of convergence of $\hat{H}_\alpha - H_\alpha(f)$, as a function of n, will be identical to that of $\hat{I}_\alpha - I_\alpha(f)$.

Let $\hat{f}$ be a density estimate of f based on n i.i.d. observations (from density f).

We have the following upper bound for plug-in estimators $I_\alpha(\hat{f})$:

Proposition 14. *For $\mathcal{F}$ as defined in Proposition 12,*

$$\sup_{f\in\mathcal{F}} E\left|I_\alpha(\hat{f}) - I_\alpha(f)\right| \leq C_1\, n^{-\frac{\alpha\beta}{2\beta+d}} \tag{2.50}$$

for $C_1 = C_1(\beta, L, d) > 0$.

Proof. The proof relies on the well known minimax rates for density estimation available in the literature (see, for example, [72]). Specifically, these rates are of order $O\left(n^{-\beta/(2\beta+d)}\right)$, i.e.,

$$\sup_{f\in\mathcal{F}} E\int |\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})|\,\mathrm{d}\boldsymbol{x} \leq C_1\, n^{-\frac{\beta}{2\beta+d}}$$

for the best estimators $\hat{f}$ (for example, wavelet thresholding based estimators).

Using the above result, the inequality $|a^\alpha - b^\alpha| \leq |a-b|^\alpha$ $(a, b \geq 0)$ and successive applications of Jensen's inequality yield the desired result,

$$\begin{aligned} E\left|I_\alpha(\hat{f}) - I_\alpha(f)\right| &\leq E\int \left|\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\right|^\alpha \mathrm{d}\boldsymbol{x} \leq E\left[\int \left|\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\right| \mathrm{d}\boldsymbol{x}\right]^\alpha \\ &\leq \left[E\int \left|\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\right| \mathrm{d}\boldsymbol{x}\right]^\alpha \leq C_1\, n^{-\frac{\alpha\beta}{2\beta+d}} \end{aligned}$$

□

For $\hat{I}_\alpha$ denoting the minimal graph estimator of $I_\alpha(f)$, we have from Proposition 6 the following result:

Proposition 15. *For $\mathcal{F}$ as defined in Proposition 12, with $0 < \beta \leq 1$, $1/2 \leq \alpha \leq (d-1)/d$,*

$$\sup_{f\in\mathcal{F}} E\left|\hat{I}_\alpha - I_\alpha(f)\right| \leq C_2\, n^{-\frac{\alpha\beta}{\alpha\beta+1}\frac{1}{d}} \tag{2.51}$$

for $C_2 = C_2(\beta, L, d, \alpha) > 0$.

Comparing Propositions 14 and 15, it can be seen that upper bound in (2.51) is smaller than (2.50) for $\alpha < 2/d$. This shows that for small values of α, graph based estimators of entropy will have a faster convergence rate. For other values of α, the looseness of the bound in (2.51) does not allow to reach any conclusion.

2.7 Conclusion

In this chapter we have given upper and lower bounds on the convergence rates for length functionals of minimal-graphs satisfying continuous quasi-additivity conditions, for general multivariate densities of the vertices. These bounds make explicit the dependency of the approximation error not only as a function of the number of samples, n, but also in terms of the dimension of the space, d, and the underlying class of densities. These results may be useful for exploring the asymptotic behavior of minimal graphs, e.g., for estimation of Rényi divergence, Rényi mutual information, and Rényi Jensen difference [40]. For example, by studying how the constants involved in the bounds depend on the graph constructions, one could compare the performance of different graphs, e.g., MST, k-NNG, etc, for entropy estimation in the finite sample size case.

There are still many open problems that remain to be studied. Of great interest is the extension of these results to k-point graphs (such as the k-MST), as, not only do they provide robustness against outliers, but they also have a natural application to unsupervised clustering. Also, to complete the results given in this paper, it would be interesting to extend the rate bounds to smoother Hölder continuous densities (i.e., $\beta > 1$). With regards to future applications, we feel that these methods can be applied in problems such as independent component analysis (ICA) or clustering

techniques. Finally, establishing general weak convergence results, e.g., a central limit theorem, for these types of minimal graphs could have a significant impact in applications such as hypothesis testing and goodness of fit tests.

2.8 Appendix: Proofs of Technical Lemmas

Proof of Lemma 1. Since $g(u)$ is concave the tangent line $y(u) \stackrel{\text{def}}{=} g(u_o) + g'(u_o)(u - u_o)$ upper bounds g. Hence

$$g(u) \leq g(u_o) + g'(u_o)|u - u_o| \,.$$

On the other hand, as g is monotone and concave, the function $z(u) \stackrel{\text{def}}{=} g(u_o) + \frac{g(u_o)}{u_o}(u - u_o)1_{\{u \leq u_o\}}$ is a lower bound on g, where $1_{\{u \leq u_o\}}$ is the indicator function of the set $\{u \leq u_o\}$. Hence,

$$g(u) \geq g(u_o) - \frac{g(u_o)}{u_o}|u - u_o| \,.$$

□

Proof of Lemma 3. By the mean value theorem, there exist points $\boldsymbol{\xi}_i \in Q_i$ such that

$$\phi_i = m^d \int_{Q_i} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = f(\boldsymbol{\xi}_i) \;.$$

Note that, in what follows, $|.|$ means both the absolute value in $\mathbb{R}$ and any norm in $\mathbb{R}^d$. Using now the fact that $f \in \Sigma_d(\beta, L)$,

$$\int_{\mathcal{S}} |\phi(\boldsymbol{x}) - f(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} = \sum_{i=1}^{m^d} \int_{Q_i} |f(\boldsymbol{\xi}_i) - f(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \leq \sum_{i=1}^{m^d} \int_{Q_i} L\,|\boldsymbol{x} - \boldsymbol{\xi}_i|^\beta \mathrm{d}\boldsymbol{x} \;.$$

As $\boldsymbol{x}, \boldsymbol{\xi}_i \in Q_i$, a sub-cube with edge length m^{-1}, $\int_{Q_i} |\boldsymbol{x} - \boldsymbol{\xi}_i|^\beta \mathrm{d}\boldsymbol{x} = O(m^{-\beta-d})$. Thus,

we have

$$\int_{\mathcal{S}} |\phi(\boldsymbol{x}) - f(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \leq C\, L\, m^{-\beta} \ .$$

□

Proof of Lemma 11. This proof is inspired by [83]. Define

$$G_i(\boldsymbol{\lambda}) = G(\boldsymbol{X}_i, \boldsymbol{\lambda}) = \sum_{j=1}^{m^d} \frac{L}{2} m^{-\beta} \lambda_j g_j(\boldsymbol{X}_i) = \frac{L}{2} m^{-\beta} \boldsymbol{\lambda}^t \boldsymbol{g}(\boldsymbol{X}_i)$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{m^d})^t \in \Lambda$ and $\boldsymbol{g} = (g_1, \ldots, g_{m^d})^t$. Define also

$$\tau_i(\boldsymbol{\lambda}, \boldsymbol{\mu}) = E_{u^n} G_i(\boldsymbol{\lambda}) G_i(\boldsymbol{\mu})$$

for $\boldsymbol{\lambda}, \boldsymbol{\mu} \in \Lambda$. Note that, as $\int g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0$,

$$E_{u^n} G_i(\boldsymbol{\lambda}) = 0 \ , \tag{2.52}$$

and due to identically distributed samples assumption, $\tau_i(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \tau_1(\boldsymbol{\lambda}, \boldsymbol{\mu})$.

Now, rewrite h_n as:

$$\begin{aligned} h_n &= \sum_{\lambda \in \Lambda} 2^{-m^d} f_\lambda^n = \sum_{\boldsymbol{\lambda} \in \Lambda} w_{\boldsymbol{\lambda}} \prod_{i=1}^{n} (1 + G_i(\boldsymbol{\lambda})) \\ &= \sum_{\boldsymbol{\lambda} \in \Lambda} w_{\boldsymbol{\lambda}} \left(1 + \sum_i G_i(\boldsymbol{\lambda}) + \sum_{i<j} G_i(\boldsymbol{\lambda}) G_j(\boldsymbol{\lambda}) + \sum_{i<j<k} G_i(\boldsymbol{\lambda}) G_j(\boldsymbol{\lambda}) G_k(\boldsymbol{\lambda}) + \ldots \right) \end{aligned}$$

where $w_{\boldsymbol{\lambda}} = 2^{-m^d}$.

Using Jensen's inequality,

$$\begin{aligned}\|h_n - u^n\|_1^2 &= \left(E_{u^n}|h_n - 1|\right)^2 \leq E_{u^n}|h_n - 1|^2 \\ &= E_{u^n}\left\{\sum_{\boldsymbol{\lambda},\boldsymbol{\mu}\in\Lambda} w_{\boldsymbol{\lambda}} w_{\boldsymbol{\mu}} \left(\sum_i G_i(\boldsymbol{\lambda}) + \sum_{i<j} G_i(\boldsymbol{\lambda})G_j(\boldsymbol{\lambda}) + \ldots\right)\left(\sum_i G_i(\boldsymbol{\mu}) + \right.\right. \\ &\qquad \left.\left. + \sum_{i<j} G_i(\boldsymbol{\mu})G_j(\boldsymbol{\mu}) + \ldots\right)\right\} \qquad (2.53)\end{aligned}$$

Expanding out the product in (2.53), due to independence and (2.52), only the terms where each factor $G_i(\boldsymbol{\lambda})$ is paired with a corresponding $G_i(\boldsymbol{\mu})$ will survive. All other terms with an isolated factor will be zero. The simplified result is

$$\begin{aligned}E_{u^n}|h_n - 1|^2 &= \sum_{\boldsymbol{\lambda},\boldsymbol{\mu}\in\Lambda} w_{\boldsymbol{\lambda}} w_{\boldsymbol{\mu}} \left(\sum_i \tau_i(\boldsymbol{\lambda},\boldsymbol{\mu}) + \sum_{i<j} \tau_i(\boldsymbol{\lambda},\boldsymbol{\mu})\tau_j(\boldsymbol{\lambda},\boldsymbol{\mu}) + \ldots\right) \\ &= \sum_{\boldsymbol{\lambda},\boldsymbol{\mu}\in\Lambda} w_{\boldsymbol{\lambda}} w_{\boldsymbol{\mu}} \left(1 + \tau_1(\boldsymbol{\lambda},\boldsymbol{\mu})\right)^n - 1 \qquad (2.54)\end{aligned}$$

Regarding the double sum in (2.54) as an expectation of a pair of independent random variables $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, each distributed according to a uniform prior in Λ, we get the following bound for the total variation norm:

$$\|h_n - u^n\|_1^2 \leq E\left(1 + \tau_1(\boldsymbol{\lambda},\boldsymbol{\mu})\right)^n - 1 \leq E\exp\{n\,\tau_1(\boldsymbol{\lambda},\boldsymbol{\mu})\} - 1\ , \qquad (2.55)$$

where the last inequality comes from $\mathrm{e}^x \geq 1 + x$.

Now, note that the functions g_i have disjoint supports and, so, are orthogonal in the sense that $E_u g_i(\boldsymbol{X}_1)g_j(\boldsymbol{X}_1) = 0$, for $i \neq j$. Thus, we have

$$\tau_1(\boldsymbol{\lambda},\boldsymbol{\mu}) = \left(\frac{L}{2}\,m^{-\beta}\right)^2\ \boldsymbol{\lambda}^t\, E_{u^n}\left\{\boldsymbol{g}(\boldsymbol{X}_1)\boldsymbol{g}^t(\boldsymbol{X}_1)\right\}\boldsymbol{\mu} = \sigma^2\ \boldsymbol{\lambda}^t\boldsymbol{\mu}\ ,$$

with $\sigma^2 = \int \left(\frac{L}{2}\, m^{-\beta} g_1(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x} = \int \left[\frac{L}{2}\, m^{-\beta} g(m(\boldsymbol{x}-\boldsymbol{x}_1))\right]^2 \mathrm{d}\boldsymbol{x} = \left(\frac{L}{2}\, \|g\|_2\right)^2 m^{-(2\beta+d)}$, where $\|g\|_2^2 = \int g^2(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$. Equation (2.55) simplifies to

$$\|h_n - u^n\|_1^2 \leq E \exp\{n\,\sigma^2\ \boldsymbol{\lambda}^t \boldsymbol{\mu}\} - 1\ .$$

The above expectation is easy to compute because the choice of a uniform prior on Λ makes the coordinates λ_i independent, taking values $+1$ and -1 with probability $1/2$:

$$E \exp\{n\,\sigma^2\ \boldsymbol{\lambda}^t \boldsymbol{\mu}\} = \left(\frac{1}{2}\,\mathrm{e}^{n\,\sigma^2} + \frac{1}{2}\,\mathrm{e}^{-n\,\sigma^2}\right)^{m^d} \leq \exp\left\{\frac{1}{2}\, m^d\, (n\,\sigma^2)^2\right\}\ .$$

Lemma 11 now follows. □

2.9 Appendix: Convergence Rates for Sobolev Densities

In this Appendix we will introduce some concepts from the theory of Sobolev spaces and then show how to extend the previous results on convergence rate bounds to densities in the Sobolev class.

Let $\mathcal{L}_p(\mathbb{R}^d)$ be the space of measurable functions over $\mathbb{R}^d$ such that $\|f\|_p = (\int |f(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x})^{1/p} < \infty$. For f a real valued differentiable function over $\mathbb{R}^d$, let $\mathrm{D}_{x_j} f = \partial f/\partial x_j$ be the x_j-th partial derivative of f, and $\mathrm{D}f = [\partial f/\partial x_1, \ldots, \partial f/\partial x_d]$ be the gradient of f. The concept of derivative can be extended to non-differentiable functions. For $f \in \mathcal{L}_1(\mathbb{R}^d)$, g is called the x_j-th *weak derivative* of f [114], written as $g \stackrel{\text{def}}{=} \mathrm{D}_{x_j} f$ if

$$\int_{\mathbb{R}^d} f(\boldsymbol{x}) \mathrm{D}_{x_j} \varphi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = -\int_{\mathbb{R}^d} g(\boldsymbol{x}) \varphi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

for all functions φ infinitely differentiable with compact support. The weak derivative g is sometimes called the *generalized derivative* of f or *distributional derivative* of f. If f is differentiable, then its weak derivative coincides with the (usual) derivative.

We now define a function space whose members have weak derivatives lying in the $\mathcal{L}_p(\mathbb{R}^d)$ spaces [114]. For $p \geq 1$, define the *Sobolev space*

$$W^{1,p}(\mathbb{R}^d) = \mathcal{L}_p(\mathbb{R}^d) \cap \{f : \mathrm{D}_{x_j} f \in \mathcal{L}_p(\mathbb{R}^d),\ 1 \leq j \leq d\} \ .$$

The space $W^{1,p}$ is equipped with a norm

$$\|f\|_{1,p} = \|f\|_p + \|\mathrm{D}f\|_p \ .$$

The Sobolev space $W^{1,p}(\mathbb{R}^d)$ is a generalization of the space of continuously differentiable functions, in the sense that $W^{1,p}(\mathbb{R}^d)$ contains functions that do not have to be differentiable (in the usual sense), but can be approximated arbitrarily close in the $\|.\|_{1,p}$ norm by infinitely differentiable functions with compact support ([114, Thm. 2.3.2]).

Let ϕ be the resolution-m block density approximation of f, as defined in section 2.3.2. The following lemma establishes how close (in $\mathcal{L}_1(\mathbb{R}^d)$ sense) these resolution-m block densities approximate functions in $W^{1,p}(\mathbb{R}^d)$.

Lemma 16. *For $1 \leq p < \infty$, let $f \in W^{1,p}(\mathbb{R}^d)$ have support $\mathcal{S} \subset [0,1]^d$. Then there exists a constant $C > 0$, independent of m, such that*

$$\int_{\mathcal{S}} |\phi(\boldsymbol{x}) - f(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \leq C m^{-1}(\|\mathrm{D}f\|_p + o(1)) \ . \tag{2.56}$$

Proof. First assume that f is a continuously differentiable function. By the mean

value theorem, there exist points $\boldsymbol{\xi_i} \in Q_i$ such that

$$\phi_i = m^d \int_{Q_i} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = f(\boldsymbol{\xi_i}) \ .$$

Also by the mean value theorem there exist points $\boldsymbol{\psi_i} \in Q_i$ such that

$$|f(\boldsymbol{x}) - f(\boldsymbol{\xi}_i)| = |\mathrm{D}f(\boldsymbol{\psi_i}) \cdot (\boldsymbol{x} - \boldsymbol{\xi}_i)|, \quad \boldsymbol{x} \in Q_i \ .$$

Using the above results, Jensen inequality and Cauchy-Schwarz inequality

$$\begin{aligned}
\left(\int_{\mathcal{S}} |\phi(\boldsymbol{x}) - f(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \right)^p &\leq \int_{\mathcal{S}} |\phi(\boldsymbol{x}) - f(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} = \sum_{i=1}^{m^d} \int_{Q_i} |f(\boldsymbol{\xi_i}) - f(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \\
&= \sum_{i=1}^{m^d} \int_{Q_i} |\mathrm{D}f(\boldsymbol{\psi_i}) \cdot (\boldsymbol{x} - \boldsymbol{\xi}_i)|^p \mathrm{d}\boldsymbol{x} \\
&\leq \sum_{i=1}^{m^d} |\mathrm{D}f(\boldsymbol{\psi_i})|^p \int_{Q_i} |\boldsymbol{x} - \boldsymbol{\xi}_i|^p \mathrm{d}\boldsymbol{x} \ .
\end{aligned}$$

As $\boldsymbol{x}, \boldsymbol{\psi}_i \in Q_i$, a sub-cube with edge length m^{-1}: $\int_{Q_i} |\boldsymbol{x} - \boldsymbol{\xi}_i|^p \mathrm{d}\boldsymbol{x} = O(m^{-p-d})$. Thus, we have

$$\left(\int_{\mathcal{S}} |\phi(\boldsymbol{x}) - f(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \right)^p \leq C m^{-p} \sum_{i=1}^{m^d} |\mathrm{D}f(\boldsymbol{\psi_i})|^p m^{-d} \leq C m^{-p} \left(\int_{\mathcal{S}} |\mathrm{D}f(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} + o(1) \right) \ .$$

Since smooth functions are dense in $W^{1,p}(\mathbb{R}^d)$ ([114, Thm. 2.3.2]), using the standard limiting argument the above inequality holds for $f \in W^{1,p}(\mathbb{R}^d)$. This establishes the desired result. □

Lemma 16 now provides the necessary result to extend the convergence rate bounds derived previously to the Sobolev case. As it can be seen from section 2.3.2, the $\mathcal{L}_1$ approximation error will influence the final rate upper bound only through the

exponent β in equation (2.20). As the Sobolev approximation error (2.56) is similar to the Holder class case for $\beta = 1$, we immediately have the following proposition:

Proposition 17. *Let $d \geq 2$ and $1 \leq \gamma \leq d-1$. Assume $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. random vectors over $[0,1]^d$ with density $f \in W^{1,p}(\mathbb{R}^d)$, $1 \leq p < \infty$, having support $\mathcal{S} \subset [0,1]^d$. Assume also that $f^{\frac{1}{2}-\frac{\gamma}{d}}$ is integrable over $\mathcal{S}$. Then, for any continuous quasi-additive Euclidean functional L_γ of order γ that satisfies the add-one bound (2.8)*

$$\left[E\left|L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|^{\kappa}\right]^{1/\kappa} \leq O\left(n^{-\frac{\alpha}{\alpha+1}\,\frac{1}{d}}\right) .$$

CHAPTER 3

Intrinsic Dimension and Entropy Estimation of Manifold Data

3.1 Introduction

Consider a class of natural occurring signals, e.g., recorded speech, audio, images, or videos. Such signals typically have high extrinsic dimension, e.g., as characterized by the number of pixels in an image or the number of time samples in an audio waveform. However, most natural signals have smooth and regular structure, e.g. piecewise smoothness, that permits substantial dimension reduction with little or no loss of content information. For support of this fact one needs only consider the success of image, video and audio compression algorithms, e.g., MP3, JPEG and MPEG, or the widespread use of efficient computational geometry methods for rendering smooth three dimensional shapes.

A useful representation of a regular signal class is to model it as a set of vectors which are constrained to a smooth low dimensional manifold embedded in a high dimensional vector space. The manifold may in some cases be a linear, i.e., Euclidean, subspace but in general it is a non-linear curved surface. This raises the question of

how to infer this lower dimensional manifold structure from high-dimensional data (e.g., a sequence of images). In the recent past this problem has received substantial attention from researchers in machine learning, computer vision, signal processing and statistics [6,26,94,103,107,111], being coined as *manifold learning* by the pattern recognition and machine learning communities.

In a practical setting, the complexity of representing such manifolds in closed form is unmanageable and all that is available is a finite number of (possibly random) samples obtained from these manifolds. It is thus important to be able to determine fundamental properties of manifolds directly from this finite representation, without resorting to cumbersome algorithms that first perform manifold reconstruction. In this chapter we address the problem of estimating the *intrinsic dimension* of a manifold and the *intrinsic entropy* of the measured manifold random samples. These two quantities measure the geometric and statistical complexity of the underlying manifold space and play a central role in many applications, ranging from computational biology [27] to image processing [40].

Formally, the intrinsic dimension of a manifold is the dimension of the vector space that is homeomorphic to local neighborhoods of the manifold [13]. Informally, intrinsic dimension describes how many "degrees of freedom" are necessary to generate the observed data. The classical way to estimate such a quantity is based on linear projection techniques [44]: a linear map is explicitly constructed and dimension is estimated by applying principal component analysis (PCA), factor analysis, or multidimensional scaling (MDS) to analyze the eigenstructure of the data. These methods estimate dimension by looking at the magnitude of the eigenvalues of the data covariance and determining in some *ad-hoc* fashion the number of such eigenvalues necessary to describe most of the data. As they do not account for non-linearities, linear methods tend to overestimate intrinsic dimension. Both nonlinear PCA [52]

methods and the ISOMAP [103] try to circumvent this problem but they still rely on unreliable and costly eigenstructure estimates. Other methods have been proposed, ranging from fractal dimension [12], estimating packing numbers [49] to a maximum likelihood approach [62].

When the samples are drawn from a large population of signals one can interpret them as realizations from a multivariate distribution supported on the manifold. The intrinsic entropy of random samples obtained from a manifold is an information theoretic measure of the complexity of this distribution. As this distribution is singular in the higher dimensional embedding space it has zero entropy as defined by the standard Lebesgue integral over the embedding space. However, when defined as a Lebesgue integral restricted to the lower dimensional manifold the entropy can be finite. This finite intrinsic entropy can be useful for exploring data compression over the manifold, registering medical images or geographical information [70] or, as suggested in [40], clustering of multiple sub-populations on the manifold.

The goal of this chapter is to develop an algorithm that jointly estimates both the intrinsic dimension and intrinsic entropy on the manifold, without knowing the manifold description, given only a set of random sample points. Our approach is based on minimal Euclidean graph methods. Specifically: construct a Euclidean k-nearest neighbors (k-NN) graph or a geodesic minimal spanning tree (GMST) over all the sample points and use its growth rate to estimate the intrinsic dimension and entropy by simple linear least squares and method of moments procedure. This approach allows for the estimation of the desired quantities using algorithms with low computational complexity that avoid reconstructing the manifold or estimating multivariate distributions.

The remainder of this chapter is organized as follows. In Section 3.2 we discuss the k-NN graph and GMST together with asymptotic results for Euclidean spaces.

Section 3.3 extends these results to Riemann manifolds. The proposed algorithms are described in Section 3.4. Experimental results are reported in Section 3.5. The technical proofs of the main results presented here are complied in Sections **??** to .

3.2 Minimal Graphs on Euclidean Spaces

Let $\mathcal{X}_n = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ be n independent identically distributed (i.i.d.) random vectors in a compact subset of $\mathbb{R}^d$, with multivariate Lebesgue density f. $\mathcal{X}_n$ will also be called the set of random vertices.

As discussed in Chapter 2, by solving certain optimization problems on the set $\mathcal{X}_n$, one can obtain special graph constructions. One such example is the k-NN graph. Start by defining the (1-)nearest neighbor of $\boldsymbol{X}_i$ in $\mathcal{X}_n$ as

$$\arg \min_{\boldsymbol{X} \in \mathcal{X}_n \setminus \{\boldsymbol{X}_i\}} d(\boldsymbol{X}, \boldsymbol{X}_i) \ ,$$

where distances between points are measured in terms of some suitable distance function $d(\cdot, \cdot)$. For general integer $k \geq 1$, the k-nearest neighbor of a point is defined in a similar way. The k-NN graph puts an edge between each point in $\mathcal{X}_n$ and its k-nearest neighbors. Let $\mathcal{N}_{k,i}(\mathcal{X}_n)$ be the set of k-nearest neighbors of $\boldsymbol{X}_i$ in $\mathcal{X}_n$. The total edge length of the k-NN graph is defined as:

$$L_\gamma^{k\text{-NN}}(\mathcal{X}_n) = \sum_{i=1}^{n} \sum_{\boldsymbol{X} \in \mathcal{N}_{k,i}(\mathcal{X}_n)} d^\gamma(\boldsymbol{X}, \boldsymbol{X}_i) \ , \tag{3.1}$$

where γ is a power weighting constant.

Another example is the MST problem, where the goal is to find a graph of minimum total edge length among the graphs $\mathcal{T}$ which span the sample $\mathcal{X}_n$. The minimum

total edge length is defined as:

$$L_\gamma^{\mathrm{MST}}(\mathcal{X}_n) = \min_{T\in\mathcal{T}} \sum_{e\in T} w^\gamma(e) \ , \tag{3.2}$$

where e is an edge in the graph and $w(e)$ is its weight. If edge e connects points $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ in $\mathcal{X}_n$, then its weight is $w(e) = d(\boldsymbol{X}_i, \boldsymbol{X}_j)$.

If $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^d$ and $d(\boldsymbol{X}, \boldsymbol{Y}) = |\boldsymbol{X} - \boldsymbol{Y}|$, i.e., the Euclidean distance between $\boldsymbol{X}$ and $\boldsymbol{Y}$, then both the MST graph and the k-NN graph fall under the framework of continuous quasi-additive Euclidean functionals discussed in Chapter 2. By showing that they satisfy subadditive, superadditive and continuous properties, their almost sure (a.s.) asymptotic behavior (also convergence in the mean) follows easily from the *umbrella* theorems for such graphs (cf. Chapter 2):

Theorem 1 ([81, 109]). *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. random vectors with values in $[0,1]^d$ and Lebesgue density f. Let $d \geq 2$, $1 \leq \gamma < d$ and define $\alpha = (d-\gamma)/d$. Then*

$$\lim_{n\to\infty} \frac{L_\gamma(\mathcal{X}_n)}{n^\alpha} = \beta_{d,L_\gamma} \int_{[0,1]^d} f^\alpha(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \qquad a.s. \ ,$$

where $L_\gamma(\mathcal{X}_n)$ is given by equation (3.1) or (3.2) with Euclidean distance, and, β_{d,L_γ} is a constant independent of f. Furthermore, the mean length $E\left[L_\gamma(\mathcal{X}_n)\right]/n^\alpha$ converges to the same limit.

Theorem 1 indicates that the limiting behavior of the graph length functional is determined by the *extrinsic* Rényi α-entropy of the multivariate Lebesgue density f:

$$H_\alpha^{\mathbb{R}^d}(f) = \frac{1}{1-\alpha} \log \int_{\mathbb{R}^d} f^\alpha(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \ . \tag{3.3}$$

In the limit, when $\alpha \to 1$ the usual Shannon entropy, $-\int_{\mathbb{R}^d} f(\boldsymbol{x}) \log f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$, is obtained. This remarkable asymptotic behavior motivates the name *entropic graphs*

given in [40].

Assume now that the random set $\mathcal{Y}_n = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\}$ is constrained to lie on a compact smooth m-dimensional manifold $\mathcal{M}$. The distribution of $\boldsymbol{Y}_i$ becomes singular with respect to Lebesgue measure and an application of Theorem 1 results in a zero limit for the length functional of the particular graph. However, this behavior can be modified by changing the way distances between points are measured. For this purpose, we use the framework of Riemann manifolds.

3.3 Entropic Graphs on Riemann Manifolds

Given a smooth manifold $\mathcal{M}$, a Riemann metric g is a mapping which associates to each point $\boldsymbol{y} \in \mathcal{M}$ an inner product $g_{\boldsymbol{y}}(\cdot, \cdot)$ between vectors tangent to $\mathcal{M}$ at $\boldsymbol{y}$ [13]. A *Riemann manifold* $(\mathcal{M}, g)$ is just a smooth manifold $\mathcal{M}$ with a given Riemann metric g. As an example, when $\mathcal{M}$ is a submanifold of the Euclidean space $\mathbb{R}^d$, the naturally induced Riemann metric on $\mathcal{M}$ is just the usual dot product between vectors.

A Riemann metric g endows $\mathcal{M}$ with a distance $d_g(\cdot, \cdot)$ via geodesics and a measure μ_g via the volume element [13]. Given the geodesic distance, one can define nearest neighbor relations or edge weights in terms of d_g instead of the usual Euclidean distance $|\cdot|$ and, consequently, define the total edge length $L_\gamma(\mathcal{Y}_n)$ as in (3.1) or (3.2), with the correspondence $d \to d_g$.

We can now extend Theorem 1 to general compact Riemann manifolds. This extension, Theorem 2, states that the asymptotic behavior of $L_\gamma(\mathcal{Y}_n)$ is no longer determined by the density of $\boldsymbol{Y}_i$ relative to the Lebesgue measure of $\mathbb{R}^d$, but depends instead on the the density of $\boldsymbol{Y}_i$ relative to μ_g.

Theorem 2. *Let $(\mathcal{M}, g)$ be a compact smooth Riemann m-dimensional manifold.*

Suppose $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ *are i.i.d. random elements of* $\mathcal{M}$ *with bounded density* f *relative to* μ_g*. Let* L_γ *be the total edge length of the MST graph or the* k*-NN graph with lengths computed using the geodesic distance* d_g*. Assume* $m \geq 2$*,* $1 \leq \gamma < m$ *and define* $\alpha = (m - \gamma)/m$*. Then,*

$$\lim_{n\to\infty} \frac{L_\gamma(\mathcal{Y}_n)}{n^\alpha} = \beta_{m,L_\gamma} \int_{\mathcal{M}} f^\alpha(\boldsymbol{y})\, \mu_g(\mathrm{d}\boldsymbol{y}) \qquad a.s.\ , \tag{3.4}$$

where β_{m,L_γ} *is a constant independent of* f *and* $\mathcal{M}$*. Furthermore, the mean length* $E\left[L_\gamma(\mathcal{Y}_n)\right]/n^\alpha$ *converges to the same limit.*

Now, the limiting behavior of $L_\gamma(\mathcal{Y}_n)$ is related to the *intrinsic* Rényi α-entropy of the multivariate density f on $\mathcal{M}$:

$$H_\alpha^{(\mathcal{M},g)}(f) = \frac{1}{1-\alpha} \log \int_{\mathcal{M}} f^\alpha(\boldsymbol{y})\, \mu_g(\mathrm{d}\boldsymbol{y})\ . \tag{3.5}$$

An immediate consequence of Theorem 2 is that, for known m,

$$\hat{H}_\alpha^{(\mathcal{M},g)}(\mathcal{Y}_n) = \frac{m}{\gamma} \left[\log \frac{L_\gamma(\mathcal{Y}_n)}{n^{(m-\gamma)/m}} - \log \beta_{m,L_\gamma} \right] \tag{3.6}$$

is an asymptotically unbiased and strongly consistent estimator of the intrinsic α-entropy $H_\alpha^{(\mathcal{M},g)}(f)$.

The proof of Theorem 2 is given in Appendix 3.7. The intuition behind it comes from the fact that a Riemann manifold $\mathcal{M}$, with associated distance and measure, looks locally like $\mathbb{R}^m$ with Euclidean distance $|\cdot|$ and Lebesgue measure λ. This implies that on small neighborhoods of the manifold, the total edge length $L_\gamma(\mathcal{Y}_n)$ behaves like a Euclidean length functional. As $\mathcal{M}$ is assumed compact, it can be covered by a finite number of such neighborhoods. This fact, together with subadditive and superadditive properties [109] of L_γ, allows for repeated applications of

Theorem 1 resulting in (3.4).

3.3.1 Approximating Geodesic Distances on Submanifolds of $\mathbb{R}^d$

Although Theorem 2 provides a characterization of the asymptotic behavior of entropic graphs over random points supported on a manifold, one further step is missing in order to make it applicable to a wide class of practical problems. This extra step comes from the computation of the length functionals which depends on finding geodesic distances between sample points, which in turn require knowing the manifold $\mathcal{M}$. However, in the general manifold learning problem, $\mathcal{M}$ (or any representation of it) is not known in advance. Consequently, the geodesic distances between points on $\mathcal{M}$ cannot be computed exactly and have to be estimated solely from the data samples.

In [18], the geodesic minimal spanning tree (GMST) algorithm was proposed, where the pairwise geodesic distances between sample points are estimated by running Dijkstra's shortest path algorithm over a global graph G of "neighborhood relations" among all sample points of the manifold. Two methods, called the ϵ-rule and the k-rule [103], are available for constructing G. The first method connects each point to all points within some fixed radius ϵ and the other connects each point to all its k-nearest neighbors. The graph G defining the connectivity of these local neighborhoods is then used to approximate the geodesic distance between any pair of points as the shortest path through G that connects them. Finally, this results in a distance matrix whose (i, j) entry is the geodesic distance estimate for the (i, j)-th pair of points. If $\hat{d}(\boldsymbol{Y}_i, \boldsymbol{Y}_j)$ is the estimate of the geodesic length of edge $e_{ij} = (\boldsymbol{Y}_i, \boldsymbol{Y}_j)$ obtained by this algorithm, then the GMST is defined as the minimal

graph over $\mathcal{Y}_n$ whose length is:

$$\hat{L}_\gamma^{\mathrm{GMST}}(\mathcal{Y}_n) = \min_{T \in \mathcal{T}} \sum_{e \in T} \hat{d}^\gamma(e) \ . \tag{3.7}$$

By using geodesic information, the GMST length functional encodes global structure about the nonlinear manifold. The geodesic distances between sample points on the manifold are uniformly well approximated by $\hat{d}$ in the following sense:

Theorem 3. *Let $(\mathcal{M}, g)$ be a compact Riemann submanifold of $\mathbb{R}^d$. Suppose $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are i.i.d. random vectors of $\mathcal{M}$, with density bounded away from zero. Then, with probability 1,*

$$\max_{\substack{1 \le i,j \le n \\ i \neq j}} \left| \frac{\hat{d}(\boldsymbol{Y}_i, \boldsymbol{Y}_j)}{d_g(\boldsymbol{Y}_i, \boldsymbol{Y}_j)} - 1 \right| \to 0 \quad \textit{as } n \to \infty \ . \tag{3.8}$$

This theorem is proven in Appendix 3.8. We remark that there exist alternative algorithms for computing geodesic distances that can also provide guarantees similar to Theorem 3. Of particular interest for future work is the method proposed in [65] for estimating geodesic distances that accounts for noisy samplings of the manifold.

Unlike the MST, the k-NN graph is only influenced by local distances. For fixed k, the maximum nearest neighbor distance of all points in $\mathcal{Y}_n$ goes to zero as the number n of samples increases. For n sufficiently large, this implies that the k-NN of each point will fall in a neighborhood of the manifold where geodesic curves are well approximated by the corresponding straight lines between end points. This suggests using simple Euclidean k-NN distances ($|\boldsymbol{Y}_i - \boldsymbol{Y}_j|$) as surrogates for the corresponding true nearest neighbor geodesic distances ($d(\boldsymbol{Y}_i, \boldsymbol{Y}_j)$). In fact, we prove in Appendix 3.9 that the geodesic k-NN distances are uniformly well approximated by the corresponding Euclidean k-NN distances in the following sense:

Theorem 4. *Let $(\mathcal{M}, g)$ be a compact Riemann submanifold of $\mathbb{R}^d$. Suppose $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$*

are i.i.d. random vectors of $\mathcal{M}$. Then, with probability 1,

$$\max_{\substack{1\le i\le n \\ \boldsymbol{Y}\in\mathcal{N}_{k,i}(\mathcal{Y}_n)}} \left| \frac{|\boldsymbol{Y}-\boldsymbol{Y}_i|}{d_g(\boldsymbol{Y},\boldsymbol{Y}_i)} - 1 \right| \to 0 \quad \textit{as } n\to\infty \ . \tag{3.9}$$

Finally, the asymptotic behavior of the GMST or the Euclidean k-NN graph is a simple consequence of Theorem 2 and Theorems 3 and 4:

Corollary 5. *Let $(\mathcal{M}, g)$ be a compact smooth Riemann m-dimensional manifold. Suppose $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n$ are i.i.d. random elements of $\mathcal{M}$ with bounded density f relative to μ_g. Let $\hat{L}_\gamma$ be the total edge length of the GMST graph or the Euclidean k-NN graph defined over $\mathcal{Y}_n$. Then,*

$$\lim_{n\to\infty} \frac{\hat{L}_\gamma(\mathcal{Y}_n)}{n^\alpha} = \beta_{m,L_\gamma} \int_{\mathcal{M}} f^\alpha(\boldsymbol{y})\, \mu_g(\mathrm{d}\boldsymbol{y}) \qquad a.s. \ , \tag{3.10}$$

where β_{m,L_γ} is a constant independent of f and $\mathcal{M}$. Furthermore, the mean length $E\left[L_\gamma(\mathcal{Y}_n)\right]/n^\alpha$ converges to the same limit.

Proof. For example, for the k-NN case,

$$\hat{L}_\gamma(\mathcal{Y}_n) = \sum_{i=1}^{n} \sum_{\boldsymbol{Y}\in\mathcal{N}_{k,i}} \left(\frac{|\boldsymbol{Y}-\boldsymbol{Y}_i|}{d_g(\boldsymbol{Y},\boldsymbol{Y}_i)} \right)^\gamma d_g^\gamma(\boldsymbol{Y},\boldsymbol{Y}_i) \ .$$

The uniform convergence expressed by Theorem 4 implies that

$$\hat{L}_\gamma(\mathcal{Y}_n) = (1+o(1))^\gamma L_\gamma(\mathcal{Y}_n) \ .$$

Corollary 5 now follows from an application of Theorem 2. The GMST case is similar. □

We remark that Corollary 5 differs from Corollary 1 presented in [18], in that the latter discusses the asymptotic behavior of the total edge length of the MST

as a function of the samples embedded on the m-dimensional Euclidean space that parameterizes the manifold (assuming a global conformal mapping), as opposed to the samples supported on the manifold itself considered here.

With regards to computational complexity, the geodesic free property of the k-NN algorithm makes it computationally inexpensive as compared with other manifold learning algorithms. In this case, complexity is dominated by determining nearest neighbors, which can be done in $O(n \log n)$ time for n sample points. This contrasts with the GMST, which, as forISOMAP, requires a costly $O(n^2 \log n)$ implementation of the geodesic pairwise distance estimation step.

3.4 Joint Intrinsic Dimension/Entropy Estimation

The asymptotic characterization of the GMST or k-NN length functional stated in Corollary 5 provides the framework for developing consistent estimators of both intrinsic dimension and entropy. The key observation is to notice that the growth rate of the length functional is strongly dependent on m while the constant in the convergent limit is equal to the intrinsic α-entropy. We use this strong growth dependence as a motivation for a simple estimator of m. Define $l_n = \log \hat{L}_\gamma(\mathcal{Y}_n)$. According to Corollary 5, l_n has the following approximation

$$l_n = a \ \log n + b + \epsilon_n \ , \tag{3.11}$$

where

$$\begin{aligned} a &= (m - \gamma)/m \ , \\ b &= \log \beta_{m,L_\gamma} + \gamma/m \ H_\alpha^{(\mathcal{M},g)}(f) \ , \end{aligned} \tag{3.12}$$

$\alpha = (m - \gamma)/m$ and ϵ_n is an error residual that goes to zero a.s. as $n \to \infty$.

Using the additive model (3.11), we propose a simple non-parametric least squares strategy based on resampling from the population $\mathcal{Y}_n$ of points in $\mathcal{M}$. Specifically, let $p_1, \ldots, p_Q$, $1 \leq p_1 < \ldots, < p_Q \leq n$, be Q integers and let N be an integer that satisfies $N/n = \rho$ for some fixed $\rho \in (0, 1]$. For each value of $p \in \{p_1, \ldots, p_Q\}$ randomly draw N bootstrap datasets $\mathcal{Y}_p^j$, $j = 1, \ldots, N$, with replacement, where the p data points within each $\mathcal{Y}_p^j$ are chosen from the entire data set $\mathcal{Y}_n$ independently. From these samples compute the empirical mean of the functionals $\bar{L}_p = N^{-1} \sum_{j=1}^{N} \hat{L}_\gamma(\mathcal{Y}_p^j)$. Defining $\bar{\boldsymbol{l}} = [\log \bar{L}_{p_1}, \ldots, \log \bar{L}_{p_1}]^T$ we write down the linear vector model

$$\bar{\boldsymbol{l}} = A \begin{bmatrix} a \\ b \end{bmatrix} + \epsilon \tag{3.13}$$

where

$$A = \begin{bmatrix} \log p_1 & \ldots & \log p_Q \\ 1 & \ldots & 1 \end{bmatrix}^T .$$

We now take a method-of-moments (MOM) approach in which we use (3.13) to solve for the linear least squares (LLS) estimates $\hat{a}, \hat{b}$ of a, b followed by inversion of the relations (3.12). After making a simple large n approximation, this approach yields the following estimates:

$$\begin{aligned} \hat{m} &= \text{round}\{\gamma/(1 - \hat{a})\} \\ \hat{H}_\alpha^{(\mathcal{M},g)} &= \frac{\hat{m}}{\gamma} \left(\hat{b} - \log \beta_{\hat{m}, L_\gamma} \right) . \end{aligned} \tag{3.14}$$

By running the algorithm M times independently over the population $\mathcal{Y}_n$, one obtains M estimates, $\{\hat{m}_i, \hat{H}_i\}_{i=1}^M$, that can be averaged to obtain final regularized dimension and entropy estimators, $\hat{m} = \sum \hat{m}_i/M$ and $\hat{H} = \sum \hat{H}_i/M$. The role of

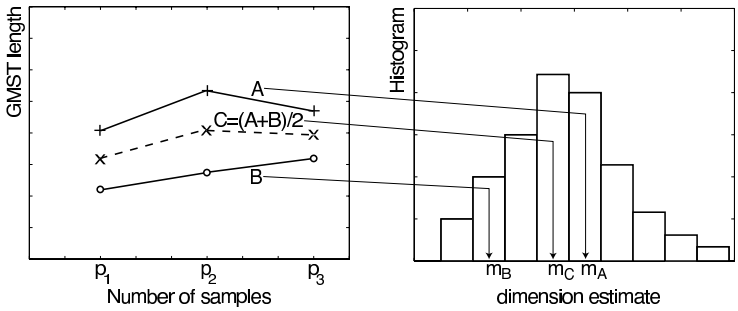


Figure 3.1: Computing the dimension estimators by averaging over the length functional values, i.e., $(M, N) = (1, N)$ (dashed line), or by averaging over the dimension estimates, i.e., $(M, N) = (M, 1)$ (solid lines).

parameter M, together with parameter N, is to provide a tradeoff between the bias and variance performance of the estimators for finite n. The two cases of interest (considered in the next section) are $(M, N) = (1, N)$ and $(M, N) = (M, 1)$. In the first case, the smoothing is performed on the graph length functional values before dimension and entropy are estimated, resulting in low variance but possibly high bias. In the second case, the smoothing is performed directly on the dimension and entropy estimates, resulting in higher variance but less bias.

Fig. 3.1 shows a graphical illustration of the smoothing step of the algorithm. Left panel shows $N = 2$ resampled graph lengths, labeled "+" and "o", along with their average labeled "x", for graphs built on $p_1 < p_2 < p_3$ randomly chosen vertices. For $(M, N) = (1, N)$, a linear least squares fit to the average graph trajectory, $C = (A + B)/2$, is used to compute the dimension estimate $\hat{m}_C$. For $(M, N) = (M, 1)$, dimension estimates $\hat{m}_A$ and $\hat{m}_B$ are computed from sub-trajectories A and B, forming a histogram from which a final estimate can be computed. The proposed algorithm is summarized in Table 3.1.

The constants β_{m,L_γ} in the above estimators depend only on m, γ and the particular entropic graph construction algorithm, e.g., GMST or k-NN. Due to the slow growth of $\{\beta_{m,L_\gamma}\}_{m>0}$ in the large n regime for which the above estimates were de-

Table 3.1: Graph resampling algorithm for estimating intrinsic dimension m and intrinsic entropy $H_\alpha^{(\mathcal{M},g)}$.

Initialize: Using entire database of signals $\mathcal{Y}_n$ construct NN graph (and geodesic distance matrix for GMST)

Select parameters: $M > 0$, $N > 0$, $Q > 0$ and $p_1 < \ldots < p_Q \leq n$

$\overline{m} = 0$, $\overline{H} = 0$;
for $M' = 1, \ldots, M$
  for $p = p_1, \ldots, p_Q$
    $\overline{L} = 0$;
    for $N' = 1, \ldots, N$
      Randomly select a subset of p signals $\mathcal{Y}_p$ from $\mathcal{Y}_n$;
      Compute graph total edge length L_p over $\mathcal{Y}_p$;
      $\overline{L} = \overline{L} + L_p$;
    end for
    Compute sample average graph length;
    $\hat{E}[\hat{L}(\mathcal{Y}_p)] = \overline{L}/N$;
  end for
  Estimate dimension $\hat{m}_{M'}$ and α-entropy $\hat{H}_{M'}$ from $\{\hat{E}[\hat{L}(\mathcal{Y}_p)]\}_{p=p_1}^{p_Q}$ via LLS/NLLS;
  $\overline{m} = \overline{m} + \hat{m}_{M'}$, $\overline{H} = \overline{H} + \hat{H}_{M'}$;
end for
$\hat{m} = \overline{m}/M$, $\hat{H} = \overline{H}/M$

rived, $\beta_{m,\gamma}$ is not required for the dimension estimator. On the other hand, the value of β_{m,L_γ} is required to obtain unbiased estimates of entropy. β_{m,L_γ} is the limit of the normalized length functional of the corresponding Euclidean entropic graph for a uniform distribution on the unit cube $[0,1]^m$. As, closed form expressions are not available, it can be determined by performing Monte Carlo simulations of the entropic graph length on the unit cube $[0,1]^m$ for uniform random samples. Another approach, applicable to the GMST, is to use analytical approximations and bounds for the MST over $[0,1]^d$, e.g. available in [109].

3.5 Experimental Results

We illustrate the performance of the entropic graph algorithm on manifolds of known dimension as well as on real high dimensional data ses of faces images and handwritten digits. In all the simulations we fixed the parameters $\gamma = 1$ and $p_1 = n - Q, \ldots, p_Q = n - 1$. With regards to intrinsic dimension estimation, we compare our algorithms to ISOMAP. In ISOMAP, similarly to PCA, intrinsic dimension is usually estimated by detecting a knee in the residual fitting error curve as a function of subspace dimension.

3.5.1 S-Shaped Surface

The first manifold considered is the standard 2-dimensional S-shaped surface [94] embedded in $\mathbb{R}^3$ (Figure 3.2). Figure 3.3 shows the evolution of the average GMST length $\bar{L}_n$ as a function of the number of samples, for a random set of i.i.d. points uniformly distributed on the surface.

To compare the dimension estimation performance of the GMST method to ISOMAP we ran a Monte Carlo simulation. For each of several sample sizes, 30 in-

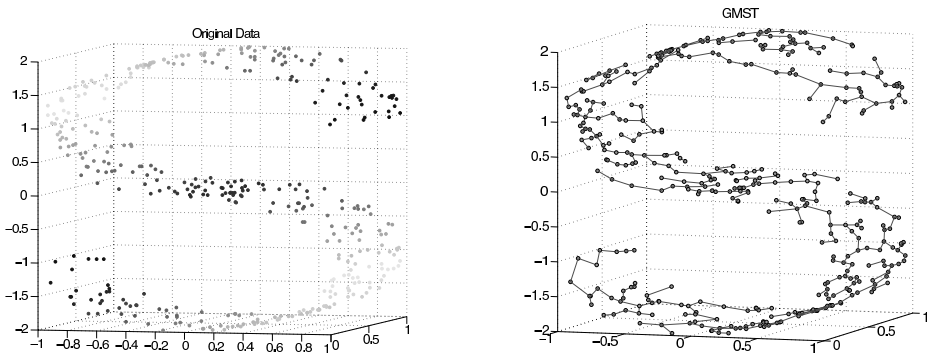
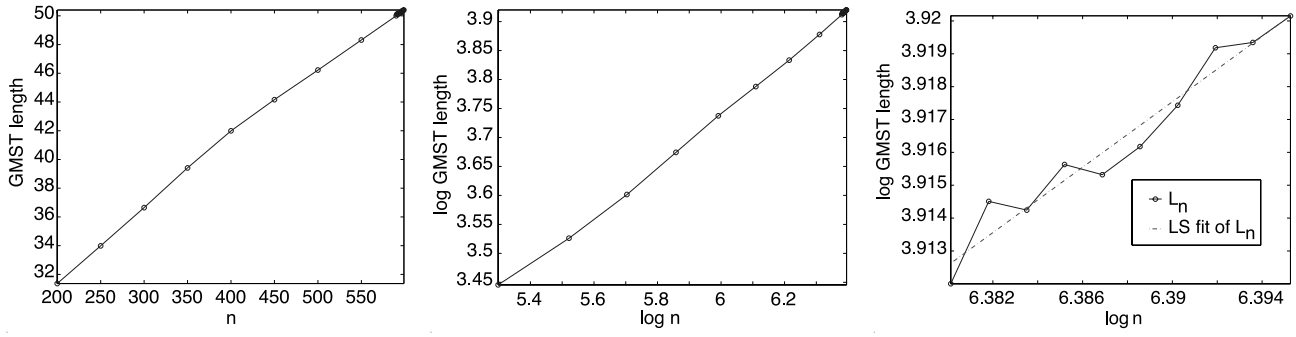


Figure 3.2: The S-shaped surface manifold and corresponding GMST ($k = 7$) graph on 400 sample points.

dependent sets of i.i.d. random vectors uniformly distributed on the surface were generated. We then counted the number of times that the intrinsic dimension was correctly estimated. To automatically estimate dimension with ISOMAP, we follow a standard PCA order estimation procedure. Specifically, we graph the residual variance of the MDS fit as a function of the PCA dimension and try to detect the "elbow" at which residuals cease to decrease "significantly" as estimated dimension increases [103]. The elbow detector is implemented by a simple minimum angle threshold rule. Table 3.2 shows the results of this experiment. As it can be observed, the GMST algorithm outperforms ISOMAP in terms of dimension estimation error rates for small sample sizes. Figure 3.4 shows the histogram of the entropy estimates for the same experiment.

Table 3.2: Number of correct ISOMAP and GMST dimension estimates over 30 trials as a function of the number of samples for the S-shaped manifold ($k = 7$).

n	200	400	600
ISOMAP	23	29	30
GMST ($M = 1, N = 5, Q = 10$)	29	30	30



(a) (b) (c)

Figure 3.3: Illustration of GMST dimension estimation for $(M, N) = (1, N)$: (a) plot of the average GMST length $\bar{L}_n$ for the S-shaped manifold as a function of the number of samples; (b) log-log plot of (a); (c) blowup of the last ten points in (b) and its linear least squares fit. The estimated slope is $\hat{a} = 0.4976$ which implies $\hat{m} = 2$. ($k = 7$, $M = 1$, $N = 5$).

Table 3.3: Number of correct dimension estimates over 30 trials as a function of the number of samples for the torus ($M = 1, N = 5, Q = 10$).

n	200	400	600
GMST	29	30	30
5-NN	29	30	30

3.5.2 Torus

Next, we consider the case of the 2-dimensional torus embedded in $\mathbb{R}^3$ (Figure 3.5). This manifold presents some challenges as it does not satisfy any of the usual isometric or conformal embedding constraints required by ISOMAP or Hessian eigenmaps [26], among others. We tested the algorithms over 30 generations of uniform random samples over the torus for different sample sizes n, and counted the number of correct dimension estimates. We note that in all the simulations ISOMAP always overestimated the intrinsic dimension as 3. The results for the GMST and k-NN are shown in Table 3.3. Table 3.4 shows the entropy estimates obtained by both methods on uniform samples supported on the torus. The true ($\alpha = 1/2$) entropy is

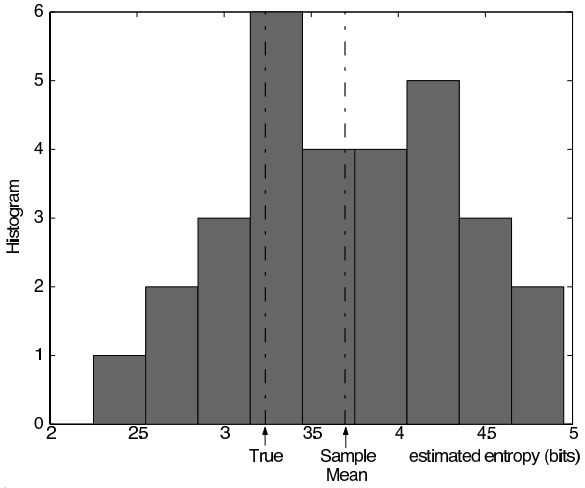


Figure 3.4: Histogram of GMST entropy estimates over 30 trials of 600 samples uniformly distributed on the S-shaped manifold ($k = 7, M = 1, N = 5, Q = 10$). True entropy ("true") was computed analytically from the area of S curve supporting the uniform distribution of manifold samples.

Table 3.4: Entropy estimates for the torus ($n = 600$, $M = 1, N = 5, Q = 10$).

	emp. mean	std. deviation
GMST	10.0	0.55
5-NN	9.6	0.93

$H_{1/2} = \log(120\pi^2) \approx 10.21$.

3.5.3 Hyper-Planes

We also investigated linear m-dimensional hyper-planes in $\mathbb{R}^{m+1}$ for which PCA methods are designed. We consider hyper-planes of the form $x_1 + \ldots + x_{m+1} = 0$. Table 3.5 shows the results of running a Monte Carlo simulation under the same conditions as in the previous subsection. When $M = 1$ (i.e., least squares applied to the average length functional values), the GMST method showed a tendency to underestimate the correct dimension at smaller sample sizes. However, by taking $N = 1$ instead (i.e., averaging of least squares dimension estimates), this negative bias was eliminated and the GMST performed as well as the ISOMAP, which was

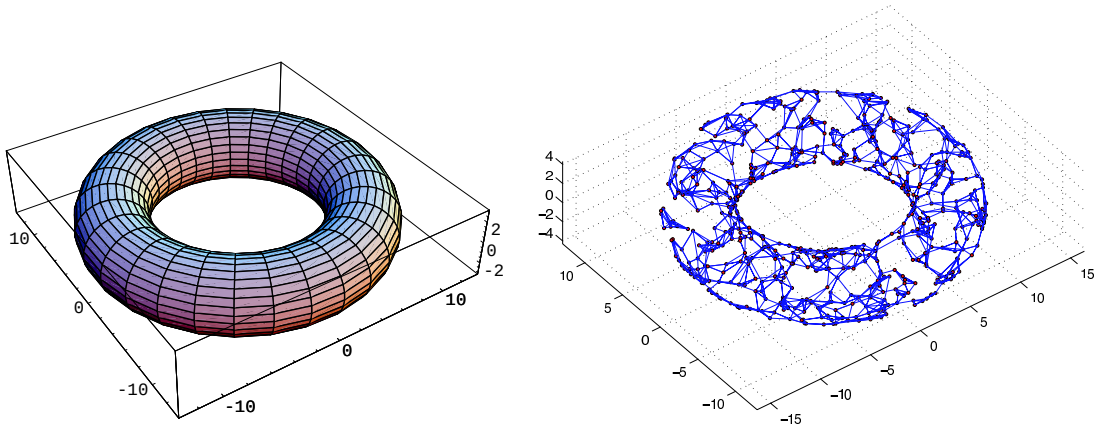


Figure 3.5: The 2D-torus and the 4-NN graph on 500 points sampled uniformly from the torus.

observed to correctly predict the dimension for all sample sizes investigated.

Of course, as expected, the number of samples required to achieve the same level of accuracy increases with the manifold dimension. This is the usual curse of dimensionality phenomenon: as the dimension increases, more samples are needed for the asymptotic regime in (3.10) to settle in and validate the limit in Corollary 5.

3.5.4 Yale Face Database B

We applied the GMST method to a real data set, and, consequently, of unknown manifold structure, intrinsic dimension and intrinsic entropy. We chose the set of 256 gray levels images of several individuals taken from the Yale Face Database B [29]. This is a publicly available database[1] containing a number of portfolios of face images under 585 different viewing conditions for each subject (Figure 3.6). Each portfolio consists of 9 poses and 65 illumination conditions (including ambient lighting) for each subject. The images were taken against a fixed background which we did not bother to segment out. This is justified since any fixed structures throughout

[1] `http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html`

Table 3.5: Number of correct GMST dimension estimates over 30 trials as a function of the number of samples for hyper-planes ($k = 5$).

Hyper-plane dimension	Q	M	N	n		
				600	800	1000
2	10	1	5	30	30	30
		5	1	30	30	30
3	10	1	5	24	24	27
		5	1	25	26	27
	15	1	10	30	30	30
		10	1	30	30	30
4	15	1	10	24	25	26
		10	1	27	28	28
	20	1	10	25	28	29
		10	1	29	29	30

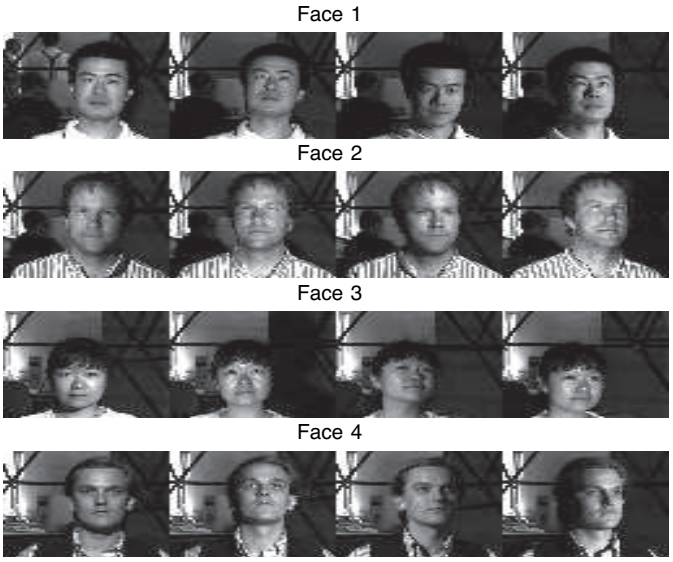


Figure 3.6: Samples from Yale face database B [29].

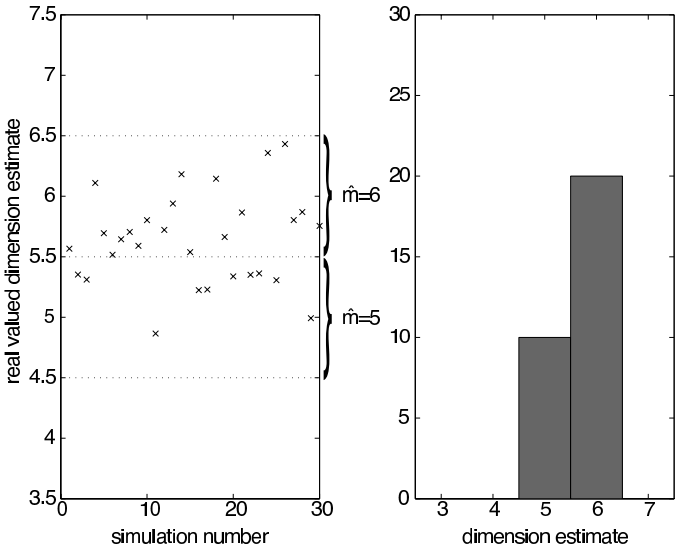


Figure 3.7: GMST real valued intrinsic dimension estimates and histogram for face 2 in the Yale face database B ($k = 7$, $M = 1$ $N = 10$, $Q = 20$).

the images would not change the intrinsic dimension or the intrinsic entropy of the dataset. We randomly selected 4 individuals from this data base and subsampled each person's face images down to a 64×64 pixels image. We normalized the pixel values between 0 and 1.

Figure 3.7 displays the results of running 30 trials of the algorithm using face 2. The first panel shows the real valued estimates of the intrinsic dimension, i.e., estimates obtained before the rounding operation in (3.14). Any value that falls in between the dashed lines will then be rounded to the integer at the midpoint. The second panel of Figure 3.7 shows the histogram for these rounded estimates over the 30 generated trials. The intrinsic dimension estimate is between 5 and 6. Figure 3.8 shows the corresponding residual variance plots used by ISOMAP to estimate intrinsic dimension. From these plots it is not obvious how to determine the "elbow" at which the residuals cease to decrease "significantly" with added dimensions. This illustrates one of the major drawbacks of ISOMAP (and other spectral based methods like PCA) as an intrinsic dimension estimator, as it relies on a specific eigenstruc-

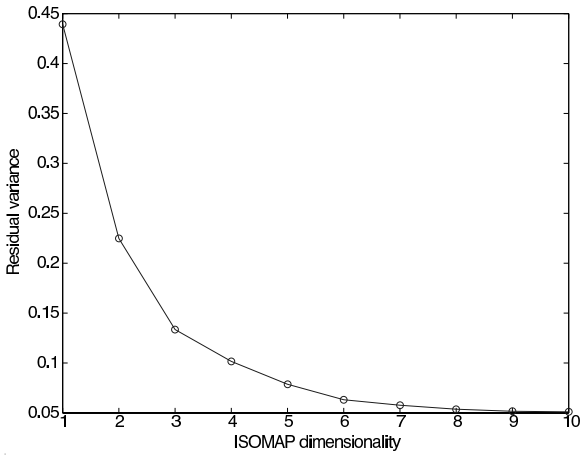


Figure 3.8: ISOMAP ($k = 7$) residual variance for face 2 in the Yale face database B.

Table 3.6: GMST dimension estimates $\hat{m}$ and entropy estimates $\hat{H}$ for four faces in the Yale Face Database B.

	Face1	Face2	Face3	Face 4
$\hat{m}$	6	6	7	7
$\hat{H}$ (bits)	24.9	26.4	25.8	28.0

ture that may not exist in real data. The simple minimum angle threshold rule on ISOMAP produced estimates between 3 and 6. Table 3.6 summarizes the results of the GMST method for the four faces. The intrinsic entropy estimates expressed in log base 2 were between 24.9 and 28 bits. As α is close to one, these values suggest that the portfolio of a person's face image could be accurately compressed using at most $28/(64 \times 64) \approx 0.007$ bits/pixel.

3.5.5 MNIST Database of Handwritten Digits

The MNIST database[2] consists of 256 gray levels images of handwritten digits obtained by optical character recognition. This publicly available database has became

[2] `http://yann.lecun.com/exdb/mnist/`

Figure 3.9: Samples from digits 0 to 9 in the MNIST database.

one of the benchmarks for testing new digit recognition algorithms [60], containing extensive test and training sets of all digits. Each digit in the database consists of a 28×28 pixel image that was size normalized and translated so that its center of mass lies in the center of the image. For the purpose of dimensionality estimation, we chose the first 1000 samples of digits 0 to 9 (Figure 3.9) in the training set.

Figure 3.10 shows the histogram of the dimension estimates for 30 simulations of the 5-NN algorithm applied to the samples of digits 0 to 9 (separately). Figure 3.11 shows the boxplot of the entropy estimates for the same scenario. Although the histograms show high variability, most of the estimates are between 9 and 15. It is interesting to notice that digit 1 exhibits the lowest dimension estimate, between 9 and 10, while all the other digits exhibit dimensions between 12 and 14. The lower complexity of digit 1 can also be seen from Figure 3.11, where its entropy estimate is much lower that all other digits. Also of interest is the bimodal behavior of the histogram of digit 7, with one mode concentrated at 10, 11 and the other at 13. After looking at the images selected in the realizations that resulted in the lower dimensional mode estimates, we realized that these images, although classified as a 7, are also very close to digit 1, thus contributing to lowering the dimension

estimates. This effect can also be observed in the boxplot of entropy estimates of Figure 3.11, where the high variance of the entropy estimate of digit 7 and consequent overlap of confidence intervals with digit 1 suggest the presence of images with a lower complexity.

For comparison purposes, we show in Figure 3.12 the eigenvalue plots for digits 2 and 3 used by ISOMAP to estimate intrinsic dimension. Even though it is not obvious how to assign a single dimension estimate from this plot - one of the main disadvantages of using spectral methods to estimate dimension - it is clear that the dataset should be at most 10-dimensional, as the residual variance ceases to decrease significantly after that value. The difference between the estimates predicted by entropic graphs and ISOMAP might be justified by the isometric assumption required by ISOMAP. The digits database contains nonlinear transformations, such as width distortions of each digit, that are not described by isometries. As consequence, ISOMAP underestimated these extra degrees of freedom, resulting in a lower dimension estimate than the entropic graphs, that are valid for a broader class of manifolds.

Finally, we present in Figure 3.13 the results of applying the proposed algorithm to the merged samples of digits 2 and 3. As it can be seen, the histogram of the dimension estimates shows an increase of its mode by one, being dominated by the dimensionality of the most complex digit (3). The entropy boxplot shows an increase of the median entropy estimate by roughly one bit. This should be expected, as compressing the augmented data set requires only one extra bit to identify which digit is being coded and then the individual codes for each digit can be used.

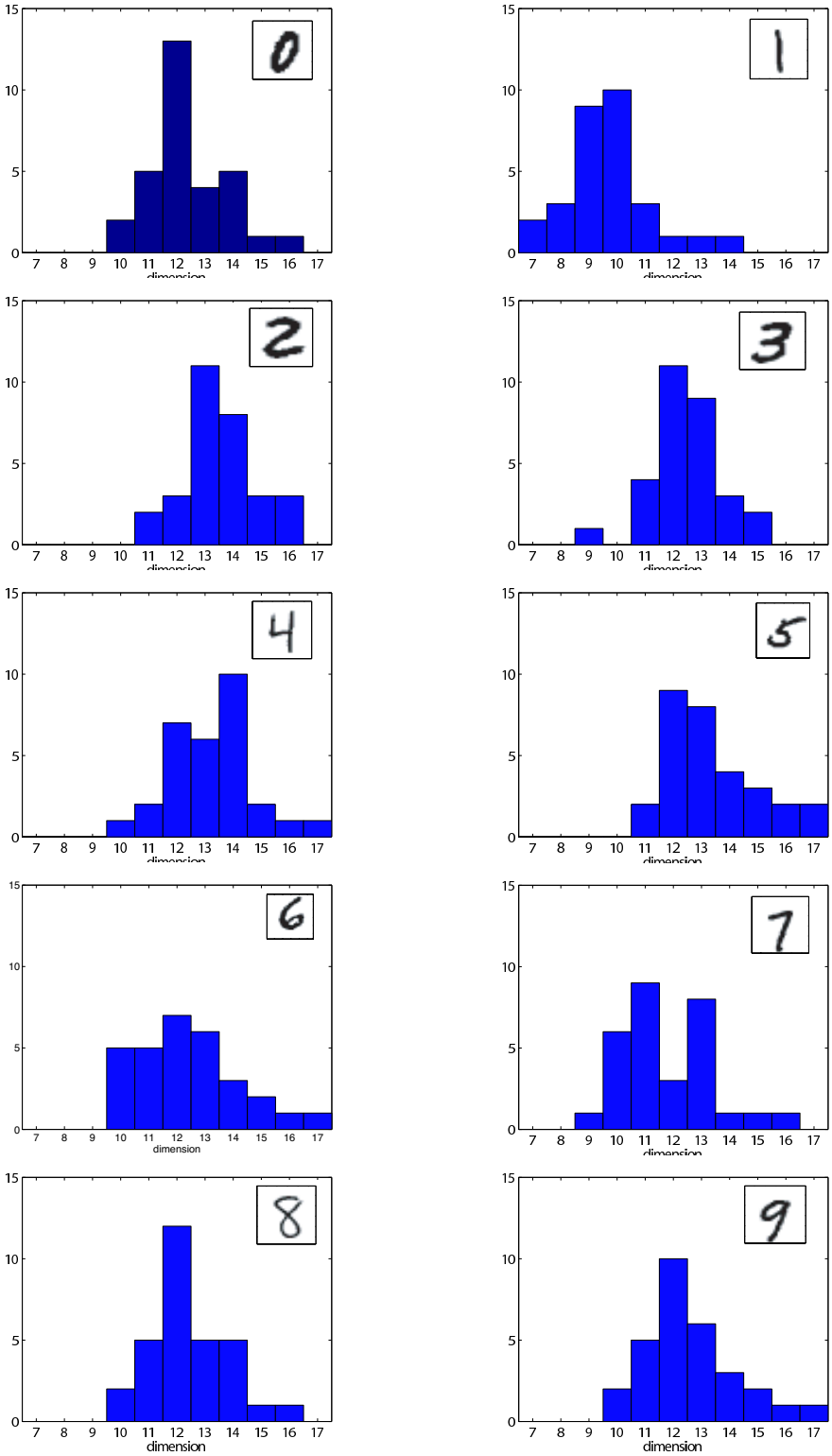


Figure 3.10: Histograms of intrinsic dimensionality estimates for digits 0 to 9 in the MNIST database using a 5-NN graph ($M = 1, N = 10$, $Q = 15$).

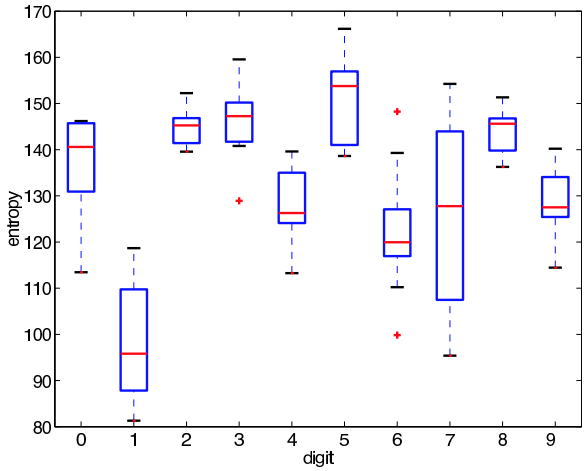


Figure 3.11: Boxplot of entropy estimates for digits 0 to 9 in the MNIST database using a 5-NN graph ($M = 1, N = 10$, $Q = 15$).

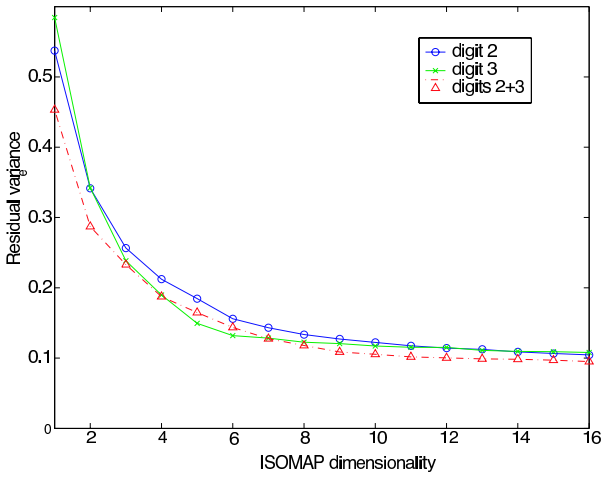


Figure 3.12: ISOMAP ($k = 6$) residual variance for digits 2 and 3 in the MNIST database.

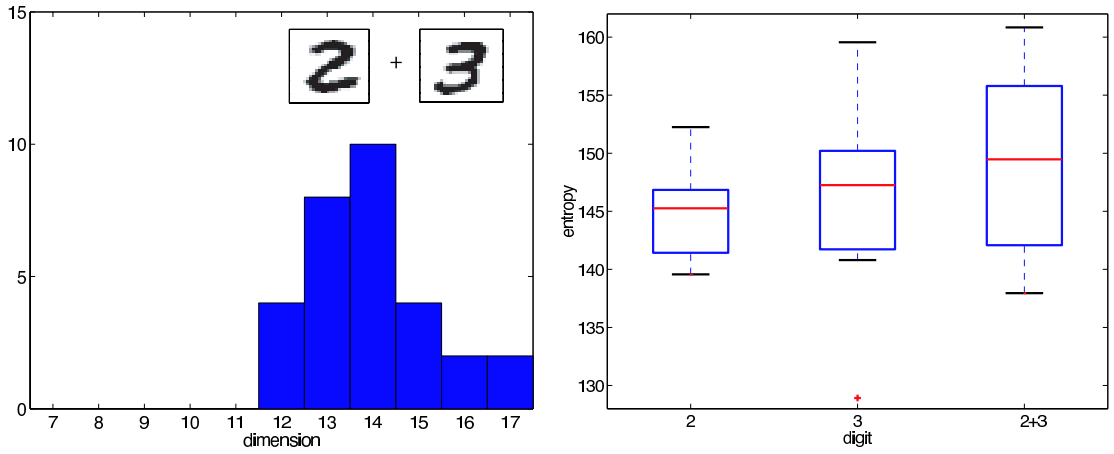


Figure 3.13: Histogram of intrinsic dimensionality estimates and boxplot of entropy estimates for digits 2+3 in the MNIST database using a 5-NN graph ($M = 1, N = 10$, $Q = 15$).

3.6 Conclusion

We have discussed the use of computational geometry graph constructions and geometric probability tools for the estimation of intrinsic dimension and entropy of shape spaces based solely on a finite random sampling of the underlying shapes. In particular, we have shown the strong statistical consistency of estimators based on k-nearest neighbor graphs or minimal spanning trees under the very general assumption of high dimensional data supported on a compact Riemann manifold. These results provide a departure from usually strong assumptions of linear, isometric or conformal embeddings expressed in the previous literature on the subject.

We are currently working on extending the proposed methods to data sets that exhibit a varying complexity across the data, characterized by a changing intrinsic dimension. This will allow the analysis of interesting datasets, like images composed of textures of different complexity or computational biology models of protein interaction [27]. Future work also includes developing bias correction mechanisms to improve the bootstrapping resampling step of the algorithm and account for depen-

dencies in the sampling process.

3.7 Appendix: Proof of Theorem 2

In this section, Theorem 2 is proven. We first introduce two auxiliary lemmas and take a small detour to discuss Euclidean boundary functionals, which are a key tool in proving asymptotic results for continuous quasi-additive Euclidean functionals [109].

The first lemma formalizes the intuition that a a Riemann manifold $\mathcal{M}$, with associated distance d_g and measure μ_g, looks locally like $\mathbb{R}^m$ with Euclidean distance $|\cdot|$ and Lebesgue measure λ:

Lemma 6 ([80, Lemma 5.1]). *Let $(\mathcal{M}, g)$ be a smooth Riemann m-dimensional manifold. For any $\boldsymbol{x} \in \mathcal{M}$ and $\varepsilon > 0$, there exists a chart (U, ϕ) for $\mathcal{M}$, with $\boldsymbol{x} \in U$, such that*

$$(1+\varepsilon)^{-1} \, |\phi(\boldsymbol{y}) - \phi(\boldsymbol{z})| \leq d_g(\boldsymbol{y}, \boldsymbol{z}) \leq (1+\varepsilon) \, |\phi(\boldsymbol{y}) - \phi(\boldsymbol{z})| \qquad \forall \, \boldsymbol{y}, \boldsymbol{z} \in U \quad (3.15)$$

and for any measurable subset $B \subset U$

$$(1-\varepsilon) \, \lambda(\phi(B)) < \mu_g(B) < (1+\varepsilon) \, \lambda(\phi(B)) \, . \quad (3.16)$$

Recall that a chart (U, ϕ) consists of a neighborhood U such that $\phi : \mathcal{M} \cap U \to \mathbb{R}^m$ determines a parametric representation of $\mathcal{M} \cap U$ in the Euclidean m-dimensional space, i.e., for $\boldsymbol{y} \in \mathcal{M} \cap U$, $\phi(\boldsymbol{y})$ represents $\boldsymbol{y}$ in an Euclidean m-dimensional coordinate system.

Boundary Functionals on Jordan Measurable Sets

We now informally introduce the notions of boundary functional. For formal definitions and details, we refer the reader to [109].

By appropriate canonical modifications of an Euclidean subadditive functional $L(F)$, it is possible to construct an associated *boundary functional* $L_B(F, R)$ on any subset R of $[0,1]^d$ [109]. Informally, in a boundary functional all the edges connecting point on the boundary of R (∂R) have zero length, so that ∂R can be seen as single point: all edges joined to the boundary are joined to one another, or, in other words, joining edges using ∂R adds no additional cost to the functional.

The importance of boundary functionals resides in the fact that they are superadditve, a property that many of the standard total edge functionals lack. If R is partitioned into sets R_1 and R_2 then L_B is superadditive if

$$L_B(F, R) \geq L_B(F \cap R_1, R_1) + L_B(F \cap R_2, R_2) \ .$$

When R, R_1, R_2 are rectangles, translation invariance and homegeneity properties of any Euclidean functional, endow $L_B(\cdot, R)$ with a self similarity property, in a way that, for a uniform sample, the value of the functional on a set of the partition is statistically similar to its value on any other partition set. However, when R is an arbitrary set, this self similarity property is lost. We now show that if R is Jordan measurable a superadditive functional has the same type of asymptotic behavior as when R is a rectangle.

Lemma 7. *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. random vectors with values in $R \subset [0,1]^d$ and bounded Lebesgue density f. Assume R is Jordan measurable. Let $L_B(\cdot, R)$ be a*

continuous superadditive Euclidean boundary functional of order γ on $\mathbb{R}^d$. Then

$$\liminf_{n\to\infty} \frac{L_B(\mathcal{X}_n, R)}{n^\alpha} \geq \beta_{d,L} \int_R f^\alpha(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \qquad a.s. \tag{3.17}$$

Furthermore, the same result holds for the mean length $E\left[L_B(\mathcal{X}_n, R)\right]/n^\alpha$.

Proof. The proof of this result relies on the fact that a Jordan measurable set is "well approximated" from below by an union of disjoint cubes. We then use the known results about the behavior of Euclidean functionals over cubes.

Let $\varepsilon > 0$. As R is Jordan measurable, there exists a finite number of disjoint cubes $\{Q_i\}$ (with faces parallel to the axis) such that $Q = \cup_i Q_i \subset R$ and $\lambda(R \backslash Q) < \varepsilon$. By superadditivity,

$$L_B(\mathcal{X}_n, R) \geq \sum_i L_B(\mathcal{X}_n \cap Q_i, Q_i)\ . \tag{3.18}$$

Let $p_i = \int_{Q_i} f\, \mathrm{d}\lambda$. By the strong law of large numbers, $\mathcal{X}_n \cap Q_i$ consists of $n(p_i + o(1))$ i.i.d. points in Q_i distributed with density $p_i^{-1} f$. By the usual umbrella theorem,

$$\frac{L_B(\mathcal{X}_n \cap Q_i, Q_i)}{(p_i n)^\alpha} \to \beta_{d,L} \int_{Q_i} (p_i^{-1} f)^\alpha\, \mathrm{d}\lambda \qquad a.s. \tag{3.19}$$

We also have

$$\left| \int_R f\, \mathrm{d}\lambda - \int_Q f\, \mathrm{d}\lambda \right| \leq \|f\|_\infty\, \lambda(R \setminus Q) < \varepsilon\, \|f\|_\infty\ , \tag{3.20}$$

where $\|f\|_\infty = \sup\{f(\boldsymbol{x}) : \boldsymbol{x} \in R\}$ is finite by assumption. Combining (3.18), (3.19) and (3.20) results in

$$\liminf_{n\to\infty} \frac{L_B(\mathcal{X}_n, R)}{n^\alpha} \geq \beta_{d,L} \sum_i \int_{Q_i} f^\alpha\, \mathrm{d}\lambda \geq \beta_{d,L} \left(\int_R f^\alpha\, \mathrm{d}\lambda - \varepsilon\, \|f\|_\infty \right)\ .$$

Letting $\varepsilon \to 0$ produces the desired result. □

Remark 1. If L_B is close in the mean (2.6) to the underlying smooth subadditive Euclidean functional, then lim inf and the inequality in equation (3.17) can be replaced, respectively, by lim and an equality.

Proof of Theorem 2

Before proving Theorem 2, we note that both the MST and the k-NN functional and respective boundary functionals defined on a Riemann manifold satisfy strong forms of subadditivity and superadditivity. Namely, if $R_1, R_2 \in \mathcal{M}$ are arbitrary sets that partition $\mathcal{M}$, then

$$L_B(F \cap R_1, R_1) + L_B(F \cap R_2, R_2) \leq L_B(F, \mathcal{M}) = L(F) \leq L(F \cap R_1) + L(F \cap R_2) + C \ , \tag{3.21}$$

where C is an error term independent of R_1 and R_2 (C is zero for the k-NN case). Note that the usual subadditivity and superadditivity conditions needed to prove umbrella theorems for Euclidean functionals only require that these conditions hold for partitions made of rectangles.

Proof of Theorem 2. Let $\varepsilon > 0$. For $\boldsymbol{x} \in \mathcal{M}$ let $(U_{\boldsymbol{x}}, \phi_{\boldsymbol{x}})$ be the chart specified by Lemma 6. Without loss of generality, $U_{\boldsymbol{x}}$ may be chosen such that $\phi_{\boldsymbol{x}}(U_{\boldsymbol{x}})$ is an open ball in $\mathbb{R}^m$ (this can be achieved by possibly shrinking the set $U_{\boldsymbol{x}}$ whose existence is guaranteed by Lemma 6). By compactness of $\mathcal{M}$, there exists a finite collection of such sets, say $\{U_i\}$, that cover $\mathcal{M}$. Define the set sequence $\{V_j\}$ by $V_1 = U_1$ and $V_j = U_j \setminus \cup_{1 \leq i \leq j-1} V_i$, for $j \geq 2$. The sets V_j are disjoint, form a partition of $\mathcal{M}$, and $V_j \subset U_j$, for all j.

Let $p_j = \int_{V_j} f \, \mathrm{d}\mu_g$ and $\mathcal{X}_{n,j} = \phi_j(\mathcal{Y}_n \cap V_j)$. By the strong law of large numbers,

$\mathcal{X}_{n,j}$ consists of $n(p_j + o(1))$ i.i.d. points in $\phi_j(V_j)$ distributed with density

$$g_j(\boldsymbol{u}) = p_j^{-1} h_j(\phi_j^{-1}(\boldsymbol{u})) f(\phi_j^{-1}(\boldsymbol{u})) \ , \quad \boldsymbol{u} \in \phi_j(V_j) \ ,$$

where h_j is the function defined in the proof of Lemma 6 in [80] (c.f. Lemma 5.1). h_j accounts for the differential changes in volume between V_j and $\phi_j(V_j)$, i.e., $\mu_g(B) = \int_{\phi(B)} h_j(\phi_j^{-1}(u))\,\mathrm{d}\boldsymbol{u}$, for $B \subset U_j$. Recall from [80] that $1 - \varepsilon < h_j(\boldsymbol{x}) < 1 + \varepsilon$ for $\boldsymbol{x} \in V_j$.

We are now ready to apply sub and superadditivity to the partition $\{V_j\}$. By (3.21)

$$\sum_j L_B(\mathcal{Y}_n \cap V_j, V_j) \le L_B(\mathcal{Y}_n, \mathcal{M}) = L(\mathcal{Y}_n) \le \sum_j L(\mathcal{Y}_n \cap V_j) + C' \ . \tag{3.22}$$

As the sets V_j were chosen such that the geodesic lengths and Euclidean lengths are ε-close, we have by (3.15)

$$L(\mathcal{Y}_n \cap V_j) \le (1 + \varepsilon)\, L(\mathcal{X}_{n,j}) \ . \tag{3.23}$$

As $L(\mathcal{X}_{n,j})$ satisfies the usual quasi-additive continuous Euclidean properties, it follows from the usual umbrella theorem that

$$\frac{L(\mathcal{X}_{n,j})}{(p_j\, n)^\alpha} \to \beta_{d,L} \int_{\phi_j(V_j)} g_j^\alpha(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} \qquad a.s. \tag{3.24}$$

Changing the integration back to μ_g and using the fact that h_j is $(1 \pm \varepsilon)$-valued,

$$\begin{aligned} p_j^\alpha \int_{\phi_j(V_j)} g_j^\alpha(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} &= \int_{\phi_j(V_j)} f^\alpha(\phi_j^{-1}(\boldsymbol{u}))\, h_j^{\alpha-1}(\phi_j^{-1}(\boldsymbol{u}))\, h_j(\phi_j^{-1}(\boldsymbol{u}))\,\mathrm{d}\boldsymbol{u} \\ &= \int_{V_j} f^\alpha(\boldsymbol{y})\, h_j^{\alpha-1}(\boldsymbol{y})\,\mu_g(\mathrm{d}\boldsymbol{y}) \\ &\leq (1-\varepsilon)^{\alpha-1} \int_{V_j} f^\alpha(\boldsymbol{y})\,\mu_g(\mathrm{d}\boldsymbol{y}) \end{aligned} \tag{3.25}$$

Combining the upper bound in (3.22) with (3.23)-(3.25), we get:

$$\limsup_{n\to\infty} \frac{L(\mathcal{Y}_n)}{n^\alpha} \leq (1+\varepsilon)(1-\varepsilon)^{\alpha-1} \int_{\mathcal{M}} f^\alpha(\boldsymbol{y})\,\mu_g(\mathrm{d}\boldsymbol{y})\ . \tag{3.26}$$

The lower bound implicit in equation (3.4) follows in a similar way. Start by noticing that, due to (3.15),

$$L_B(\mathcal{Y}_n \cap V_j, V_j) \geq (1+\varepsilon)^{-1} L_B(\mathcal{X}_{n,j}, \phi_j(V_j))\ .$$

Recall that V_j is a finite intersection of sets U_i with smooth boundary (U_i was constructed to be the inverse image of a ball trough the smooth transformation ϕ_j). So, the set $\phi_j(V_j)$ will have smooth piecewise boundary and, consequently, will be Jordan measurable. Lemma 7 can now be applied to conclude that:

$$\liminf_{n\to\infty} \frac{L_B(\mathcal{X}_{n,j}, \phi_j(V_j))}{(p_j\, n)^\alpha} \geq \beta_{d,L} \int_{\phi_j(V_j)} g_j^\alpha(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} \qquad a.s.$$

Repeating the same arguments used above, we have

$$\liminf_{n\to\infty} \frac{L(\mathcal{Y}_n)}{n^\alpha} \geq (1+\epsilon)^{-1}(1+\varepsilon)^{\alpha-1} \int_{\mathcal{M}} f^\alpha(\boldsymbol{y})\,\mu_g(\mathrm{d}\boldsymbol{y})\ . \tag{3.27}$$

Finally, combining equations (3.26) and (3.27) and letting $\epsilon \to 0$ establishes

Theorem 2. □

3.8 Appendix: Proof of Theorem 3

Here, we prove Theorem 3 for the case when geodesic distances are estimated using the "ϵ-rule" [9]. This rule estimates geodesic distances by running Dijkstra's shortest path algorithm over the graph constructed by putting an edge between each point and all points within a fixed radius ϵ. Of course, for the algorithm to be consistent as the number of samples n grows, ϵ has to decrease to 0 as $n \to \infty$. In particular, our proof shows that $\epsilon_n = o\left(n^{-\xi/m}\right)$, for some $0 < \xi < 1$, is sufficient to guarantee consistency.

Proof of Theorem 3. According to [9], proving the consistency result expressed by equation (3.8) reduces to showing that the "δ-sampling" condition holds with probability one. This condition states that for all $\boldsymbol{x} \in \mathcal{M}$ there is a sample $\boldsymbol{x}_i$ such that $d_g(\boldsymbol{x}, \boldsymbol{x}_i) \leq \delta$.

In the following, we use the same notation as defined in the *Sampling Lemma* of [9]. In particular, $B_i(\delta)$ is the metric ball in $\mathcal{M}$ of radius δ, centered at some point $\boldsymbol{p}_i$; $V_{min}(\delta)$ is the volume of the smallest metric ball in $\mathcal{M}$ of radius δ. For Riemann submanifolds of $\mathbb{R}^d$ without boundary, $V_{min}(\delta) \asymp \delta^m$; V is the volume of $\mathcal{M}$; $f_{min} = \inf_{\boldsymbol{y} \in \mathcal{M}} f(\boldsymbol{y}) > 0$.

Begin by covering $\mathcal{M}$ with a finite family of metric balls of radius $\delta/2$, choosing the centers $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots$ such that

$$\boldsymbol{p}_{i+1} \notin \cup_{j=1}^{i} B_i(\delta/2)$$

and stopping when this is no longer possible. As no two centers $\boldsymbol{p}_i$ are within dis-

tance $\delta/2$ of each other, the balls $B_i(\delta/4)$ are disjoint and, consequently, at most $V/V_{min}(\delta/4)$ points can be chosen before the process terminates.

The δ-sampling condition will be satisfied if each ball B_i contains at least one sample, as the diameter of B_i is δ and every $\boldsymbol{x} \in \mathcal{M}$ belongs to a ball B_i. The probability of this event is:

$$P(\delta\text{-sampling condition holds}) \geq P(\text{no ball } B_i \text{ is empty}) \geq 1 - \sum_i P(B_i \text{ is empty}) \ . \tag{3.28}$$

Under the i.i.d. assumption on the samples, the probability $P(B_i \text{ is empty})$ can be computed as:

$$\begin{aligned} P(B_i \text{ is empty}) = \left(1 - \int_{B_i} f \mathrm{d}\mu_g\right)^n &\leq \left(1 - V_{min}(\delta/2)\, f_{min}\right)^n \\ &\leq \exp\left\{-n\, V_{min}(\delta/2)\, f_{min}\right\} \ , \end{aligned} \tag{3.29}$$

where the last inequality follows from the inequality $\log(1-x) \leq -x$. Substituting equation (3.29) in (3.28) and using the asymptotic value for $V_{min}(\delta/2)$ results in:

$$\begin{aligned} P(\delta\text{-sampling condition holds}) &\geq 1 - \frac{V}{V_{min}(\delta/4)} \exp\left\{-n\, V_{min}(\delta/2)\, f_{min}\right\} \\ &= 1 - C_1\, V \delta^{-m} \exp\left\{-C_2\, f_{min}\, n\, \delta^m\right\} \ , \end{aligned} \tag{3.30}$$

where C_1 and C_2 are constants.

Now, choose $\delta = \delta_n$ as a function of the number of samples such that $\delta_n \to 0$ and $n\,\delta_n^m \to \infty$ as $n \to \infty$. For example, $\delta_n = O\left(n^{-\xi/m}\right)$, for some $0 < \xi < 1$, will satisfy these conditions. Then choose a sequence ϵ_n such that $\epsilon_n \to 0$ and $\epsilon_n/\delta_n \to 0$ as $n \to \infty$. For example, $\epsilon_n = o\left(n^{-\xi/m}\right)$. Given $\lambda > 0$, there exists an integer n_0 such that for all $n > n_0$, ϵ_n is small enough to satisfy conditions 5, 6 and 7 of *Main*

Theorem A of [9]. This theorem, together with equation (3.30), implies that

$$P\left(\max_{\substack{1\le i,j\le n\\ i\neq j}}\left|\frac{\hat{d}(\boldsymbol{Y}_i,\boldsymbol{Y}_j)}{d_g(\boldsymbol{Y}_i,\boldsymbol{Y}_j)}-1\right|\ge\lambda\right)\le C_1\,V\delta_n^{-m}\exp\left\{-C_2\,f_{min}\,n\,\delta_n^m\right\}\ ,$$

for $n > n_0$. As the choice of δ_n implies that $\sum_{n>n_0}\delta_n^{-m}\exp\{-C_2\,f_{min}\,n\,\delta_n^m\}<\infty$, the desired result follows by the Borel-Cantelli Lemma. □

3.9 Appendix: Proof of Theorem 4

Proof of Theorem 4. Without loss of generality, assume that $\mathcal{M}\in[0,1]^d$. We first prove that $M_{n,k}=M_{n,k}(\mathcal{Y}_n)$, the length of the longest k-NN link, converges to zero with probability 1.

Given $\varepsilon>0$, partition $[0,1]^d$ into a finite number of cubes, $\{Q_j\}$, with edge length at most ε. Let $p_j=\int_{Q_j\cap\mathcal{M}}f(\boldsymbol{y})\,\mu_g(\mathrm{d}\boldsymbol{y})$. By the strong law of large numbers, there will be $n\,(p_j+o(1))$ points in Q_j with probability 1. This implies, for $p_j>0$, that there exists an integer N_j such that for all $n>N_j$, $n\,(p_j+o(1))\ge k$. Let $N=\max_j N_j$. Ignoring the cubes with $p_j=0$ (with probability 1 they will have no points), each cube has at least k points for $n>N$. This implies that for all $n>N$, $M_{n,k}<O(\varepsilon)$, i.e, $M_{n,k}\to 0$ as $n\to\infty$. With this result in hand, Theorem 4 follows directly by an application of Corollary 4 from [9]. □

CHAPTER 4

Classification Constrained Dimensionality Reduction

4.1 Introduction

Following the approach of chapter 3, we continue the study of high dimensional complex data sets. Overlapping with the problem of inferring geometrical or statistical quantities from high dimensional data sets, as intrinsic dimension or entropy, is the problem of finding appropriate "compact" representations of the complex data. In this chapter, we address the problem of extracting lower-dimensional features relevant for classification tasks. By taking into account not only the geometric constraints resulting from a low dimensional manifold embedding of the data in a high dimensional space, but also the constraints resulting from labeled samples, we develop a general nonlinear dimensionality reduction algorithm, aimed at constructing a lower dimensional representation of the original data set.

Although part of a long and rich history, past approaches to the general problem of dimensionality reduction did not explicitly consider the manifold structure possibly present in many high dimensional data sets. It was only after the semi-

nal papers of Tenenbaum *et al* [103] and Roweis and Saul [94] that the usefulness of this approach captured the eye of researchers, resulting in a renewed interest in the area from the machine learning, computer science and statistics communities. Classical approaches to dimensionality reduction include Principal Component Analysis (PCA) [44] and Multidimensional Scaling (MDS) [21]. They are based on solving a global optimization problem, namely finding eigenvectors of a data similarity matrix, but can be reliably applied only when the data is linearly embedded in the high-dimensional spaces. Recent approaches of kernel PCA [97] allow for extra nonlinear relationships among the data, but still ignore any explicit manifold structure. Within the class of *manifold learning* algorithms, many methods have been proposed in the past five years. They range from methods aimed at preserving local manifold structure to global methods. Local methods include Locally Linear Embedding (LLE) [94], Laplacian Eigenmaps [6], Hessian Eigenmaps (HLLE) [26] and Local Space Tangent Analysis [111]. They are based on local approximation of the geometry of the manifold, still preserving the global optimization formulation. Global methods include ISOMAP [103] and Semidefinite Embedding (SDE) [107]. These methods preserve global manifold properties, like geodesic distances, but are restricted to strong isometric or locally isometric data embeddings.

Although dimensionality reduction is usually invoked as a tool to improve classification, regression, denoising or visualization tasks, among other applications, current algorithms do not use this information to find a particular lower dimensional representation of the data. For example, in the classification problem, the lower dimensional embeddings found by many popular algorithms generally induce a nonlinear mixing of the classes, resulting in an harder problem in the embedded domain than in the original high dimensional space. However, incorporating classification information in the specification of the data embedding can lead to improvements in

classification performance. In particular, by designing a classifier based on a "good" lower dimensional embedding of the data, instead of the high dimensional space, one might break the well known *curse of dimensionality*.

The goal of this chapter is to introduce a dimensionality reduction method where the class labels of data points having a manifold structure are incorporated in the construction of a lower dimensional data embedding. It seems intuitive that such class dependent manifold embedding algorithm can improve the performance of supervised and semi-supervised learning tasks. This is accomplished by modifying the Laplacian approach to manifold learning through the introduction of class dependent constraints.

Currently, the only other approaches to classification that take advantage of the manifold structure of the data are from the semi-supervised learning perspective [7, 101, 113]. However, the perspective of these approaches is one of regularization, instead of the dimensionality reduction perspective followed here.

The outline of this chapter is as follows. In Section 4.2 we formulate the problem of dimensionality reduction as a global optimization program and discuss Laplacian eigenmaps. Section 4.3 described the embedding method proposed by adding class label constraints to the optimization program. Some illustrative examples are presented in Section 4.4 and Section 4.5 describes the application of the proposed method to semi-supervised learning.

4.2 Graph Laplacians and Manifold Embeddings

Let $\mathcal{X}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^d$ be a set of n points constrained to lie on an m-dimensional submanifold $\mathcal{M}$ of $\mathbb{R}^d$. The manifold learning problem consists in finding an embedding of $\mathcal{X}_n$ into a subset $\mathcal{Y}_n = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ of a lower m-dimensional

space $\mathbb{R}^m$ (usually $m \ll d$), without any prior knowledge about $\mathcal{M}$ besides its finite sampling $\mathcal{X}_n$.

A common framework used to represent the geometric information about $\mathcal{M}$ carried by its sampling $\mathcal{X}_n$ is through the use of adjacency graphs. Let $G = (V, E)$ be an undirected weighted graph, whose vertex set V is given by the data points, i.e., $V = \mathcal{X}_n$, and E is the set of edges in the graph. The edge set E is associated with an $n \times n$ weight matrix $\boldsymbol{W}$ specifying adjacency relations between vertices, such that w_{ij} is a function of the similarity between points i and j. The weights are assumed nonnegative and symmetric. Although there are many choices for G, throughout this paper we consider nearest neighbor (NN) graphs with a weight matrix derived from the heat kernel [7]. The construction of this graph proceeds as follows:

1. For a fixed neighborhood parameter $k \in \mathbb{N}$, construct a k-NN graph on $\mathcal{X}_n$, i.e., put and edge between points i and j if i is one of the k-NN's of j or j is one of the k-NN's of i.

2. For a fixed scale parameter $\epsilon > 0$, assign weight

$$w_{ij} = \exp\left\{-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\epsilon}\right\} ,$$

   if vertices i and j are connected and $w_{ij} = 0$ otherwise.

Following the Laplacian eigenmaps approach [6], we formulate manifold learning as the problem of minimizing the cost function

$$E(\mathcal{Y}_n) = \sum_{ij} w_{ij} \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2 \tag{4.1}$$

in the embedding points $\mathcal{Y}_n \subset \mathbb{R}^d$. This cost function naturally accounts for the geometry of $\mathcal{X}_n$, as mapping close points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the manifold to faraway points

$\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ in $\mathbb{R}^m$ results in a large penalization, whereas there is no penalty for far away points in the manifold. Equation (4.1) can be rewritten as

$$E(\mathcal{Y}_n) = 2\,\mathrm{tr}\left(\mathsf{Y}\,\mathsf{L}\,\mathsf{Y}^T\right) \ , \tag{4.2}$$

where $\mathsf{Y} = [\boldsymbol{y}_1 \dots \boldsymbol{y}_n]$ and $\mathsf{L} = \mathsf{D} - \mathsf{W}$, with D a diagonal matrix with entries $\mathsf{D}_{ii} = \sum_j w_{ji}$. L is known as the graph Laplacian of G. After imposing appropriate constraints to remove arbitrary translations and scalings in the embedding, finding a lower dimensional embedding of $\mathcal{X}_n$ reduces to solving the following optimization problem:

$$\arg \min_{\substack{\mathsf{Y}\,\mathsf{D}\,\mathbf{1} = \mathbf{0} \\ \mathsf{Y}\,\mathsf{D}\,\mathsf{Y}^T = \mathsf{I}}} \mathrm{tr}\left(\mathsf{Y}\,\mathsf{L}\,\mathsf{Y}^T\right) \ , \tag{4.3}$$

where I is the $n \times n$ identity matrix and $\mathbf{1}$ is a column vector of ones.

As L is positive semidefinite, the solution to problem (4.3) is given by the m generalized eigenvectors associated with the m smallest positive generalized eigenvalues that solve

$$\mathsf{L}\,\boldsymbol{v} = \lambda\,\mathsf{D}\,\boldsymbol{v} \ . \tag{4.4}$$

This is equivalent to solving a regular eigenvalue problem for matrix $\tilde{\mathsf{L}} = \mathsf{D}^{-1/2}\,\mathsf{L}\,\mathsf{D}^{-1/2}$, the so-called normalized graph Laplacian. If $\mathsf{V} = [\boldsymbol{v}_1 \dots \boldsymbol{v}_m]$ is the collection of such eigenvectors, then the embedded points are given by $\boldsymbol{y}_i = (v_{i1}, \dots, v_{im})^T$, $1 \leq i \leq n$.

4.3 Constraining the Manifold Embedding

Assume now that each point of $\mathcal{X}_n \in \mathcal{M}$ (or a subset of them) is associated with a class label, i.e., $\boldsymbol{x}_i$ has label $c_i \in \{-1, 1\}$. For simplicity, we only consider the problem of two classes, although the extension of the method proposed here to a

multi-class scenario is straightforward.

We are interested in finding a lower dimensional embedding for $\mathcal{X}_n$ that, unlike common manifold learning algorithms, takes into account the class structure of the data. The goal is to obtain an embedding that tries to separate classes in order to improve training and generalization capabilities of a classifier fitted to the lower dimensional embedded data.

Although we also use graph Laplacians, we do not follow the approach advocated in [7]. In [7], dimensionality reduction and classification are treated as a function fitting problem, relying on the the eigenvectors of the Laplacian as a natural basis to represent functions on the graph sampling of the manifold.

The method developed here is based on the idea of *maximum alignment* [112] between classes and data points. This idea proceeds as follows. Start by associating with each class a new node on the adjacency graph, called *class center*, inserting an edge of unit weight between this node and all data points with the same class label. Now, if we view the graph edges as springs that pull together nodes in the graph, determining an embedding corresponds to finding data coordinates in an m-dimensional space that minimize the stresses in the system of springs. This will lead to points with the same class label trying to cluster together around the class center, while attempting to preserve the geometric neighborhood structure of the manifold. In this way, the class centers are maximally aligned with the data points.

We now formalize this idea. Let $\boldsymbol{z}_k \in \mathbb{R}^m$ be the class center associated with class k and C be the class membership matrix, i.e., $c_{ki} = 1$ if $\boldsymbol{x}_i$ has label k and $c_{ki} = 0$ otherwise. As before, we find the embedding by minimizing the cost function

$$E(\mathcal{Z}_n) = \sum_{ki} c_{ki} \left\| \boldsymbol{z}_k - \boldsymbol{y}_i \right\|^2 + \beta \sum_{ij} w_{ij} \left\| \boldsymbol{y}_i - \boldsymbol{y}_j \right\|^2 , \tag{4.5}$$

where $\mathcal{Z}_n = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ and $\beta \geq 0$ is a regularization parameter. Large values of β will produce embeddings that ignore class labels, while small values will produce embeddings that ignore the manifold structure of the data. Of course, in the latter case, points will tend to collapse into the class centers, producing lower dimensional data with little value to train a classifier with good generalization performance.

By defining $Z = [\boldsymbol{z}_1\, \boldsymbol{z}_2\, \boldsymbol{y}_1 \ldots \boldsymbol{y}_n]$, determining the lower dimensional embedding of $\mathcal{X}_n$ can be once again made equivalent to the following optimization problem:

$$\arg \min_{\substack{Z\,D\,\mathbf{1} = \mathbf{0} \\ Z\,D\,Z^T = I}} \operatorname{tr}\left(Z\,L\,Z^T\right) \ , \tag{4.6}$$

where L is the $(n+2) \times (n+2)$ graph Laplacian associated with weight matrix

$$W' = \begin{bmatrix} I & C \\ C^T & \beta W \end{bmatrix} .$$

The solution of problem (4.6) is again given by the matrix of the generalized eigenvectors associated with the m smallest positive generalized eigenvalues of L, where the first rows correspond to the coordinates of the class centers and the following rows determine the embedding of the original data points.

We remark that the method proposed here extends naturally to the semi-supervised setting, where only partial labeling is available. In this case, the points $\boldsymbol{x}_i$ for which there are no label information will have the corresponding columns of matrix C set to zero, thus imposing no additional constraints.

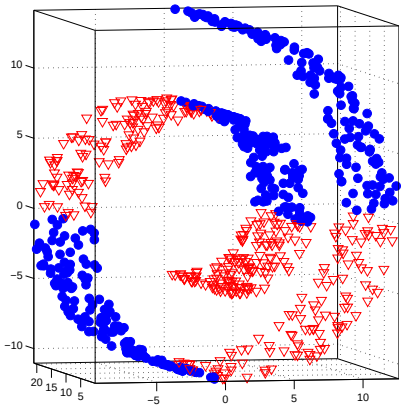


Figure 4.1: Swiss roll manifold with 400 points from each of 2 classes, marked as '▽' (red) and '•' (blue).

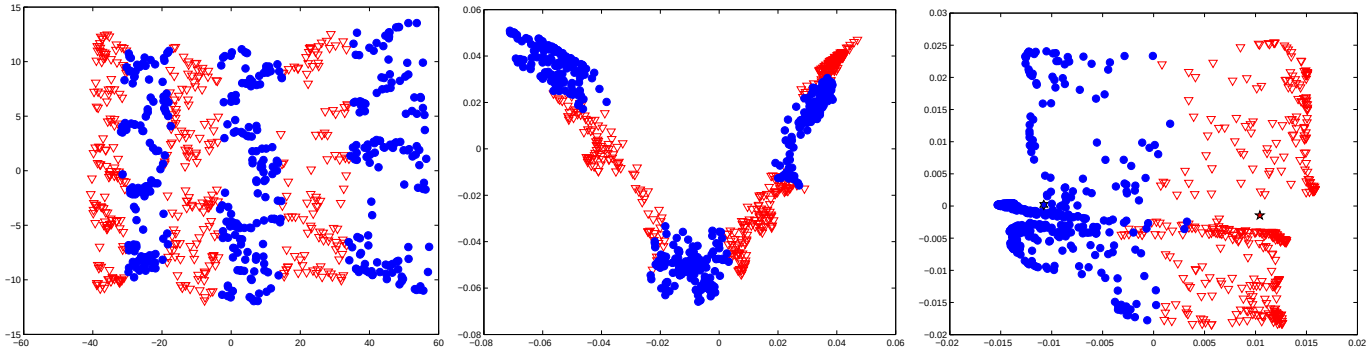
4.4 Examples

We now show through simple examples how the proposed classification constrained dimensionality reduction (CCDR) algorithm works. All the simulations presented here have $\beta = 1$, neighborhood parameter $k = 12$ and the scale parameter ϵ of the heat kernel is set automatically according to [56]:

$$\epsilon = \frac{10}{n} \sum_{i=1}^{n} \min_{j:\, \boldsymbol{x}_j \neq \boldsymbol{x}_i} \|\boldsymbol{x}_j - \boldsymbol{x}_i\|^2 \ .$$

Consider the standard 2-dimensional swiss roll manifold in $\mathbb{R}^3$. We sample 400 points uniformly on the manifold from each of two classes, as shown in figure 4.1. As it can be deduced, there is no linear projection of the data into a 2-dimensional subspace that separates the classes.

Figure 4.2 shows the results of applying standard manifold learning methods, ISOMAP [103] and Laplacian Eigenmaps, together with the proposed CCDR algorithm to the data set of Figure 4.1. Recall that both ISOMAP and Laplacian Eigenmaps do not account for label information when computing the embedding. As a result, although ISOMAP (Figure 4.2(a)) is able to recover an isometric embedding



(a) ISOMAP (b) Laplacian Eigenmpas (c) CCDR

Figure 4.2: Applying dimensionality reduction algorithms to the Swiss roll data set of Figure 4.1. ISOMAP was computed using 8-NN, while both Laplacian Eigenmaps and CCDR used 12-NN.

of the data into the plane, it fails at finding a simple separation of the classes. The Laplacian eigenmaps method (Figure 4.2(b)) gives similar results, albeit finding an arc-length type parameterization of the data. On the contrary, the CCDR algorithm (Figure 4.2(c)) computes an embedding where classes are almost linearly separable.

To quantify this behavior, we designed a very simple classifier. To classify a new sample, add it to the graph formed by the training set, with unknown label (add a zero column to matrix C), compute the constrained (or simple Laplacian) embedding. and then classify the sample using a simple NN-classifier on the embedded points. We compare this to a baseline NN-classifier on the full dimensional data set. In all the experiments a 3-NN classifier was used. We tested 50 sample points per training set and repeated for 20 random training sets. Table 4.1 shows the average error rates as a function of the number of training samples. As it can be seen, the CCDR algorithm outperforms the other methods. Supporting the claim that dimensionality reduction without guidance can harm classification performance, it can be observed that the full dimensional NN-classifier does better than a NN-classifier based on the Laplacian embedding.

Table 4.1: Error rates for classification using pre-processing dimensionality reduction versus full dimensional data

no. of train. samples	CCDR	Laplacian	3-NN
300	**4.4** %	6.4 %	5.0 %
400	**3.6** %	5.0 %	4.4 %
500	**2.6** %	3.6 %	3.4 %

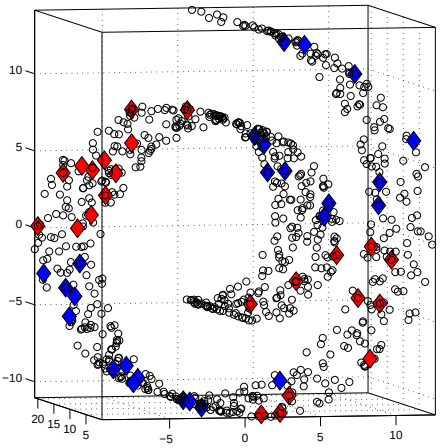


Figure 4.3: Swiss roll manifold with 50 samples labeled out of a total of 400 training samples. Labeled and unlabeled samples are marked as '◊' (red and blue) and '○' (black), respectively.

4.5 Semi-supervised Learning of Manifold Data

In many applications, although it might be easy to collect a large database of unlabeled training examples, the operation of labeling examples can be too expensive as it can depend on time consuming or costly experiments. These include object and speech recognition or text and genetic databases classification, among others. In these cases, one has to resort to using only a few labeled samples within a large database of unlabeled training points. See Figure 4.3. However, the geometric structure of the overall data set can be combined with this information to improve labeling of the remaining unlabeled examples.

The framework presented in this chapter, combining geometric structure preser-

vation and class discrimination, can be used to improve semi-supervised learning machines. Adopting the method proposed in [7], we have the following algorithm. Firstly, compute the constrained embedding of the entire data set, inserting a zero column in C for each unlabeled sample. Secondly, fit a classifier to the labeled embedded points. For example, fit a linear classifier by minimizing the quadratic error loss:

$$\ell(\boldsymbol{a}) = \sum_{\substack{i\,:\,\boldsymbol{x}_i \text{ is} \\ \text{labeled}}} \left(c_i - \boldsymbol{a}^T \boldsymbol{y}_i\right)^2 .$$

Thirdly and finally, for an unlabeled point $\boldsymbol{x}_j$, label it using the fitted (linear) classifier:

$$c_j = \begin{cases} 1 & \text{if } \boldsymbol{a}^T \boldsymbol{y}_j \geq 0 \\ -1 & \text{if } \boldsymbol{a}^T \boldsymbol{y}_j < 0 \end{cases} .$$

We compare this algorithm to the simple Laplacian equivalent, where the embedding is found using Laplacian eigenmaps, and to a baseline NN-classifier, where points are labeled according to the nearest labeled neighbors in the full dimensional space. We chose the best k-NN classifier for $k = 1, 2, 3$. Figure 4.4 shows the error rates as a function of the number of labeled points, for a total of 1000 points on the Swiss roll. For each fixed number of labeled points, we drew 20 independent data sets and randomly assigned labels, although guaranteeing balanced classes. As it can be seen, for a small number of labeled points, CCDR performs almost as bad as the Laplacian algorithm, as label information is not enough to produce an embedding substantially different from the original Laplacian eigenmaps, where classes are highly mixed (see Fig. 4.2(b)). However, as the number of samples increases beyond 100, the class constraints start to take effect, driving an embedding that can achieve almost linear separation of classes and thus outperforming all other methods.

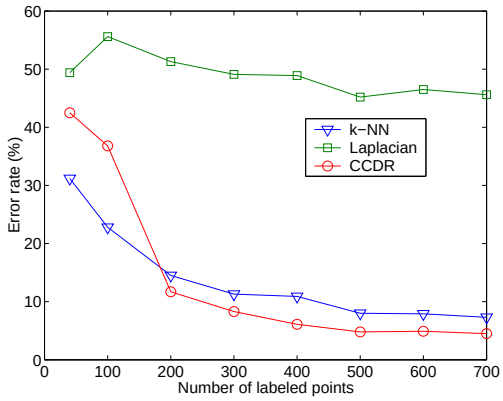


Figure 4.4: Percentage of errors for labeling unlabeled samples as a function of the number of labeled points, out of a total of 1000 points on the Swiss roll.

4.6 Conclusion

Several issues should be addressed in order to make the method proposed widely applicable to real life problems. Of prime importance is the development of out-of-sample extensions of the embeddings to new points. After these issues are solved, it would be important to see the effect of the proposed algorithm when used with state of the art classifiers, like support vector machines (SVM). This will contribute to understanding the tradeoff between feature extraction and classification and dimensionality reduction and dimensionality expansion via kernel machines. Also of interest, is the study of the influence of the regularization parameter β in the classification performance.

We are currently applying this method to high-dimensional databases, such as the MNIST database of handwritten digits.

CHAPTER 5

Distributed Weighted-Multidimensional Scaling for Node Localization in Sensor Networks

5.1 Introduction

In this chapter, we shift the focus from using random graphs to extract properties of the sample distribution to using random graphs to extract spatial information.

If data is collected by sensors distributed over different locations, the adjacency graphs introduced in Chapter 4 can be used to encode the spatial dependencies, as opposed to geometrical and statistical dependencies. Dense networks, like modern and future sensor networks or the Internet, collect massive amounts of measurements that fit the profile of high-dimensional data sets discussed in this thesis. In particular, the spatial distribution of the sensors introduces statistical dependencies in the collected data. As such, implementing dimensionality reduction methods on such data sets to extract relevant information can not only greatly reduce storage needs but also dramatically decrease communication costs within the network.

Part of the research presented in this chapter was joint work with Neal Patwari.

This chapter addresses the problem of localization in a sensor network using these tools, together with the multidimensional scaling principle to dimensionality reduction.

5.1.1 Localization in Sensor Networks

For monitoring and control applications using wireless sensor networks, automatic localization of every sensor in the network will be a key enabling technology. Sensor data must be registered to its physical location to permit deployment of energy-efficient routing schemes, source localization algorithms, and distributed compression techniques. Moreover, for applications such as inventory management and manufacturing logistics, localization and tracking of sensors are the primary purposes of the wireless network. For large-scale networks of inexpensive, energy-efficient devices, it is not feasible to include GPS capability on every device or to require a system administrator to manually enter all device coordinates. In this chapter, we consider the location estimation problem for which only a small fraction of sensors have *a priori* coordinate knowledge, and range measurements between many pairs of neighboring sensors permit the estimation of all sensor coordinates. While angle measurements have also been used for sensor localization, in this chapter, we limit the discussion to localization based on range measurements.

Two major difficulties hinder accurate sensor location estimation: first, accurate range measurements are expensive; and second, centralized estimation becomes impossible as the scale of the network increases. This chapter proposes a distributed localization algorithm, based on a weighted version of multidimensional scaling (MDS), which naturally incorporates local communication constraints within the sensor network. Its key features are:

1. A weighted cost function that allows range measurements that are believed to

be more accurate to be weighted more heavily.

2. An adaptive neighbor selection method that avoids the biasing effects of selecting neighbors based on noisy range measurements.

3. A majorization method which has the property that each iteration is guaranteed to improve the value of the cost function.

Simulation results and experimental channel measurements show that even when using only a small number of range measurements between neighbors and relying on fading-prone received signal-strength (RSS), the proposed algorithm can be nearly unbiased with performance close to the Cramér-Rao lower bound.

5.1.2 Sensor Localization Requirements

For a network of thousands or even millions of sensors, the large scale precludes centralized location estimation. Sending pair-wise range measurements from each sensor to a single point and then sending back estimated device coordinates would overwhelm the capacity of low-bandwidth sensor networks and waste energy. Decentralized algorithms are vital for limiting communication costs (which are usually much higher than computation costs) as well as for balancing the communication and computational load evenly across the sensors in the network. Furthermore, when a sensor moves, the ability to recalculate location locally rather than globally will result in energy savings which, over time, may dramatically extend the lifetime of the sensor network.

Sensor energy is also conserved by limiting transmission power. For a given channel between a pair of wireless sensors, the SNR of the received signal can be improved by increasing the transmit power. Range measurement accuracy improves at higher SNR [11, 51], thus imposing a tradeoff between energy cost and accuracy.

There is also a tradeoff between device cost and range accuracy: using ultrawideband (UWB) [17, 28] or hybrid ultrasound/RF techniques [32] can achieve accuracies on the order of centimeters, but at the expense of high device and energy costs. Alternatively, very inexpensive wireless devices can measure RF RSS just by listening to network packet traffic, but range estimates from RSS incur significant errors due to channel fading. All range measurements tend to degrade in accuracy with increasing distance. In particular, RSS-based range measurements experience errors whose variance is proportional to the actual range. Accurate localization algorithms must take into account the range dependence of the ranging variance.

Finally, measurement of ranges between every pair of devices would require $O(n^2)$ measurements for n sensors. The distributed weighted-multidimensional scaling (dwMDS) algorithm reduces measurement costs by requiring range measurements only between a small number of neighboring sensors.

5.1.3 Multidimensional Scaling

The goal of multidimensional scaling is to find a low dimensional representation of a group of objects (e.g., sensor positions), such that the distances between objects fit as well as possible a given set of measured pairwise "dissimilarities" that indicate how dissimilar objects are (e.g., inter-sensor RSS). MDS has found many applications in chemical modeling, economics, sociology, political science and, especially, mathematical psychology and behavioral sciences [21]. More recently, MDS has been used by the machine learning community for manifold learning [103]. In the sensor localization context, MDS can be applied to find a map of sensor positions (in 2-D or 3-D) when dissimilarities are measurements of range obtained, for example, via RSS or Time of Arrival (TOA).

For the last 70 years many approaches to solving the MDS problem have been

formulated (see [8, 21, 23, 35] and references therein). On the one hand, when the measured dissimilarities are equal to the true distances between sensors, classical MDS provides a closed formed solution by singular value decomposition of the centered squared dissimilarity matrix (see Section 5.3). On the other hand, when dissimilarities are measured in noise, other techniques should be used, usually based on iteratively minimizing a loss function between dissimilarities and distances. This framework encompasses techniques such as alternating nonlinear least squares (ALSCAL) [102], nonlinear least squares via majorizing functions (SMACOF) [36], nonmetric scaling [54,55] or maximum likelihood formulations [86,115]. Common to all these methods is the need for a central processing unit to gather all the available dissimilarities and perform the function minimization.

In contrast, we present a distributed MDS algorithm, which operates by minimizing multiple local loss functions. The local nonlinear least squares problem is solved using quadratic majorizing functions as in SMACOF. Since each local cost distributes additively over the network, each sensor contributes to the minimization of the global MDS loss function. In this way, our algorithm produces a sequence of position estimates with corresponding non-increasing global cost and limited communication between sensors.

5.1.4 Related Work

Many aspects of the sensor localization problem have been addressed in recent literature. Notably, bounds on estimation performance have been derived for the cases when pair-wise measurements are RSS, TOA, Angle Of Arrival (AOA), or a combination [14, 67, 68, 74, 79]. Furthermore, centralized algorithms based on multidimensional scaling [98], convex optimization [25], and maximum-likelihood [67] have demonstrated good estimation performance.

Research has also demonstrated the feasibility of distributed localization algorithms, which are required for scalability and balancing computational costs across large sensor networks. Distributed localization algorithms presented in the literature can be grouped into two types: adapted trilateration algorithms, and successive refinement algorithms. In the first type, devices estimate the distance to multiple known-location devices, using either a direct measurement, or if none exists, an estimate based on the shortest path to the known-location devices [69, 73, 96]. Then, using these range estimates to the anchors, the device uses trilateration to estimate its location. In the successive refinement approaches to localization, each device locally estimates its location from measured ranges to its neighbors. Each device begins with its own local coordinate system, and later merges it with neighboring coordinate systems [106]. The devices successively refine their location estimates [1, 95], effectively finding a solution to a global optimization problem that uses all ranges measured between neighbors.

The distributed algorithm presented here falls in the successive refinement category, which finds a minimum of a global cost function. In the dwMDS approach, however, the special cost function structure avoids the complicated step of merging local maps and a majorization algorithm is used to ensure that each iteration decreases the global cost function. This global improvement is guaranteed even though sensors operate individual updates based only on information received from their few closest neighbors.

Although using a different formulation then the one proposed here, the following papers also apply MDS-type techniques to sensor localization:

– *Plain MDS*: In [98], devices have connectivity information (whether or not two devices are in range). The distance between two connected nodes is defined to be 1, while the distance between two nodes not in range is set to the number

of hops in the shortest path between them (similar to Isomap [103]). The matrix of distances between each pair of devices is used by an MDS algorithm to estimate the coordinates of the devices. Compared to the present chapter, this centralized MDS method weights each distance equally. Unlike [98], the method proposed here avoids the (usually inaccurate) estimation of distances between out-of-range sensors.

– *Local MDS*: In [45], a local version of MDS is used to compute maps of many local arrangements of nodes. These local maps are pieced together to obtain global maps. This method tends to perform better than the global MDS method when node density is non-uniform, or "holes" in coverage exist. The local calculations allow a distributed implementation, but weights are restricted to be either 0 or 1. The formulation introduced in the present chapter removes that restriction, by allowing arbitrary non-negative weights, and naturally bypasses the complex step of fusing the local maps into a global map.

– *Manifold Learning*: Centralized manifold learning techniques are used in [78] to estimate sensor locations, without explicit range estimation, in cases where sensor data has correlation structure that is monotone in inter-sensor distance. Classical MDS is used to estimate physical location coordinates from the high-dimensional sensor data. The present chapter uses direct measurements of range between pairs of neighboring devices to estimate locations.

Common to most sensor localization methods is the process of selecting sensor neighborhoods for range measurements. Most methods propose using only ranges measured between nearby neighbors, in order to limit communication costs and computational complexity. However, when ranges are measured with noise, the act of choosing neighbors based on these measurements will tend to select devices whose

measured distances are shorter than the true distances. This chapter addresses this biasing effect and proposes a two-stage neighbor selection process that can be used to unbias location estimates even in high-noise environments. We remark that, to our knowledge, this problem has not been previously considered in the sensor localization literature.

5.1.5 Outline

The remainder of this chapter is organized as follows. Section 5.2 provides a formal statement of the sensor localization problem considered here. In Section 5.3 we describe the solution to the classical MDS formulation and discuss its shortcomings in a distributed and sensor network environment. The proposed algorithm is introduced in Section 5.4. In Section 5.5 we discuss statistical models for TOA and RSS measurements to show why a weighted MDS solution is important. Section 5.6 discusses the bias effect associated with using these noisy range measurements to select neighboring devices and proposes a solution. In Section 5.7 we show results on both simulated measurement data and on measured range data recorded for a 44-node sensor network in an indoor office environment. Finally, Section 5.8 concludes the chapter with a discussion about the proposed method, improvements and future work.

5.2 Problem Statement

To be specific about sensor localization, we now formally state the estimation problem addressed in this chapter.

Consider a network of $N = n+m$ devices, living in a D-dimensional space ($D = 2$ or 3, although the proposed formulation can handle arbitrary D-dimensional local-

ization, as long as $D < N$). Let $\{\boldsymbol{x}_i\}_{i=1}^N$, $\boldsymbol{x}_i \in \mathbb{R}^D$, be the actual vector coordinates of sensors, or, equivalently, define the matrix of coordinates $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N]$. The last m sensors ($i = n+1, \ldots, N$) have perfect *a priori* knowledge of their coordinates and are called *anchor nodes*. The first n sensors ($i = 1, \ldots, n$) have either no knowledge or some imperfect *a priori* coordinate knowledge and are called *unknown-location nodes*. Imperfect *a priori* knowledge about sensor $i \leq n$ is encoded by parameters r_i and $\overline{\boldsymbol{x}}_i$, where, with accuracy r_i, $\boldsymbol{x}_i$ is believed to lie around $\overline{\boldsymbol{x}}_i$ (see Section 5.4 for a precise definition of these parameters). If no such knowledge is available, $r_i = 0$. Summarizing, three distinct sets of sensors can be considered in this formulation based on their *a priori* information: perfect ($i > n$), imperfect ($i \leq n, r_i > 0$), or zero coordinate knowledge ($i \leq n, r_i = 0$). Note that one or two of these sets might be empty, e.g., no anchor nodes available and/or no prior information on sensors locations. These and other notation used throughout this chapter is gathered in Table 5.1.

The localization problem we consider is the estimation of the coordinates $\{\boldsymbol{x}_i\}_{i=1}^n$ given the coordinates of the anchor nodes, $\{\boldsymbol{x}_i\}_{i=n+1}^N$, imperfect *a priori* knowledge, $\{(r_i, \overline{\boldsymbol{x}}_i)\}_{i=1}^n$ and many pairwise range measurements, $\{\delta_{ij}^{(t)}\}$, taken over time $t = 1 \ldots K$. We use the terms 'dissimilarity' and 'range measurement' interchangeably, in order to seamlessly merge terms common to MDS and localization literature. The available range measurements (i, j) are some subset of $\{1 \ldots N\}^2$. We assume that this subset of range measurements results in a connected network; otherwise, each connected subset should be considered individually.

The method developed is general enough to adapt to any range measurement method, such as TOA, RSS, or proximity. We focus in particular on RSS-based range measurements, due to its desirability as a low-device cost method, but we also test the method using TOA range measurements in Section 5.7.

Notation	Description
D	Dimensions of location estimates ($D = 2$ unless noted)
$N = n + m$	Total number of sensors
n	Sensors with imperfect or no *a priori* coordinate information
m	Sensors with perfect *a priori* coordinate knowledge ('anchor' nodes)
P_{ij}	Power received (dB) at sensor i transmitted by sensor j
P_{thr}	Minimum received power for successful reception
d_{thr}	Distance at which mean received power $= P_{thr}$
d_R	Threshold distance for neighborhood selection
$\boldsymbol{x}_i$	Actual coordinate vector of sensor i, $i = 1 \ldots n + m$
X	Actual coordinate matrix, $[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n+m}]$
$d_{ij}, d_{ij}(\mathsf{X})$	Actual distance between sensors i and j in matrix X
$\delta_{ij}^{(t)}$	Range measured at time t between sensors i and j
$w_{ij}^{(t)}$	Weight given to the range measured at time t between sensors i and j
$\overline{\delta}_{ij}$	Weighted average measured range between sensors i and j
$\overline{w}_{ij}$	Weight given to the average measured range between sensors i and j
S	Global objective function to be minimized
S_i	Local objective function to be minimized at sensor $i = 1 \ldots n$
$\boldsymbol{x}_i^{(k)}$	Estimated coordinates of sensor i at iteration k
$\mathsf{X}^{(k)}$	Estimated coordinate matrix at iteration k

Table 5.1: Symbols used in text and derivations

5.3 Classical Metric Scaling

If we assume that we measure all the pairwise dissimilarities $\{\delta_{ij}\}_{i,j=1}^{N}$ between points, and that these correspond to the true Euclidean distances, then

$$\delta_{ij} = d_{ij} = d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\| = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{x}_i - \boldsymbol{x}_j)} \; . \tag{5.1}$$

By writing the squared distances as $d_{ij}^2 = \boldsymbol{x}_i^T\boldsymbol{x}_i - 2\boldsymbol{x}_i^T\boldsymbol{x}_j + \boldsymbol{x}_j^T\boldsymbol{x}_j$, one can recover the matrix of inner products between points in the following way. Defining $\boldsymbol{\psi} = [\boldsymbol{x}_1^T\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N^T\boldsymbol{x}_N]^T$, the squared distance matrix, $\mathsf{D} = [d_{ij}^2]_{i,j=1}^{N}$, can now be written as

$$\mathsf{D} = \boldsymbol{\psi}\boldsymbol{e}^T - 2\mathsf{X}^T\mathsf{X} + \boldsymbol{e}\boldsymbol{\psi}^T \; ,$$

where $\boldsymbol{e}$ is the N-dimensional vector of all ones. Defining H to be the centering operator, $\mathsf{I} - \boldsymbol{e}\boldsymbol{e}^T/N$, it follows that

$$\mathsf{B} = -\mathsf{H}\,\mathsf{D}\,\mathsf{H} = \mathsf{H}\,\mathsf{X}^T\mathsf{X}\,\mathsf{H} \; .$$

After multiplication with H, the columns of $\mathsf{X}^T\mathsf{X}$ have zero mean. Now, given B, one can recover matrix X, up to a translation and orthogonal transformation, as the solution to the following variational problem:

$$\min_{\mathsf{Y}} \|\mathsf{B} - \mathsf{Y}^T\mathsf{Y}\|_F^2 \; , \tag{5.2}$$

where $\|.\|_F$ is the Frobenius norm and the minimum is taken over all $D \times N$ rank-D matrices. The solution of (5.2) is given by

$$\mathsf{X} = \mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_D^{1/2})\,\mathsf{U}^T \; , \tag{5.3}$$

where

$$B = U \operatorname{diag}(\lambda_1, \ldots, \lambda_D)\, U^T$$

is the (rank-D) singular vale decomposition (SVD) of matrix B.

The above derivation exposes the shortcomings of classical MDS. First, obtaining matrix B requires the knowledge of all the pairwise dissimilarities, a scenario highly unlikely in a dense sensor network due to power and/or bandwidth constraints. Second, due to a lack of any special sparse structure, computing matrix B and its SVD requires that all the dissimilarities be communicated and processed by a central processing unit, a communication-intensive operation in most sensor networks. Finally, (5.3) assumes that the true distances between points are available. For the realistic case in which range measurements are corrupted by multiplicative type noise (see Section 5.5), classical metric scaling minimizes the squared error between the squared distances d_{ij}^2 and δ_{ij}^2 (rather than the distances themselves) which tends to amplify the measurement errors, resulting in poor noise performance.

5.4 Distributed Weighted Multidimensional Scaling

We propose a distributed weighted MDS algorithm (dwMDS) that fits the sensor networks framework of distributed computations and restricted communications and also accounts for measurement errors. In particular, each sensor will only need to communicate relevant information with its neighbors (one hop away) in order to achieve a localization solution, as opposed to a multi-hop communication protocol or higher power transmissions necessary to reach a fusion center in a centralized algorithm.

5.4.1 The dwMDS Cost Function

Motivated by the variational formulation of classical metric scaling (cf. (5.2)), we seek to estimate sensor positions by minimizing the following global cost function (a.k.a. STRESS function [21]):

$$S = 2 \sum_{1 \le i \le n} \sum_{i < j \le n+m} \sum_{1 \le t \le K} w_{ij}^{(t)} \left(\delta_{ij}^{(t)} - d_{ij}(\mathsf{X}) \right)^2 + \sum_{1 \le i \le n} r_i \left\| \boldsymbol{x}_i - \overline{\boldsymbol{x}}_i \right\|^2 . \qquad (5.4)$$

where the actual Euclidean distance $d_{ij}(\mathsf{X})$ is given by (5.1), and we assume that for each pair (i, j), up to K dissimilarity measurements are available. The arbitrary weight $w_{ij}^{(t)}$ $(t = 1, \ldots, K)$ can be selected to quantify the predicted accuracy of measurement $\delta_{ij}^{(t)}$. If no such measurement is available between i and j, or its accuracy is zero, then $w_{ij}^{(t)} = 0$. We assume that $w_{ij}^{(t)} \geq 0$, $w_{ii}^{(t)} = 0$ and $w_{ij}^{(t)} = w_{ji}^{(t)}$, i.e., the weights are symmetric. Vector $\overline{\boldsymbol{x}}_i$ contains prior information about the location of node i $(1 \leq i \leq n)$, while r_i determines the influence of such information on the overall cost, by quantifying how accurate this prior location is. If there is no prior information, then $r_i = 0$. Note, function (5.4) differs from the standard MDS objective function in that we have added a penalty term to account for prior knowledge about node locations.

We stress that the variational formulation of (5.4) implies a nonparametric view of the location problem – the sensor positions are estimated by minimizing S (w.r.t. $\{\boldsymbol{x}_i\}$), where no model assumptions are made about the statistical behavior of the observed dissimilarities. This permits the use of data-dependent weighting schemes (see Section 5.4.3), resulting in a cost function that can automatically adapt to different measurement models. Nevertheless, we remark that equation (5.4) can also be seen from a statistical viewpoint. Under a Bayesian perspective, (5.4) can be interpreted as the log posterior density of the nodes locations given the observed dissimilarities,

$\log f(\{\boldsymbol{x}_i\}|\{\delta_{ij}\})$, if we assume that the dissimilarities $\{\delta_{ij}^{(t)}\}$ are i.i.d. Gaussian with mean d_{ij} and variance $(2\,w_{ij}^{(t)})^{-1}$ and points $\{\boldsymbol{x}_i\}$ have a Gaussian prior with mean $\overline{\boldsymbol{x}}_i$ and variance $(2\,r_i)^{-1}$.

After simple manipulations, S can be rewritten as follows:

$$S = \sum_{i=1}^{n} S_i + c \;, \tag{5.5}$$

where local cost functions S_i are defined for each unknown-location node (ie. $1 \leq i \leq n$),

$$S_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} \overline{w}_{ij} \left(\overline{\delta}_{ij} - d_{ij}(\mathsf{X})\right)^2 + \sum_{j=n+1}^{n+m} 2\,\overline{w}_{ij} \left(\overline{\delta}_{ij} - d_{ij}(\mathsf{X})\right)^2 + r_i \|\boldsymbol{x}_i - \overline{\boldsymbol{x}}_i\|^2 \;, \tag{5.6}$$

and c is a constant independent of the nodes locations X. In (5.6), the K weights and range measurements between i and j are summarized by a single weight $\overline{w}_{ij} = \sum_{t=1}^{K} w_{ij}^{(t)}$ and measurement $\overline{\delta}_{ij} = \sum_{t=1}^{K} w_{ij}^{(t)}\,\delta_{ij}^{(t)} / \overline{w}_{ij}$. As S_i only depends on the measurements available at node i and the positions of neighboring nodes, i.e., nodes for which $w_{ij}^{(t)} > 0$ (for some t), it can be viewed as the local cost function at node i. We note that if $m = 0$ (i.e., no anchor nodes are available) and $r_i = 0$, for all i (i.e., no prior information on the nodes locations), then $\partial S / \partial \boldsymbol{x}_i = 2\,\partial S_i / \partial \boldsymbol{x}_i$. This implies that the influence of $\boldsymbol{x}_i$ on the local cost S_i determines its influence on the global cost S. Motivated by this cost structure, we propose an iterative scheme in which each sensor updates its position estimate by minimizing the corresponding local cost function S_i, after observing dissimilarities and receiving position estimates from its neighboring nodes.

5.4.2 Minimizing the dwMDS Cost Function

Unlike classical MDS, no closed form expression exists for the minimum of the cost function S or S_i. By assuming that each node has received position estimates from neighboring nodes, we minimize $S_i = S_i(\boldsymbol{x}_i)$ iteratively using quadratic majorizing functions as in SMACOF (Scaling by MAjorizing a COmplicated Function [36]). This method has the attractive property of generating a sequence of non-increasing STRESS values.

A majorizing function $T_i(\boldsymbol{x}, \boldsymbol{y})$ of $S_i(\boldsymbol{x})$ is a function $T_i : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ that satisfies: (i) $S_i(\boldsymbol{x}) \leq T_i(\boldsymbol{x}, \boldsymbol{y})$ for all $\boldsymbol{y}$, and (ii) $S_i(\boldsymbol{x}) = T_i(\boldsymbol{x}, \boldsymbol{x})$. This function can then be used to implement an iterative minimization scheme. Starting at an initial condition $\boldsymbol{x}_0$, the function $T_i(\boldsymbol{x}, \boldsymbol{x}_0)$ is minimized as a function of $\boldsymbol{x}$. The newly found minimum, $\boldsymbol{x}_1$, can then be used to define a new majorizing function $T_i(\boldsymbol{x}, \boldsymbol{x}_1)$. This process is then repeated until convergence (see [36] for details). The trick is to use a simple majorizing function that can be minimized analytically, e.g., a quadratic function. Following [36], we start by rewriting S_i as:

$$S_i(\boldsymbol{x}_i) = \eta_\delta^2 + \eta^2(X) - 2\,\rho(X) \ ,$$

where

$$\eta_\delta^2 = \sum_{\substack{j=1 \\ j\neq i}}^{n} \overline{w}_{ij}\,\overline{\delta}_{ij}^2 + \sum_{j=n+1}^{n+m} 2\,\overline{w}_{ij}\,\overline{\delta}_{ij}^2 \ , \tag{5.7}$$

$$\eta^2(X) = \sum_{\substack{j=1 \\ j\neq i}}^{n} \overline{w}_{ij}\,d_{ij}^2(X) + \sum_{j=n+1}^{n+m} 2\,\overline{w}_{ij}\,d_{ij}^2(X) + r_i\|\boldsymbol{x}_i - \overline{\boldsymbol{x}}_i\|^2 \ , \tag{5.8}$$

$$\rho(X) = \sum_{\substack{j=1 \\ j\neq i}}^{n} \overline{w}_{ij}\,\overline{\delta}_{ij}\,d_{ij}(X) + \sum_{j=n+1}^{n+m} 2\,\overline{w}_{ij}\,\overline{\delta}_{ij}\,d_{ij}(X) \ . \tag{5.9}$$

Term (5.7) does not depend on $\boldsymbol{x}_i$ and term (5.8) is quadratic in $\boldsymbol{x}_i$. Only term (5.9) depends on $\boldsymbol{x}_i$ through a more complicated (sum of square roots) function. Define $T_i(\boldsymbol{x}, \boldsymbol{y})$ as:

$$T_i(\boldsymbol{x}_i, \boldsymbol{y}_i) = \eta_\delta^2 + \eta^2(X) - 2\,\rho(X, Y)\ , \tag{5.10}$$

where

$$\rho(X, Y) = \sum_{\substack{j=1 \\ j\neq i}}^{n} \overline{w}_{ij}\, \frac{\overline{\delta}_{ij}}{d_{ij}(Y)} (\boldsymbol{x}_i - \boldsymbol{x}_j)^T (\boldsymbol{y}_i - \boldsymbol{y}_j) + \sum_{j=n+1}^{n+m} 2\,\overline{w}_{ij}\, \frac{\overline{\delta}_{ij}}{d_{ij}(Y)} (\boldsymbol{x}_i - \boldsymbol{x}_j)^T (\boldsymbol{y}_i - \boldsymbol{y}_j)\ . \tag{5.11}$$

Using the fact that, by Cauchy-Schwarz inequality,

$$d_{ij}(X) = \frac{d_{ij}(X)\, d_{ij}(Y)}{d_{ij}(Y)} \geq \frac{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T (\boldsymbol{y}_i - \boldsymbol{y}_j)}{d_{ij}(Y)}\ ,$$

it is easily seen that T_i majorizes S_i. Minimizing S_i through a majorizing algorithm is now a simple task of finding the minimum of T_i:

$$\frac{\partial T_i(\boldsymbol{x}_i, \boldsymbol{y}_i)}{\partial \boldsymbol{x}_i} = 0\ . \tag{5.12}$$

An expression for this gradient is given in Appendix 5.9. If $X^{(k)}$ is the matrix whose columns contain the position estimates for all points at iteration k, one can derive an update for the position estimate of node i using equation (5.12):

$$\boldsymbol{x}_i^{(k+1)} = a_i \left(r_i\, \overline{\boldsymbol{x}}_i + X^{(k)} \boldsymbol{b}_i^{(k)} \right)\ , \tag{5.13}$$

where

$$a_i^{-1} = \sum_{\substack{j=1 \\ j\neq i}}^{n} \overline{w}_{ij} + \sum_{j=n+1}^{n+m} 2\,\overline{w}_{ij} + r_i\ , \tag{5.14}$$

and $\boldsymbol{b}_i^{(k)} = [b_1, \ldots, b_{n+m}]^T$ is a vector whose entries are given by:

$$\begin{aligned} b_j &= \overline{w}_{ij} \left[1 - \overline{\delta}_{ij}/d_{ij}(\boldsymbol{X}^{(k)})\right] && j \leq n \ , \ j \neq i \\ b_i &= \sum_{\substack{j=1 \\ j \neq i}}^{n} \overline{w}_{ij}\, \overline{\delta}_{ij}/d_{ij}(\boldsymbol{X}^{(k)}) + \sum_{j=n+1}^{n+m} 2\, \overline{w}_{ij}\, \overline{\delta}_{ij}/d_{ij}(\boldsymbol{X}^{(k)}) && . \\ b_j &= 2\, \overline{w}_{ij} \left[1 - \overline{\delta}_{ij}/d_{ij}(\boldsymbol{X}^{(k)})\right] && j > n \end{aligned} \quad (5.15)$$

As the weights $w_{ij}^{(t)}$ are zero for nodes j not in the relative neighborhood of node i, only the corresponding entries of vector $\boldsymbol{b}$ will be nonzero, and the update rule for $\boldsymbol{x}_i$ will depend only on this neighborhood (as opposed to the whole matrix $\boldsymbol{X}^{(k)}$). This fact provides the structure necessary for a distributed implementation: each node updates its location estimate according to equation (5.13); it then transmits this update to its set of neighbors; after receiving the same information from its neighbors, the node reiterates its location estimate.

We remark that, unlike the centralized SMACOF algorithm described in [36], the computation of (5.13) does not require the evaluation of a $n \times n$ Moore-Penrose matrix inverse.

We also point out that the minimization algorithm proposed can be seen as a special case of optimization transfer methods through surrogate objective functions [57], which also include the popular EM algorithm.

5.4.3 Algorithm

The proposed algorithm is summarized in Figure 5.1. We make the following comments:

1. The choice of weighting function w_{ij} should reflect the accuracy of measured dissimilarities, such that less accurate measurements are down-weighted in the overall cost function. If a noise measurement model is available, w_{ij} can be tai-

Inputs: $\{\delta_{ij}^{(t)}\}$, $\{w_{ij}^{(t)}\}$, m, $\{r_i\}$, $\{\overline{\boldsymbol{x}}_i\}$, ϵ, initial condition $\boldsymbol{X}^{(0)}$
Initialize: $k = 0$, $S^{(0)}$, compute a_i from equation (5.14)
repeat
  $k \leftarrow k + 1$
  for $i = 1$ **to** n
    compute $\boldsymbol{b}_i^{(k-1)}$ from equation (5.15)
    $\boldsymbol{x}_i^{(k)} = a_i \left(r_i \, \overline{\boldsymbol{x}}_i + \boldsymbol{X}^{(k-1)} \boldsymbol{b}_i^{(k-1)} \right)$
    compute $S_i^{(k)}$
    $S^{(k)} \leftarrow S^{(k)} - S_i^{(k-1)} + S_i^{(k)}$
    communicate $\boldsymbol{x}_i^{(k)}$ to neighbors of node i (i.e., nodes for which $w_{ij} > 0$)
    communicate $S^{(k)}$ to node $i + 1 \,(\mathrm{mod}\, n)$
  end for
until $S^{(k-1)} - S^{(k)} < \epsilon$

Figure 5.1: Algorithm for decentralized weighted-multidimensional scaling

lored to the variance predictions of the model. For example, one might select $w_{ij} = 1/(c_1 \delta_{ij} + c_2)^2$ if the measurements are Gaussian distributed with standard deviation increasing linearly with the true distances, i.e., $\sigma = c_1 d_{ij} + c_2$. When a reliable model is not available, one can adopt a model-independent adaptive weighting scheme. This is the approach adopted in this chapter. Inspired by the weighting frequently used in locally weighted regression methods (LOESS) [16], we propose the following weight assignment:

$$w_{ij} = \begin{cases} \exp\left\{-\delta_{ij}^2 / h_{ij}^2\right\} & \text{, if } \delta_{ij} \text{ is measured} \\ 0 & \text{, otherwise} \end{cases}, \tag{5.16}$$

where $h_{ij} = \max\left\{\max_k \delta_{ik}, \max_k \delta_{kj}\right\}$. This choice of w_{ij}, which equalizes the (nonzero) weight distribution in all sensors, has robust performance as shown in the experiments reported in Section 5.7. Other weighting schemes are also possible, ranging from alternative monotone functions to a naive choice of unit weights for measured distances.

2. The question of how to adaptively choose the neighbors of each node (i.e., which weights are made positive) in order to decrease communication costs or improve localization performance is addressed in Section 5.6.

3. The values of r_i should be chosen according to prior information. For example: if prior information about node i was obtained using GPS, then r_i should reflect the accuracy of the GPS sensor used; or, if anchor nodes are subject to small displacements, like vibrations, r_i should quantify the mean squared error between the node's average position and the its expected movements. Under a Bayesian perspective, the choice of r_i is analogous to the problem of subjective prior choice in a Bayesian model and thus can be guided by similar principles. Also, as in Bayesian problems, as more measurements per sensor are collected, the influence of the values of r_i are discounted in the final solution. Section 5.7 includes a discussion on the influence of r_i in experimental results.

4. Regarding the initialization of the algorithm, every node requires an initial estimate of its position. This can be done using the algorithms proposed in [95] or [106]: each node builds its local coordinate system, which is then passed along the network until a rough global map of the network is built. In the experiments reported in Section 5.7, we use a naive random initialization and found that the algorithm was robust with respect to these "rough" initial position estimates.

5. In the description of the algorithm, it was assumed for notational convenience, that the algorithm cycles through the network in an ordered fashion (i.e., messages are passed between nodes in the order $1, 2, \ldots, n$). However, many other non-cyclic update rules are possible. In particular, one possibility is for (spatial) clusters of sensors to iterate among themselves until their position estimates

stabilize. These estimates can then be transmitted to the neighboring clusters, before starting a new iteration step.

6. Although the majorization approach used guarantees a non-increasing sequence of STRESS vales, it may converge to a local minimum of this cost function, instead of the global one, like any gradient search method. This behavior can be alleviated to some extent by using some of the advanced search techniques proposed in [36].

Computational Complexity and Energy Consumption

Regarding computational complexity, it is easily seen that the algorithm in Figure 5.1 scales as $O(n\,L)$, where L is the total number of iterations required until the stopping rule is satisfied. This compares favorably to classical MDS, which requires $O(n^2\,T)$ operations, where T is the number of steps required by the Lanczos method to compute the necessary SVD.

However, in sensor network applications, far more important than computational complexity, is the amount of communication required by the algorithm, as the energy consumed by a single wireless transmission can far outweigh the energy necessary for local computations. As we are interested in how communication complexity scales with the size of the network, we adopt the model proposed in [85]. In this model, the average total energy used by a general data processing algorithm, as function of the number of nodes n, is given by

$$\mathcal{E}(n) = b(n) \times h(n) \times e(n) \ ,$$

where $b(n)$ is the average number of bits/packets transmitted, $h(n)$ is the average number of hops over which communication occurs, and $e(n)$ is the average amount

of energy required to transmit one bit/packet over one hop.

For simplicity, we assume, in the following analysis, that the sensors are uniformly distributed over a square or cube of unit side length, for, respectively, a $D = 2$ or $D = 3$ dimensional network. The proposed algorithm requires that each node transmits its position estimate to other nodes from which it obtained range measurements. If we assume that a node is able to sense all other nodes within a threshold distance d_{thr}, then the average number of neighbors a node can communicate with is upper bounded by $c_1(n-1)\,d_{thr}^D$, where c_1 is the volume of the D-dimensional unit sphere (nodes close to the border of the unit square or cube actually have fewer expected neighbors). As this operation occurs for each iteration for every node, an upper bound on the average number of transmitted bits is $b_{\text{dwMDS}}(n) \leq O(n^2\,L\,d_{thr}^D)$. Each communication to its neighbors can be made in one hop, so $h_{\text{dwMDS}}(n) = 1$. Thus, the average energy required for communication by the proposed algorithm is:

$$\mathcal{E}_{\text{dwMDS}}(n) \leq O\left(n^2\,L\,d_{thr}^D\,e_{\text{dwMDS}}(n)\right) \ . \tag{5.17}$$

Notice that $e_{\text{dwMDS}}(n)$ depends on d_{thr} (in a nonlinear way).

We remark that the same bound (5.17) on energy consumption is also valid for a sensor network with nodes distributed over a uniform grid of side length $O(n^{-1/D})$ (see Fig. 5.3). This scenario makes it easier to compare the proposed method to a centralized algorithm, assuming a multi-hop communication protocol. To simplify the analysis, we consider the threshold distance $d_{thr} = O(n^{-1/D})$. In this case, each node communicates only with its immediate neighbors in the uniform grid, making the average hop distance the same in the centralized and distributed case. This implies that the same energy is required to transmit a bit/packet over one hop, i.e., $e_{\text{dwMDS}}(n) = e_{\text{centr}}(n)$. For $d_{thr} = O(n^{-1/D})$, each node will, on the average, receive

range measurements from a fixed number of neighborhood nodes, no matter how big the network is. The centralized algorithm must transmit them to a fusion center. After a simple calculation, it can be shown that this results in $b_{\text{cebtr}}(n) = O(n)$ bits transmitted. For the uniform grid geometry, a simple calculation shows that the average number of hops from a node to the fusion center is $h_{\text{centr}}(n) = O(n^{1/D})$. Finally, we obtain the average energy required by a centralized algorithm,

$$\mathcal{E}_{\text{centr}}(n) = O\left(n^{1+1/D} e_{\text{centr}}(n)\right) \ . \tag{5.18}$$

Substituting for the assumed d_{thr} in expression (5.17), we obtain the ratio between energies required by a centralized versus a distributed algorithm, in the uniform grid case:

$$\frac{\mathcal{E}_{\text{centr}}(n)}{\mathcal{E}_{\text{dwMDS}}(n)} = O\left(\frac{n^{1/D}}{L}\right) \ . \tag{5.19}$$

For dense networks of the same size n, and fixing *a priori* the maximum number of iterations allowed, a centralized algorithm will require an order of $n^{1/D}$ more energy than the proposed distributed algorithm. Note that the costs of the centralized algorithm are not evenly distributed - nodes near the fusion center will disproportionately bear the forwarding costs.

To conclude this section, we remark that, for $D = 2$ and $d_{thr} = O(n^{-1/D})$, the proposed algorithm has a transport requirement of $O(n^2 \, L \, d_{thr}^D) \times d_{thr} = O(\sqrt{n})$ bit-meters/sec, which is the same as the transport capacity of a wireless network on a unit area region [37]. This suggests that the implementation of the proposed algorithm is pratically feasible, even with more resource aggressive update rules (e.g., parallel updates of all nodes), for a large sensor network.

5.5 Range Measurement Models

For concreteness, we assume throughout the rest of this chapter that range measurements between sensors are obtained either via RSS or TOA or a combination of the two. Both RSS and TOA can be measured via RF or by acoustic media; both media are subject to multipath and shadow fading phenomena which impair range estimates.

5.5.1 Time-of-Arrival

For a TOA receiver, the objective is to identify the time-of-arrival (TOA) of the direct line-of-sight (DLOS) path.[1] The power in the DLOS path is attenuated by any obstacles in between the transmitter and the receiver, and often, later-arriving non-line-of-sight(NLOS) multipath components arrive at the receiver with equal or greater power than the DLOS. As the distance between two devices increases, late-arriving paths contribute an increasing proportion of the overall received power. This increase has been observed in measured power-delay profiles - for example, excess delay and RMS delay spread tend to increase with path length [20, 38, 87]. Specifically motivated by radiolocation applications, researchers have used ns-synchronized measurement equipment to accurately identify the DLOS signal and show that NLOS signals' proportion of the total received power increases with path length [75]. This NLOS signal power serves as self-interference, in combination with other noise and interference, which effectively decreases the SNR of the TOA measurements as the range increases. Previous research has suggested using weighted least-squares algorithms to improve localization performance [15].

[1] This is a different goal than for a communications receiver, which aims to synchronize to the time which maximizes the SNR, regardless of whether the signal power comes from the DLOS path or later arriving paths.

5.5.2 Received Signal Strength

Similarly, range measurements based on RSS degrade with distance. Objects in the environment between the transmitter and receiver have the effect of multiplying the signal energy by attenuation factors. The cumulative effect of many such multiplications, by a central limit argument, results in a log-normal distribution of RSS (or equivalently received power) at the receiver [19]. If $P_{ij}(\text{mW})$, the received power in mW at sensor i transmitted by sensor j, is log-normal, then received power in decibels, $P_{ij} = 10 \log_{10} P_{ij}(\text{mW})$, is Gaussian. Furthermore, RF channel measurements have shown that the variance of P_{ij} is largely constant over path length [87] [79]. Thus P_{ij} is typically modeled as

$$\begin{aligned} P_{ij} &\sim \mathcal{N}(\bar{P}_{ij}, \sigma^2_{dB}) \\ \bar{P}_{ij} &= P_0 - 10 n_p \log_{10}(d_{ij}/d_0) \end{aligned} \tag{5.20}$$

where $\bar{P}_{ij}$ is the mean power in decibel milliwatts at distance d_{ij}, σ^2_{dB} is the variance of the shadowing, and $P_0(\text{dBm})$ is the received power at a reference distance d_0. Typically $d_0 = 1$ meter, and P_0 is calculated from the free space path loss formula [87]. The path loss exponent n_p is a parameter determined by the environment.

From this model for received power as a function of distance d_{ij}, the maximum likelihood estimator of distance is:

$$\delta_{ij} = d_0 10^{(P_0 - P_{ij})/(10 n_p)}. \tag{5.21}$$

If $P_{ij} = \bar{P}_{ij}$, then $\delta_{ij} = d_{ij}$. When $P_{ij} \neq \bar{P}_{ij}$, we can see why distance errors increase proportionally with distance. Consider a constant dB error in the received power measurement: $\Delta = \bar{P}_{ij} - P_{ij}$. For this error, $\delta_{ij} = d_{ij} 10^{\Delta/(10 n_p)}$, thus the

actual distance is multiplied by a constant factor. In fact, the range estimation error, $\delta_{ij} - d_{ij}$, is directly proportional to d_{ij} by the constant factor $10^{\Delta/(10n_p)} - 1$. Assuming constant standard deviation σ_{dB} of received power with distance, the range estimation error standard deviation will also increase proportionally with distance.

This characteristic of RSS-based range estimation leads to very high errors at large path lengths, which have limited its application in traditional location systems. However, in a dense sensor network, the distances between neighboring sensors is small, and a weighted least-squares estimator can be designed to fully utilize the accuracy of the range measurements made between the closest neighbors. A method for achieving this is proposed in the next section.

5.6 Adaptive Neighborhood Selection

Typically, neighbors are selected by choosing those devices which are closer than a threshold distance. But, since the exact distance is not known, we need to use noisy measurements to select neighbors. Range measurements, whether made via TOA, RSS, or proximity, are all subject to errors. In this section we discuss the biasing effects of selecting neighbors via noisy distance measurements, and how we can unbias the selection.

When distance is measured in noise, the act of thresholding neighbors based on the measured distance will tend to select the devices with smaller measured distances. For example, consider two devices separated by distance R, when R is also the threshold distance. With some positive probability (due to noise), the measured distance, δ, will be greater than R, and the two will not be considered neighbors. Alternatively, if $\delta \leq R$, the two will be considered neighbors, and δ will be used in the localization algorithm. The problem is that the expected value of δ, for devices

separated by R which consider themselves neighbors, is less than R. Thus, the measured distance is negatively biased because of the effect of thresholding. Note that selecting the K-nearest-neighbors effectively has an adaptive threshold, and thus does not avoid this biasing effect.

This bias has not been specifically addressed in the sensor localization literature, because its effects are not severe in certain systems. Some proposed sensor localization systems measure very accurate distances, eg., using TOA in UWB or a combination of RF and ultrasound media – for these systems, the effect of selecting neighbors based on measured distances will be minimal. Alternatively, if neighbors are selected based on independent means (eg., based on RSS or connectivity when range estimates are based on TOA), then the biasing effect is avoided[2]. Finally, when studies show results for the case in which all devices are connected to every other device, the thresholding step (and its biasing effect) is eliminated. In this chapter, we consider both noisy RSS measurements and small neighborhoods, so we cannot avoid the biasing effect. We limit our discussion to RSS measurements in this section, since low device costs and energy consumption are very attractive device characteristics of RSS, but the discussion is also applicable to systems which use noisy TOA-based range measurements for neighbor selection.

5.6.1 RSS-based Biasing Effect

When discussing thresholding based on RSS, we must make a distinction between the physical limits of the receiver and the threshold which we use to select neighbors, because generally, the two do not need to be the same. If a device has a large radio range to be robust to low device densities, it may want a stricter threshold when

[2]Note, however, for the RSS/TOA example, that if both are available, we may wish to use a combination of both; and if not, RSS and TOA for a link are correlated because objects in the environment tend to degrade both measurements simultaneously.

there are very many devices with which it can communicate. Denote P_{thr} to be the received power level below which a receiver cannot demodulate packets. (For most digital receivers with large frames and FEC, the frame error rate is very close to zero or very close to one for the vast majority of SNR, and the transition region is narrow. Thus, to a good approximation, for a constant noise level, we can state that packets above P_{thr} are received and demodulated correctly, while those below are not [77].) Denote P_R to be the received power level below which we do not include the transmitting device as a neighbor. Clearly, $P_R \geq P_{thr}$. Equivalently, we can define distances d_{thr} and d_R from (5.20) to be the range at which the mean received power is equal to P_{thr} and P_R, respectively.

Whether or not we select neighbors based on connectivity (measured power is greater than P_{thr}) or select them based on a power threshold (measured power is greater than P_R), the biasing effect will be the same. In following, we use P_R and d_R to indicate the thresholds (which may be set equal to P_{thr} and d_{thr} if desired).

Let $E\left[\delta_{ij}|P_{ij} > P_R\right]$ be the expected value of the range estimate between devices i and j given that the two are neighbors (i.e., the received power P_{ij} is greater than P_R). Using the RSS measurement model (see Section 5.5.2), it can be shown that

$$E\left[\delta_{ij}|P_{ij} > P_R\right] = \|x_i - x_j\| \, C \, \frac{1 - \Phi\left(\sqrt{\beta}\log\frac{\|x_i - x_j\|}{d_R} + \frac{1}{\sqrt{\beta}}\right)}{1 - \Phi\left(\sqrt{\beta}\log\frac{\|x_i - x_j\|}{d_R}\right)} \,, \tag{5.22}$$

where $\Phi(.)$ is the cumulative distribution function of a standard Gaussian random variable, $\beta = \frac{10n_p}{\sigma_{dB}\log 10}$, $C = \exp\left(\frac{1}{2\beta^2}\right)$ and σ_{dB}, n_p are channel parameters. Equation (5.22) is plotted in Fig. 5.2 as a function of the ratio of the true distance to d_R. Ideally, the range estimator should have a mean value equal to the actual range. However, as the range increases, the expected value of δ_{ij} (given that i and j are neighbors) deviates from linear and asymptotically becomes constant. There is a strong negative

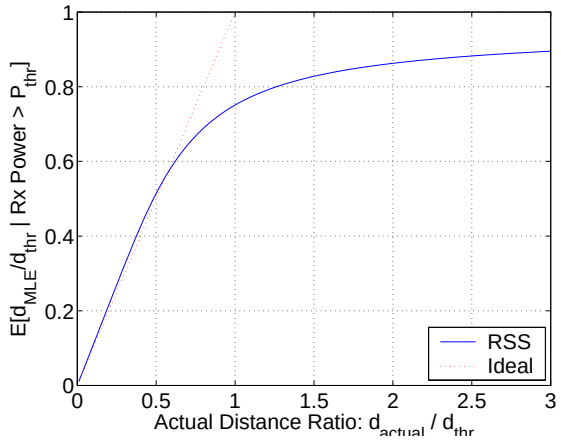


Figure 5.2: The expected value of the RSS-based estimate of range given that that two devices are neighbors (- - -), and the ideal unbiased performance (—). The channel has $\sigma_{dB}/n = 1.7$ and $d_R = 1$ (or equivalently, distances are normalized by d_R).

bias for devices separated by d_R or greater.

5.6.2 Two-Stage Selection Algorithm

Motivated by the negative bias phenomenon displayed in Fig. 5.2, we propose a two stage neighborhood selection process, based on the predicted distances between sensors.

In the first step, the dwMDS algorithm from Fig. 5.1 is run with a neighborhood structure based on the available range measurements, i.e., set $w_{ij} = 0$ if $\delta_{ij} > d_R$. After convergence, this step provides an interim estimate $\{\hat{\boldsymbol{x}}_i\}$ of the sensors locations. With high probability, the predicted distances between the estimated sensor locations will be negatively biased.

In the second step, these predicted distances from the estimated sensor locations are used to compute a new neighborhood structure, by assigning $w_{ij} = 0$ if $\|\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{x}}_j\| > d_R$. Some neighbors with low range measurements will be dropped, and some neighbors with possibly longer range measurements will be added. Then, using $\{\hat{\boldsymbol{x}}_i\}$

as an initial condition and the new neighborhood structure, the dwMDS algorithm is re-run, resulting in the final location estimates. Note that the predicted distances $\|\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{x}}_j\|$ are used only to select neighbors (i.e., which weights are positive) – the measured ranges δ_{ij} are still used to determine the weight values.

We remark that this 2-step algorithm does not imply twice the computation. The dwMDS algorithm is based on majorization, and each iteration brings it closer to convergence. Since the first step only needs to provide coarse localization information, it does not need to be very accurate, and so the dwMDS algorithm can be stopped quickly with a large ϵ. Next, the second step begins with very good (although biased) coordinate estimates, so the second run of the dwMDS algorithm will likely require fewer iterations to converge.

Note that for some of the devices which are considered neighbors in the 2nd run of the algorithm, the measured range δ_{ij} will actually be greater than d_R. Thus, to use this 2-step algorithm, d_R must be sufficiently less than the physical communication limit of the devices, d_{thr}, so that other range measurements can be considered. If we consider the non-circular (real-world) coverage area of a device, d_{thr} can be considered to be the mean radius of the coverage area, while d_R should be set to the minimum radius of the coverage area.

5.7 Experimental Localization Results

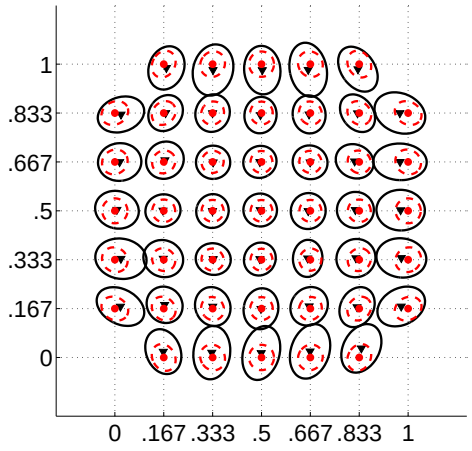
We apply the proposed MDS algorithm to the location problem in a network, using both simulated data and real data collected on an experimental sensor network.

5.7.1 Simulations

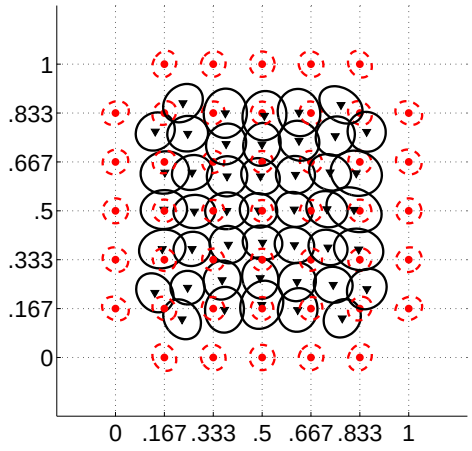
In this section, all the simulated data were generated from the RSS measurement model presented in Section 5.5.2, with channel parameters $\sigma_{dB}/n_p = 1.7$.

We first demonstrate the performance of the proposed algorithms on a network of 7×7 sensors arranged on a uniform grid of unit area, in which the four corner devices are anchor nodes and the remaining 45 are unknown location devices. For all experiments on this configuration, we use $d_R = 0.4\,\mathrm{m}$ (yielding an average of 14 neighbors per device). We ran 200 Monte Carlo simulation trials to determine confidence ellipses, root-mean-square error (RMSE) and bias performance (per sensor) of the location estimates. The results are displayed in Figure 5.3, where we plot the mean and 1-σ uncertainty ellipse of the estimator, and compare it to the actual device location and the Cramér-Rao lower bound (CRB) on the uncertainty ellipses which was presented for the case of RSS measurements in [79]. We remark that the CRB shown is calculated assuming full connectivity (all devices measure range to all other devices), and as such provides only a loose lower bound on the best performance achievable by any unbiased estimator. In the first experiment, we provide a baseline best-case scenario by using perfect (noise-free) distance measurements to select neighborhoods. The baseline assumes that we have an oracle to tell us when the true distance between i and j is less than a threshold, ie., $\|\boldsymbol{x}_i - \boldsymbol{x}_j\| < d_R$. This is shown in Figure 5.3(a), resulting in a RMSE of the location estimates of $0.090\,\mathrm{m}$ and an average bias of $0.019\,\mathrm{m}$.

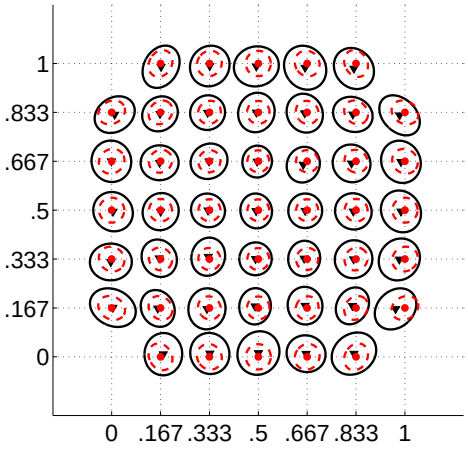
For the second experiment, we remove the assumption of perfect connectivity knowledge. Instead, we use RSS measurements to select neighbors, i.e., devices i and j are neighbors if $P_{ij} \geq P_R$, or, equivalently, if $\delta_{ij} \leq d_R$. The results are shown in Figure 5.3(b). The estimates are strongly pulled towards the center of the square, due to the negative bias of the range estimates which are 'selected' by the



(a) neighborhood selection using actual distances



(b) neighborhood selection using measured ranges



(c) adaptive neighborhood selection

Figure 5.3: Estimator mean (▾) and 1-σ uncertainty ellipse (—) for each blindfolded sensor compared to the true location (•) and CRB on the 1-σ uncertainty ellipse (- - -).

connectivity condition. Now, the RMSE is 0.162 m and the bias is 0.130 m.

A third experiment uses the adaptive neighborhood selection method proposed in Section 5.6.2. The results are displayed in Figure 5.3(c), where it can be seen that this method succeeds in removing the negative bias effect. The bias has gone back down to 0.012 m, while the RMSE is 0.092m, just slightly higher than the baseline experiment using the oracle.

Comparing Figure 5.3(c) and 5.3(a), the localization errors of the two-step algorithm are spread more evenly throughout the network compared to the first experiment – the errors for edge devices are reduced, getting closer in magnitude to those in the center. Based on the similarity of the RMSE in both experiments, we believe that the 2-step process eliminates most of the neighbor selection bias. Additionally, by changing the neighbor lists (and therefore the weights) and re-running the dwMDS algorithm, the 2nd iteration also provides the opportunity to break out local maxima, which are more likely to affect edge devices. Finally, the low variance achieved by the 2-stage algorithm is very close to the CRB which no unbiased location estimator can outperform, despite the fact that the CRB is an optimistic bound for the scenario considered here.

We also studied the influence of the threshold distance on the RMSE performance of the proposed algorithms. Figure 5.4 shows a plot of the RMSE vs. threshold distance (marked by '•'), for the 7×7 uniform grid example using adaptive neighborhood selection. It can be seen that there is an optimal threshold distance, $d_R = 0.5\,\mathrm{m}$, beyond which, no performance increase occurs. As d_R is increased beyond this optimal value, more distant sensors are included in the cost function. By the RSS measurement model, the accuracy of range measurements degrades quickly with distance, thus adding these far way sensors will not bring any gain to the estimation algorithm. Figure 5.4 also shows the behavior of the same quantities for the case

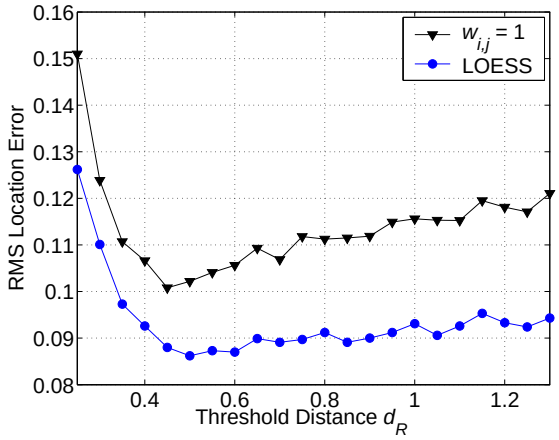


Figure 5.4: RMSE versus threshold distance for the 7×7 uniform grid example using adaptive neighborhood selection, for different weighting schemes.

of a naive weighting scheme: all measured distances have equal weights, while non observed distances have weights set zero. The worse performance of this simple scheme shows further evidence to support the claim that weights should be chosen adaptively to reflect measurement accuracy and further justifies the LOESS type weighting scheme (cf. (5.16)) adopted.

Finally, we show the influence of different choices of prior weighting r_i in the quality of the localization solution. Under the same scenario, we now consider the four corner nodes to have imperfect information. In particular, the algorithm only has access to a noisy version of the actual coordinates of these nodes, perturbed by zero mean Gaussian noise with unknown variance σ_p^2. Figure 5.5 shows the resulting RMSE, obtained by running 5000 Monte-Carlo Simulation trials for $\sigma_p = 0.025$, 0.050, 0.100 and setting $r_i = r$ for the corner nodes, where r is made to vary between 10^{-2} and 10^2.

From Figure 5.5, it can be observed that for small values of r, the RMSE levels off to a value that is constant across different values of σ_p^2. Essentially, the prior information is only being used to rotate and translate the final solution obtained by

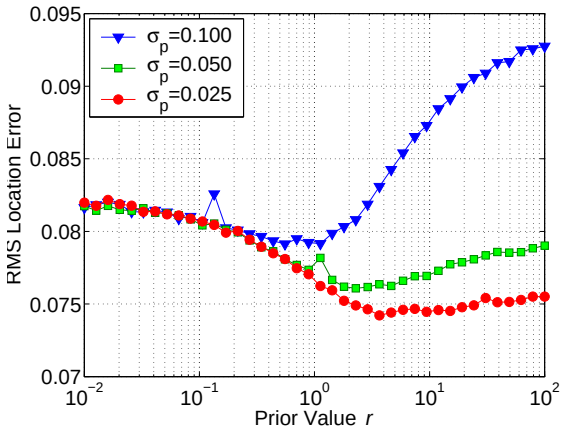


Figure 5.5: RMSE versus prior weighting of the four corner nodes in the 7×7 uniform grid example using adaptive neighborhood selection.

the MDS algorithm to best fit the estimated position of the anchor nodes. On the other hand, for high values of r, a similar phenomenon occurs, but this time due to the fact that the MDS algorithm is focusing on placing sensors with prior information at their *a priori* coordinates, to the detriment of fitting range measurements.

For intermediate values of r, there is an optimal r which best weights the relative information in the prior coordinates with respect to the weights chosen for the measured ranges. As we would expect, the optimal r is inversely proportional to σ_p^2, although the exact dependency is influenced by factors such as the prior coordinates noise distribution, the weighting scheme chosen, and the number of neighbors of each node. Further research should investigate these dependencies. However, the RMSE near the optimal is a very shallow function of r – for all three curves, there is nearly an order of magnitude range within which the RMSE is within 1% of its minimum. So, although simulation might be necessary to find the optimal r, as long as r is within the correct order of magnitude, the results will be nearly optimal. This suggest that, for little or no knowledge about the perturbations to prior coordinates used by the algorithm, choosing intermediate values of r would be a good rule. In particular, this would result in a better RMSE than possibly using either:

– a method that uses prior coordinate information only to find the best rotation of a calculated relative map [98], which is analogous to low r in the dwMDS method; or

– an MLE method which assumes that anchor coordinates are known perfectly [79], which is analogous to high r.

5.7.2 Localization in a Measured Network

To test the performance of the proposed algorithm on real-world channel measurements, we used the RSS and TOA measurements presented in [79]. This data set includes the RSS and TOA range measurements from a network of 44 devices (4 of which are anchor nodes) using a wideband direct-sequence spread-spectrum (DS-SS) transmitter and receiver pair operating at a center frequency of 2.4 GHz. The measurements were made in an open plan office building, within a 14×14 m square area. The RSS between each pair of devices was measured 10 times, from which the average was calculated and labeled as P_{ij}, for each pair (i, j).

We use the bias-corrected MLE to estimate range from the RSS, i.e.,

$$\delta_{ij} = \frac{d_0}{C} 10^{(P_0 - P_{ij})/(10 n_p)} \ . \tag{5.23}$$

We choose to divide by C in (5.23) because this estimator, as opposed to the MLE in (5.21), is unbiased, ie., $E[\delta_{ij}] = d_{ij}$. See [79] for details.

To give the reader a feeling of how challenging is to do sensor localization using RSS range measurements in a real live scenario, we plot, in Figure 5.6, the error between range measurements and real distances, i.e., $\delta_{ij} - d_{ij}$. Note that the standard deviation of the RSS-based range estimator error increases steadily with distance. But, most importantly, the error as a percentage of actual range is often high: there

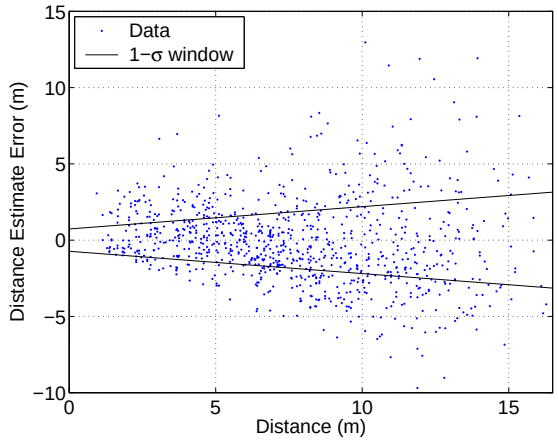


Figure 5.6: Plot of distance measurement errors Vs. distance. The $1-\sigma$ interval superimposed on the plot was obtained from the ML fit of the error measurement model $\mathcal{N}(d_{ij}, (a\, d_{ij} + b)^2)$.

	Classical MDS	**MLE** [79]	**dwMDS**
RSS	4.30 m	**2.18** m	2.48 m
TOA	1.96 m	1.23 m	**1.12** m

Table 5.2: RMSE of location estimates in experimental network

are several range errors larger than 100% of the actual range.

We compare the performance of the dwMDS algorithm with adaptive neighborhood selection to classical MDS and the MLE based solutions from [79]. Table 5.2 summarizes the RMSE of the location estimates. Figures 5.7(a) and 5.7(b) show the location estimates using classical MDS (which used all the pairwise range measurements between sensors) and the dwMDS algorithm, for the RSS measurement data set. The true and estimated sensor positions are marked by 'o' and '$\triangledown$', respectively, where the lines represent the estimation errors. The anchor nodes are marked with an 'x'. It can be observed that the dwMDS algorithm does much better than classical MDS. On the other hand, the RMSE of the dwMDS algorithm is slightly higher than the RMSE of the centralized MLE reported in [79]. However, that method not only uses all pairwise range measurements, but also relies on previously estimating

the channel parameters. If we allow d_R to increase at the expense of increasing communication costs, the dwMDS algorithm can reach an RMSE as low as 2.269 m for $d_R = 8.5$ m.

Figure 5.7(c) and 5.7(d) show again the location estimates using classical MDS and the dwMDS algorithm, but this time for the TOA measurement data set. From Table 5.2, it can be seen that the dwMDS algorithm outperforms all other location estimators. If we allow d_R to increase at the expense of increasing communication costs, the dwMDS algorithm can reach an RMSE as low as 0.940 m for $d_R = 7.5$ m. Once again, we stress that the dwMDS algorithm, unlike the MLE estimator from [79], does not use all the pairwise range measurements and does not assume knowledge of the distribution of the range measurements.

5.8 Conclusion

This chapter proposes a distributed weighted-MDS method specially suited for node localization in a wireless sensor network. First, the method reflects the distributed nature of the problem, incorporating network communication constraints in its design. In this way, the need to transmit all range measurements to a central unit is eliminated, resulting in energy savings for a dense sensor network. Second, the inhomogeneous character of range measurements in a wireless network is accounted for by introducing weights that adaptively emphasize measurements believed to be more accurate. We stress that the dwMDS algorithm is nonparametric in its nature, i.e., it does not depend on any particular channel or range measurement models. This makes it applicable to a broad range of distance measurements, e.g., RSS, TOA, proximity, without the need to tweak any parameters. We have shown via simulation that the algorithm has excellent bias and variance performance compared to

the CRB, and that its performance in a real-world sensor network is similar to the centralized MLE algorithm.

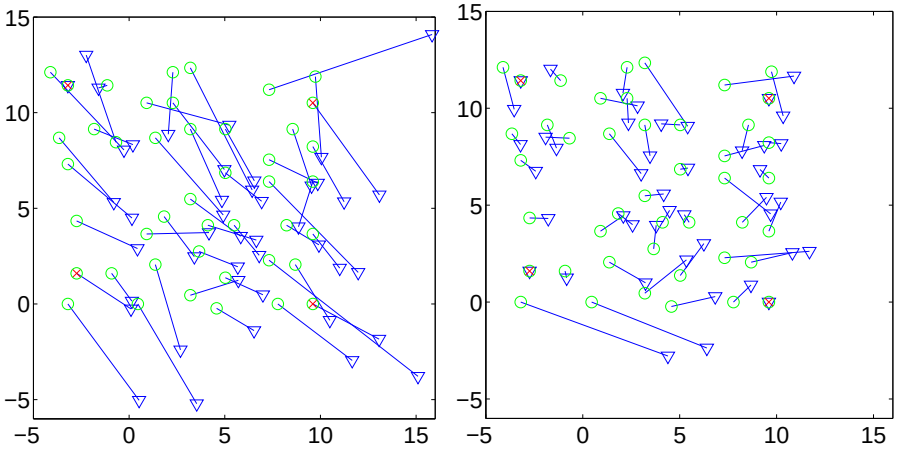
We remark that the dwMDS algorithm can be applied more generally to dimensionality reduction problems across a network of processors, such as Internet monitoring [76] or distributed sensor data compression. To make it more general, other distance metrics can be used, such as L_p $(1 \leq p \leq 2)$ metrics. In this case, a majorization technique can still be used which guarantees a non-increasing cost function. For other general distances (without any convex structure), gradient descent techniques can be used. In particular, incremental gradient methods fit well the framework considered in this chapter and might provide faster convergence rates at the cost of losing the monotonicity of the cost function. Other developments that can improve the algorithm's performance include extending the formulation to include non range measurements like AOA, or adding new terms to the cost function that model the correlations between range measurements.

5.9 Appendix

In this appendix, we give an expression for the gradient of the majorizing function T_i defined by equation (5.10).

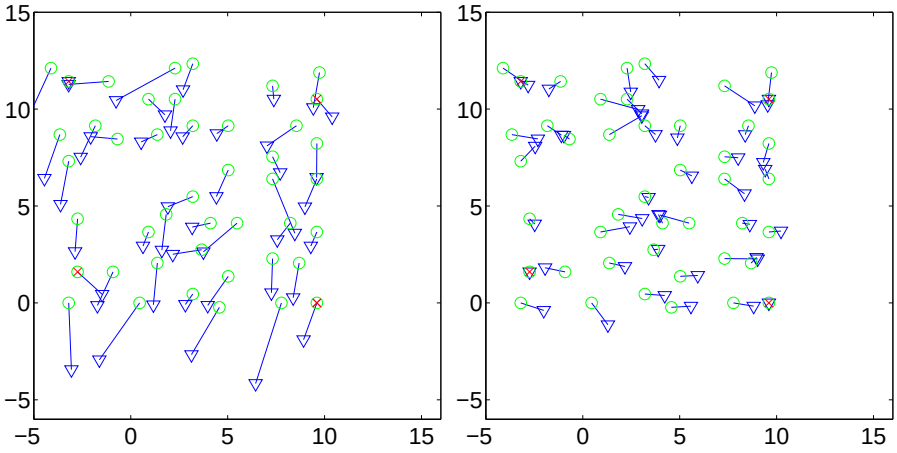
$$\frac{1}{2}\,\frac{\partial T_i(\boldsymbol{x}_i, \boldsymbol{y}_i)}{\partial \boldsymbol{x}_i} = \left(\sum_{\substack{j=1 \\ j\neq i}}^{n} \overline{w}_{ij} + \sum_{j=n+1}^{n+m} 2\,\overline{w}_{ij} + r_i \right) \boldsymbol{x}_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} \overline{w}_{ij}\,\boldsymbol{x}_j - \sum_{j=n+1}^{n+m} 2\,\overline{w}_{ij}\,\boldsymbol{x}_j -$$

$$
\begin{aligned}
&- r_i \, \overline{\boldsymbol{x}}_i - \left[\sum_{\substack{j=1 \\ j \neq i}}^{n} \overline{w}_{ij} \frac{\overline{\delta}_{ij}}{d_{ij}(Y)} + \sum_{j=n+1}^{n+m} 2 \, \overline{w}_{ij} \frac{\overline{\delta}_{ij}}{d_{ij}(Y)} \right] \boldsymbol{y}_i \\
&+ \sum_{\substack{j=1 \\ j \neq i}}^{n} \overline{w}_{ij} \frac{\overline{\delta}_{ij}}{d_{ij}(Y)} \boldsymbol{y}_j + \sum_{j=n+1}^{n+m} 2 \, \overline{w}_{ij} \frac{\overline{\delta}_{ij}}{d_{ij}(Y)} \boldsymbol{y}_j \ .
\end{aligned}
\tag{5.24}
$$



(a) Classical MDS (RSS)

(b) dwMDS (RSS)



(c) Classical MDS (TOA)

(d) dwMDS (TOA)

Figure 5.7: Location estimates using RSS and TOA range measurements from experimental sensor network. True and estimated sensor locations are marked, respectively, by 'o' and '▽', while anchor nodes are marked by 'x'. The dwMDS algorithm uses adaptive neighbor selection, with $d_R = 6\,\mathrm{m}$.

CHAPTER 6

Conclusion and Future Work

6.1 Summary

This thesis was motivated by the challenging problem of analyzing high-dimensional data and information that must be transmitted, stored and processed accurately using manageable computational resources. The research presented here was guided by the desire to accomplish two fundamental goals. First, we aimed at contributing to a deeper understanding of such data sets by studying some of their fundamental theoretical properties. Second, we aimed at developing practical algorithms that can have an impact on the analysis of real data. We proposed to achieve such research program by using random graphs as a general tool to analyze high-dimensional data sets. On the one hand, they are associated with computational efficient and scalable algorithms that can be used to model statistical and spatial constraints describing high-dimensional data. On the other hand, their theoretical performance can be predicted using tools from probability, statistics and differential geometry. Although random graphs have been present in the fields of signal processing or pattern recognition for several decades, it was not until recently that researchers have started working towards bridging the gap between statistical pattern recognition and com-

putational geometry (e.g., [64]). We hope this thesis is a contribution towards that goal.

We started this thesis by introducing minimal Euclidean graphs and discussing the asymptotic behavior of power weighted functionals of such graphs as nonparametric entropy estimators. We derived upper bounds on the convergence rates in terms of the smoothness of the probability densities involved and lower bounds using a minimax framework.

The core of this thesis dealt with detection, representation and processing of data that spans a high-dimensional space but which contain fundamental features that are concentrated on lower-dimensional subsets of this space – curves, surfaces or, more generally, lower-dimensional manifolds. In particular, we have developed a novel geometric probability approach to the problem of estimating intrinsic dimension and entropy of manifold data, based on asymptotic properties of computational efficient graphs such as Minimal Spanning Trees or k-Nearest Neighbor graphs. Unlike previous solutions to this problem, we were able to prove statistical consistency of the obtained estimators for the wide class of Riemann submanifolds of an Euclidean space. These algorithms were successfully applied to high-dimensional image databases of faces and handwritten digits.

Complementary to the aforementioned problem, we designed a method to learn low-dimensional features of high-dimensional data aimed at improving classification tasks. Current methods are only concerned with finding a low-dimensional embedding of the original data without satisfying any specific constraints. In particular, the embedding found may result in a harder classification problem than in the original space. We developed a spectral method based on graph Laplacians that produces a low-dimensional embedding, where class separation constraints were taken into account.

Taking advantage of the neighborhood graphs introduced for the analysis of high-dimensional data sets, we looked at signal processing problems arising in a sensor network environment, where communication constraints exist and distributed optimization is required. We realized that the graph structures developed before fit this framework by using them to model spatial dependencies within the network. In particular, using some of the tools derived for the analysis of high-dimensional data, we have developed a scalable distributed Multidimensional Scaling (MDS) algorithm for node localization in a sensor network. This method was successfully applied to node localization in an experimental sensor network, using both received signal strength and time of arrival range measurements.

6.2 Future Work

The research presented in this thesis is part of a long term project that encompasses building a framework to address several pressing problems in data processing today. Examples of such problems and methodologies considered include: developing dimensionality reduction algorithms to extract relevant features from a genetic database, improve face recognition systems or handwritten digits classification; creating new nonparametric tools for robust high-dimensional pattern recognition or information retrieval; developing visualization mechanisms for high-dimensional data that helps analyzing gene expression data or detect anomalies in a computer network.

Current algorithms for classification and clustering do not scale well with data dimensionality. Following the same graph theoretic approach used in the development of direct entropy estimators, which showed its wide applicability to high-dimensional data, we are now working on nonparametric estimation of divergence measures and nonlinear correlation coefficients. In particular, we are generalizing Friedman-Rafsky

goodness of fit test to discriminate between samples from two populations, by looking at edges in a MST connecting samples from different distributions. These estimators are the first step towards developing new scalable and robust tools for detection, clustering and information indexing and retrieval of high-dimensional signals. Also of great interest is exploring the close connections between the asymptotic behavior of the minimal graphs discussed in this thesis and asymptotic results for high-rate vector quantization.

Regarding dimensionality estimation, many problems remain open. We list the following:

- Analyzing the case of manifolds sampled with noise.
- Improving the resampling mechanism used by the proposed algorithm and taking into account the possible dependencies between successive subsamplings of the data.
- Accessing the performance of the intrinsic dimension estimator.

In another direction, we are now actively working on developing a theory and algorithms to estimate local intrinsic dimension of data points, as opposed to the global intrinsic dimension of the data set. This involves studying the properties of local neighborhood graphs that can guarantee adjacency relations only among data points that share the same topological properties. This can have a deep impact on problems where data dimensionality can be used to discriminate between patterns of different complexity. For example, to segment different textures in an image. Another application, for which we have encouraging preliminary results indicate, is to detect anomalies in the traffic flow of an Internet backbone network. Such extensions will require block bootstrap methods that account for dependency.

Continuing ongoing efforts on designing classification constrained embedding algorithms, the following are the next steps:

- Develop an out-of-sample extension for the non-linear mapping.
- Study the effect of the regularization parameter that trades off the preservation of geometric structure with class label information.

In the long term, we plan to work on developing the tools needed to do a quantitative analysis of performance gains in classification obtained by using dimensionality reduction methods. The final goal is to understand the connections and tradeoff between the effect of dimensionality reduction in classification and dimensionality expansion and the performance of state-of-the art classifiers such as support vector machines.

Finally, the dwMDS algorithm introduced in chapter 5 is more general than the application to sensor localization that was used to motivate and illustrate the theory. In particular, it has a natural application in dimensionality reduction for spatio-temporal data visualization. This method can be applied to manifold learning with prior information or to Internet traffic time series visualization, aimed at detecting anomalies in the network [76]. Another problem closely related to sensor network localization is the inference of molecular conformation. In this problem, one hopes to determine the 3-dimensional structure of a protein from a subset of all possible inter-atomic distances, obtained, for example, by nuclear magnetic resonance and/or prior modeling of molecule interactions [105]. We are currently exploring the application of the dwMDS algorithm to this scenario.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] J. Albowicz, A. Chen, and L. Zhang, "Recursive position estimation in sensor networks," in *IEEE Int. Conf. on Network Protocols*, Nov. 2001, pp. 35–41.

[2] S. Arora, "Nearly linear time approximation schemes for Euclidean TSP and other geometric problems," in *Proceedings of IEEE Symposium on Foundations of Computer Science*, 1997.

[3] D. Banks, M. Lavine, and H. J. Newton, "The minimal spanning tree for nonparametric regression and structure discovery," in *Proceedings of the 24th Symposium on the Interface and Computing Science and Statistics*, H. J. Newton, Ed., 1992, pp. 370–374.

[4] J. Beardwood, J. H. Halton, and J. M. Hammersley, "The shortest path through many points," *Proc. Cambridge Philosophical Society*, vol. 55, pp. 299–327, 1959.

[5] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. van der Meulen, "Non-parametric entropy estimation: an overview," *Intern. J. Math. Stat. Sci.*, vol. 6, no. 1, pp. 17–39, 1997.

[6] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, June 2003.

[7] ——, "Semi-supervised learning on Riemannian manifolds," *Machine Learning*, vol. 56, pp. 209–239, 2004, special Issue on Clustering.

[8] J. Benzécri, *L'Analyse des Données, Tome 2, L'Analyse des Correspondences*. Paris: Dunod, 1973.

[9] M. Bernstein, V. de Silva, J. C. Langford, and J. B. Tenenbaum, "Graph approximations to geodesics on embedded manifolds," Department of Psychology, Stanford University, Tech. Rep., 2000.

[10] L. Birgé and P. Massart, "Estimation of integral functionals of a density," *The Annals of Statistics*, vol. 23, no. 1, pp. 11–29, 1995.

[11] J. Caffery Jr. and G. L. Stuber, "Subscriber location in cdma cellular networks," *IEEE Trans. on Veh. Tech.*, vol. 47, no. 2, pp. 406–416, May 1998.

[12] F. Camastra and A. Vinciarelli, "Estimating the intrinsic dimension of data with a fractal-based method," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 10, pp. 1404–1407, October 2002.

[13] M. Carmo, *Riemannian geometry*. Boston: Birkhäuser, 1992.

[14] A. Catovic and Z. Sahinoglu, "The Cramer-Rao bounds of hybrid TOA/RSS and TDOA/RSS location estimation schemes," Mitsubishi Electric Research Laboratory, Tech. Rep. TR2003-143, Jan. 2004. [Online]. Available: http://www.merl.com/

[15] P.-C. Chen, "A non-line-of-sight error mitigation algorithm in location estimation," in *IEEE Wireless Comm. and Networking Conf.*, Sept. 1999, pp. 316–320.

[16] W. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Jornal of the American Statistical Association*, vol. 74, no. 368, pp. 829–836, 1979.

[17] N. S. Correal, S. Kyperountas, Q. Shi, and M. Welborn, "An ultra wideband relative location system," in *IEEE Conf. on Ultra Wideband Systems and Technologies*, Nov. 2003.

[18] J. A. Costa and A. O. Hero, "Geodesic entropic graphs for dimension and entropy estimation in manifold learning," *IEEE Trans. on Signal Processing*, vol. 52, no. 8, pp. 2210–2221, August 2004.

[19] A. J. Coulson, A. G. Williamson, and R. G. Vaughan, "A statistical basis for lognormal shadowing effects in multipath fading channels," *IEEE Trans. on Veh. Tech.*, vol. 46, no. 4, pp. 494–502, April 1998.

[20] D. Cox, "Delay Doppler Characteristics of Multipath Propagation at 910 MHz in a Suburban Mobile Radio Environment," *IEEE Transactions on Antennas and Propagation*, vol. AP-20, no. 5, pp. 625–635, Sep 1972.

[21] T. Cox and M. Cox, *Multidimensional Scaling*. London: Chapman & Hall, 1994.

[22] N. A. Cressie, *Statistics for spatial data*. Wiley, NY, 1993.

[23] M. L. Davidson, *Multidimensional Scaling*. Ney York: Wiley, 1983.

[24] M. T. Dickerson and D. Eppstein, "Algorithms for proximity problems in higher dimensions," *Comput. Geom. Theory and Appl.*, vol. 5, no. 5, pp. 277–291, 1996.

[25] L. Doherty, K. S. J. pister, and L. E. Ghaoui, "Convex position estimation in wireless sensor networks," in *IEEE INFOCOM*, vol. 3, 2001, pp. 1655 –1663.

[26] D. Donoho and C. Grimes, "Hessian eigenmaps: locally linear embedding techniques for high dimensional data," *Proc. Nat. Acad. of Sci.*, vol. 100, no. 10, pp. 5591–5596, 2003.

[27] H. Edelsbrummer, M. Facello, and J. Liang, "On the definition and the construction of pockets on macromolecules," *Discrete Applied Math.*, vol. 88, pp. 83–102, 1998.

[28] R. Fleming and C. Kushner, "Low-power, miniature, distributed position location and communication devices using ultra-wideband, nonsinusoidal communication technology," Aetherwire Inc., Semi-Annual Technical Report, ARPA Contract J-FBI-94-058, Tech. Rep., July 1995.

[29] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[30] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. on Inform. Theory*, vol. 28, pp. 373–380, 1979.

[31] A. Gersho and R. M. Gray, *Vector quantization and signal compression.* Kluwer, Boston MA, 1992.

[32] L. Girod, V. Bychkovskiy, J. Elson, and D. Estrin, "Locating tiny sensors in time and space: a case study," in *IEEE Int. Conf. on Computer Design*, 2002, pp. 214–219.

[33] S. Graf and H. Luschgy, *Foundations of Quantization for Probability Distributions*, ser. Lecture Notes in Mathematics. Springer-Verlag, Berlin Heidelberg, 2000.

[34] P. Grassberger and I. Procaccia, "Measuring the stangeness of strange attractors," *Physica D*, vol. 9, pp. 189–208, 1983.

[35] M. J. Greenacre, *Theory and Applications of Correspondence Analysis.* London: Academic Press Inc., 1984.

[36] P. Groenen, *The majorization approach to multidimensional scaling: some problems and extensions.* DSWO Press, 1993.

[37] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. on Inform. Theory*, vol. 46, no. 2, pp. 388–404, 2000.

[38] H. Hashemi, "The Indoor Radio Propagation Channel," *Proceedings of the IEEE*, vol. 81, no. 7, pp. 943–968, July 1993.

[39] A. Hero and O. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Trans. on Inform. Theory*, vol. 45, no. 6, pp. 921–1939, September 1999.

[40] A. Hero, B. Ma, O. Michel, and J. Gorman, "Applications of entropic spanning graphs," *IEEE Signal Processing Magazine*, vol. 19, no. 5, pp. 85–95, October 2002.

[41] A. Hero and O. Michel, "Estimation of Rényi information divergence via pruned minimal spanning trees," in *IEEE Workshop on Higher Order Statistics*, Caesaria, Israel, Jun. 1999.

[42] R. Hoffman and A. K. Jain, "A test of randomness based on the minimal spanning tree," *Pattern Recognition Letters*, vol. 1, pp. 175–180, 1983.

[43] C. Huber, "Lower bounds for function estimation," in *Festschrift for Lucien Le Cam*, D. Pollard, E. Torgersem, and G. Yang, Eds. Springer-Verlag, New York, 1997, pp. 245–258.

[44] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Englewood Cliffs, NJ: Prentice Hall, 1988.

[45] X. Ji and H. Zha, "Sensor positioning in wireless ad-hoc sensor networks with multidimensional scaling," in *IEEE INFOCOM (to appear)*, 2004.

[46] R. M. Karp, "The probabilistic analysis of some combinatorial search algorithms," in *Algorithms and complexity: New directions and recent results*, J. F. Traub, Ed. New York: Academic Press, 1976, pp. 1–19.

[47] ——, "Probabilistic analysis of partitioning algorithms for the traveling salesman problem," *Oper. Res.*, vol. 2, pp. 209–224, 1977.

[48] R. M. Karp and J. M. Steele, "Probabilistic analysis of heuristics," in *The Traveling Salesman Problem: A guided tour of combinatorial optimization*, E. L. Lawler, J. K. Lenstra, A. H. G. R. Kan, and D. B. Shmoys, Eds. Wiley, New York, 1985, pp. 181–206.

[49] B. Kégl, "Intrinsic dimension estimation using packing numbers," in *Neural Information Processing Systems: NIPS*, Vancouver, CA, Dec. 2002.

[50] G. Kerkyacharian and D. Picard, "Estimating nonquadratic functionals of a density using Haar wavelets," *The Annals of Statistics*, vol. 24, no. 2, pp. 485–507, 1996.

[51] S. Kim, T. Pals, R. Iltis, and H. Lee, "CDMA multipath channel estimation using generalized successive interference cancellation algorithm for radiolocation," in *37th Annual Conference on Information Sciences and Systems*, March 2002.

[52] M. Kirby, *Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns*. Wiley-Interscience, 2001.

[53] A. Korostelev and A. Tsybakov, *Minimax theory of image reconstruction*. Springer-Verlag, New York, 1993.

[54] J. Kruskal, "Multidimensional scaling by optmizing goodness-of-fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, pp. 1–27, 1964.

[55] ——, "Nonmetric multidimensional scaling: a numerical method," *Psychometrika*, vol. 29, pp. 115–129, 1964.

[56] S. Lafon, "Diffusion maps and geometric harmonics," Ph.D. dissertation, Yale University, May 2004.

[57] K. Lange, D. R. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions," *Journal of Computational and Graphical Statistics*, vol. 9, no. 1, pp. 1–20, March 2000.

[58] B. Laurent, "Efficient estimation of integral functionals of a density," *The Annals of Statistics*, vol. 24, no. 2, pp. 659–681, 1996.

[59] E. L. Lawler, J. K. Lenstra, A. H. G. R. Kan, and D. B. Shmoys, *The traveling salesman problem*. Wiley, New York, 1985.

[60] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, November 1998.

[61] O. Lepski, A. Nemirovski, and V. Spokoiny, "On estimation of the L_r norm of a regression function," *Probab. Theory Relat. Fields*, vol. 113, pp. 221–253, 1999.

[62] E. Levina and P. Bickel, "Maximum likelihood estimation of intrinsic dimension," in *Neural Information Processing Systems: NIPS*, Vancouver, CA, Dec. 2004.

[63] B. Ma, "Parametric and non-parametric approaches for multisensor data fusion," Ph.D. dissertation, University of Michigan, Ann Arbor, 2001.

[64] D. Marchette, *Random Graphs for Statistical Pattern Recognition*. Wiley, 2004.

[65] F. Mémoli and G. Sapiro, "Distance functions and geodesic distances on point clouds," *to appear in SIAM Journal of Applied Math.*, 2005, (Tech. Rep. 1902, IMA, University of Minnesota, Mineapolis).

[66] J. Mitchell, "Guillotine subdivisions approximate polygonal subdivisions: a simple new method for the geometric k-MST problem," in *Proc. of ACM-SIAM Symposium on Discrete Algorithms*, 1996, pp. 402–408.

[67] R. L. Moses, D. Krishnamurthy, and R. Patterson, "An auto-calibration method for unattended ground sensors," in *Proc. of IEEE Int. Conf. on Acoust. Speech and Signal Processing*, vol. 3, May 2002, pp. 2941–2944.

[68] ——, "A self-localization method for wireless sensor networks," *EURASIP Journal on Applied Sig. Proc.*, no. 4, pp. 348–358, Mar. 2003.

[69] R. Nagpal, H. Shrobe, and J. Bachrach, "Organizing a global coordinate system from local information on an ad hoc sensor network," in *2nd Intl. Workshop on Inform. Proc. in Sensor Networks*, April 2003.

[70] H. Neemuchwala, A. O. Hero, and P. Carson, "Image registration using entropy measures and entropic graphs," *European Journal of Signal Processing, Special Issue on Content-based Visual Information Retrieval*, vol. 85, no. 2, pp. 277–296, 2005.

[71] D. N. Neuhoff, "On the asymptotic distribution of the errors in vector quantization," *IEEE Trans. on Inform. Theory*, vol. 42, pp. 461–468, March 1996.

[72] M. Neumann, "Multivariate wavelet thresholding: a remedy against the curse of dimensionality," Preprint no. 229, Weierstrass Institute, Berlin, 1996. [Online]. Available: citeseer.nj.nec.com/neumann96multivariate.html

[73] D. Niculescu and B. Nath, "Ad hoc positioning system," in *IEEE Globecom 2001*, vol. 5, April 2001, pp. 2926 –2931.

[74] ——, "Error characteristics of ad hoc positioning systems," in *ACM MOBIHOC*, May 2004.

[75] K. Pahlavan, P. Krishnamurthy, and J. Beneat, "Wideband radio propagation modeling for indoor geolocation applications," *IEEE Comm. Magazine*, pp. 60–65, April 1998.

[76] N. Patwari and A. O. Hero, "Manifold learning visualization of network traffic data," in *SIGCOMM 2005 Workshop on Mining Network Data*, Philadelphia, August 2005.

[77] N. Patwari and A. O. Hero III, "Using proximity and quantized RSS for sensor localization in wireless networks," in *IEEE/ACM 2nd Workshop on Wireless Sensor Nets. & Applications*, Sept. 2003.

[78] ——, "Manifold learning algorithms for localization in wireless sensor networks," in *Proc. of IEEE Int. Conf. on Acoust. Speech and Signal Processing*, May 2004.

[79] N. Patwari, A. O. Hero III, M. Perkins, N. Correal, and R. J. O'Dea, "Relative location estimation in wireless sensor networks," *IEEE Trans. Sig. Proc.*, vol. 51, no. 8, pp. 2137–2148, Aug. 2003.

[80] M. Penrose, "A strong law for the largest nearest-neighbour link between random points," *J. London Math. Soc*, vol. 60, no. 2, pp. 951–960, 1999.

[81] M. Penrose and J. Yukich, "Weak laws of large numbers in geometric probability," *Annals of Applied Probability*, vol. 13, no. 1, pp. 277–303, 2003.

[82] K. Pettis, T. Bailey, A. Jain, and R. Dubes, "An intrinsic dimensionality estimator from near-neighbor information," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 1, no. 1, pp. 25–36, 1979.

[83] D. Pollard, "Asymptopia," `http://www.stat.yale.edu/~pollard/Asymptopia/`.

[84] A. J. E. M. J. R. Baraniuk, P. Flandrin and O. Michel, "Measuring time frequency information content using the rényi entropies," *IEEE Trans. on Inform. Theory*, vol. 47, no. 4, 2001.

[85] M. Rabbat and R. Nowak, "Distributed optimization in sensor networks," in *3rd Int. Symp. on Information Processing in Sensor Networks (IPSN'04)*, Berkeley, CA, April 2004.

[86] J. Ramsay, "Some statiscal approaches to multidimensional scaling data," *J. R. Statist. Soc. A*, vol. 145, part 3, pp. 285–312, 1982.

[87] T. S. Rappaport, *Wireless Communications: Principles and Practice.* New Jersey: Prentice-Hall Inc., 1996.

[88] R. Ravi, M. Marathe, D. Rosenkrantz, and S. Ravi, "Spanning trees short or small," in *Proc. 5th Annual ACM-SIAM Symposium on Discrete Algorithms*, Arlington, VA, 1994, pp. 546–555.

[89] ——, "Spanning trees – short or small," *SIAM Journal on Discrete Math*, vol. 9, pp. 178–200, 1996.

[90] C. Redmond and J. E. Yukich, "Limit theorems and rates of convergence for Euclidean functionals," *Annals of Applied Probability*, vol. 4, no. 4, pp. 1057–1073, 1994.

[91] ——, "Asymptotics for Euclidean functionals with power weighted edges," *Stochastic Processes and their Applications*, vol. 6, pp. 289–304, 1996.

[92] W. T. Rhee, "A matching problem and subadditive Euclidean functionals," vol. 3, pp. 794–801, 1993.

[93] B. D. Ripley, *Pattern recognition and neural networks*. Cambridge U. Press, 1996.

[94] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear imbedding," *Science*, vol. 290, no. 1, pp. 2323–2326, 2000.

[95] C. Savarese, J. M. Rabaey, and J. Beutel, "Locationing in distributed ad-hoc wireless sensor networks," in *Proc. of IEEE Int. Conf. on Acoust. Speech and Signal Processing*, May 2001, pp. 2037–2040.

[96] A. Savvides, H. Park, and M. B. Srivastava, "The bits and flops of the n-hop multilateration primitive for node localization problems," in *Intl. Workshop on Sensor Nets. & Apps.*, Sept. 2002, pp. 112–121.

[97] B. Schlkopf, A. Smola, and K. Miller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, 1998.

[98] Y. Shang, W. Ruml, Y. Zhang, and M. P. J. Fromherz, "Localization from mere connectivity," in *Proc. of the 4th ACM Int. Symp. on Mobile Ad Hoc Networking & Computing*, June 2003, pp. 201–212.

[99] J. M. Steele, "Growth rates of euclidean minimal spanning trees with power weighted edges," *Annals of Probability*, vol. 16, pp. 1767–1787, 1988.

[100] ——, *Probability theory and combinatorial optimization*, ser. CBMF-NSF Regional Conferences in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), 1997, vol. 69.

[101] M. Szummer and T. Jaakkola, "Partially labeled classification with markov random walks," in *Advances in Neural Information Processing Systems 14*, 2002.

[102] Y. Takane, F. Young, and J. de Leeuw, "Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features," *Psychometrika*, vol. 42, pp. 7–67, 1977.

[103] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[104] G. Toussaint, "The relative neighborhood graph of a finite planar set," *Pattern Recognition*, vol. 12, pp. 261–268, 1980.

[105] M. W. Trosset, "Applications of multidimensional scaling to molecular conformation," *Computing Science and Statistics*, vol. 29, pp. 148–152, 1998.

[106] S. Čapkun, M. Hamdi, and J.-P. Hubaux, "GPS-free positioning in mobile ad-hoc networks," in 34^{th} *IEEE Hawaii Int. Conf. on System Sciences (HICSS-34)*, Jan. 2001.

[107] K. Weinberger and L. Saul, "Unsupervised learning of image manifolds by semidefinite programming," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Washington D.C., 2004.

[108] B. Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam*, D. Pollard, E. Torgersem, and G. Yang, Eds. Springer-Verlag, New York, 1997, pp. 423–435.

[109] J. E. Yukich, *Probability theory of classical Euclidean optimization problems*, ser. Lecture Notes in Mathematics. Springer-Verlag, Berlin, 1998, vol. 1675.

[110] C. Zahn, "Graph-theoretical methods for detecting and describing Gestalt clusters," *IEEE Trans. on Computers*, vol. C-20, pp. 68–86, 1971.

[111] Z. Zang and H. Zha, "Principal manifolds and nonlinear dimension reduction via local tangent space alignment," *SIAM Journal of Scientific Computing*, vol. 26, no. 1, pp. 313–338, 2004.

[112] H. Zhong, J. Shi, and M. Visontai, "Detecting unusual activity in video," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Washington, DC, June 2004.

[113] X. Zhu, Z. Ghaharamani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. of Int. Conf. on Machine Learning*, Washington DC, August 2003.

[114] W. P. Ziemer, *Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation*, ser. Graduate Texts in Mathematics. Springer-Verlag, New York, 1989.

[115] J. Zinnes and D. MacKay, "Probabilistic multidimensional scaling: complete and incomplete data," *Psychometrika*, vol. 48, pp. 27–48, 1983.