Geometric Representations of High Dimensional Random Data

by

Sung Jin Hwang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in The University of Michigan
2012

Committee:

Professor Alfred O. Hero III, Co-Chair
Steven B. Damelin, Wayne Country Day School, Co-Chair
Professor Anna C. Gilbert
Associate Professor Clayton D. Scott
Assistant Professor Rajesh Rao Nadakuditi

© Sung Jin Hwang

2012

For Sun Hyo

Acknowledgements

First and foremost I express my sincerest gratitude to my advisor, Professor Alfred Hero, for his support and guidance with patience through many years. My development as a researcher and completion of this thesis would not have been possible without his insights, knowledge, and full support for my research. I thank Dr. Steven Damelin for the discussions we have had over the years and his reviews on my dissertation. I also extend thanks to the other committee members, Professor Anna Gilbert, Professor Clayton Scott, and Professor Rajesh Rao Nadakuditi, for their feedbacks and helps to this dissertation.

I appreciate the experience to work with many exceptional colleagues in Professor Hero's group. I thank Professor Raviv Raich and Mark Kliger for the kind helps when I first started research on machine learning. I learned many topics in statistics from discussions with Professor Ami Wiesel, Yilun Chen, Arnau Tibau Puig, Kumar Sricharan, and Kevin Xu. I also thank Kevin Carter, Yongsheng Huang, Patrick Harrington, Gregory Newstadt, Tzu-Yu Liu, Se-Un Park, Dae-Yon Jung, Hamed Firouzi, Zhaoshi Meng, Ko-Jen Hsiao, Joel LeBlanc, Dennis Wei, and Koby Todros for the valuable discussions.

I enjoyed the time I have spent in Ann Arbor. I thank Gyemin Lee, Jooseuk Kim, Hyun Jeong Cho, Manhee Jeong, Kuang-Hung Liu for all the memories we shared from the beginnings in our Ph.D. studies, and for all the funs we had solving the puzzles. I extend thanks to Jae Young Park, Yoojin Choi, Jung Hyun Bae, and Wonseok Huh. I also express my gratitude to Se Young Chun for his mentorship.

Finally I would like to thank my family for their support. I thank my wife Sun Hyo for her love and support. I would never be able to come this far without her.

Table of Contents

Dedication ii

Acknowledgements iii

List of Figures vii

Chapter 1 Introduction 1

1.1 Short introductions on topics and contributions 2

1.1.1 Power-weighted shortest paths and conformal deformations 2

1.1.2 Information curve comparison 3

1.1.3 Quantum cluster analysis 4

1.2 Presentations, posters, and publications 5

Chapter 2 Shortest Paths and Conformal Deformations 7

2.1 Introduction 7

2.2 Main result 8

2.2.1 Convergence of the length of the shortest path 8

2.3 Proofs 11

2.3.1 Percolation lemma 13

2.3.2 Path refinement 14

2.3.3 Mean convergence for Poisson point processes 17

2.3.4 Shortest path size 22

2.3.5 Mean convergence in i.i.d. cases 24

2.3.6 Concentration of measure 25

2.3.7 Convergence in Riemannian manifolds 32

2.4 Extensions 36

2.4.1 Super-additive dissimilarity 36

2.4.2 Non-compact complete manifolds 41

2.5 Future works . . . 42
2.5.1 Manifolds with boundary . . . 42
2.5.2 Conformal deformations in anisotropic diffusion maps . . . 43

Chapter 3 Information Geometric Curves **46**
3.1 Introduction . . . 46
3.2 Preliminary . . . 46
3.2.1 Information divergence . . . 47
3.2.2 L^p space . . . 52
3.3 Information geometry . . . 55
3.3.1 Parameterization . . . 55
3.3.2 Tangent bundle of parameter space . . . 56
3.3.3 Parallel transport . . . 57
3.4 Parameter space interpolation . . . 62
3.4.1 Interpolation parameter space . . . 64
3.4.2 Transversal vector field . . . 64
3.4.3 Area growth rate calculation from Fisher information . . . 66
3.4.4 Fisher information calculation . . . 67
3.5 Approximations . . . 68
3.5.1 Triangulation approximation . . . 68
3.5.2 Surface energy . . . 70
3.6 Comparison to other curve distances . . . 71
3.6.1 Sensitivity analysis . . . 72
3.7 Simulations . . . 74
3.7.1 Jeffreys' prior . . . 74
3.7.2 Action recognition . . . 77

Chapter 4 Quantum Cluster Analysis **81**
4.1 Introduction . . . 81
4.2 Mixture models and k-means . . . 82
4.2.1 Quantum mechanical generalization . . . 83
4.3 Quantum cluster analysis . . . 84
4.3.1 Quantum mechanical background . . . 84
4.3.2 Quantum states for cluster analysis . . . 85
4.3.3 Mixture models in quantum cluster analysis . . . 87
4.4 State space dimensionality . . . 88
4.5 Quantum state optimization: implementation . . . 90

4.5.1 Stereographic projection 90
4.5.2 Derivatives under stereographic projections 91
4.5.3 Switching stereographic projections 93
4.6 Experiments 94
4.6.1 Three circles dataset 94
4.6.2 Other datasets 96
4.7 Further discussions 100
4.7.1 Connections to spectral clustering 100
4.7.2 Quantum state optimization and couplings 101

Chapter 5 Conclusion **103**

Bibliography **105**

List of Figures

1.1 Given flows of the distributions $\{P_t\}$, $\{Q_t\}$, a surface is built in the statistical manifold. 4

2.1 The power-weighted shortest path runs from $(-2, 1.5)$ to $(2, 1.5)$ through 400 Gaussian sample points. The path is bent toward the central region where the density is high. 9

2.2 A run through the family tree generated by $\mathbb{X}_n$ with $p = 2$. The point x_0 is the ancestor with parameter $r_0 = 9$. This means that all the runs through the family tree are paths with power-weighted length less than $r_0^{1/p} = 3$. Here $x_{1,1} \in \mathbb{X}_1$ is among the first generations since it is within $B(x_0; r_0^{1/p})$, and $x_{2,1} \in \mathbb{X}_2$ is among the second generations since it is within the balls centered at the first generation offsprings, e.g., $x_{1,1}$. This particular run ends at $x_{4,1}$ as there is no point in the vicinity. In this example, the power-weighted path length is $\sqrt{1^2 + 2^2 + 1.5^2 + 1^2} = \sqrt{8.25} < 3$. Note that $x_{2,1}$ is also in the ball centered at x_0, so it is also a first generation offspring. Some other runs through the family tree will have the point $x_{2,1}$ as a first generation offspring. 15

2.3 T_b is the region between the upper and the lower horizontal lines. The shaded region indicates $U = \bigcup_t B(tz; \delta b)$. Because $\delta < 1/4$ and $t \le 3/4$, no point in U may have its second component norm greater than b. Hence $U \subset T_b$. Inside U there is a cylinder of radius δb and length $2^{-1}r$, so the volume of U is at least $V_{d-1}(\delta b)^{d-1}2^{-1}r$. 16

2.4 An illustration of the path paste procedure. A new path from $(0, 0)$ to $(s + t, 0)$ is created by removing $(s, 0) = \gamma_0$ and joining γ_- and γ_+. Only the end points are fixed points in the new path. 18

2.5 $\mathcal{L}_{\lambda v}\big(x_\lambda, y_\lambda; \lambda^{(\alpha-1)/d}|x_\lambda - y_\lambda|\big)$ is the shortest path that is contained in the shaded region. The radii of the circles inside the shaded region are $\lambda^{(\alpha-1)/d}|x_\lambda - y_\lambda|$. If λ is large enough so that $\lambda^{(\alpha-1)/d}|x_\lambda - y_\lambda| \le 2R_1\lambda^{(\alpha-1)/d} \le R_2 - R_1$, then the shaded region must be contained in $B(z; R_2)$ where the intensity is uniform. 21

2.6 All ξ_i's are within $B(z; 4^{-1}R_2)$. Therefore all $B(\xi_i; 5R_2/8)$ are contained in $B(z; R_2)$. Since $|\xi_i - \xi_j| \le 2^{-1}R_2$, with high probability we have $L_n(\xi_i, \xi_j) < L_n(\xi_i; 5R_2/8)$, and $L_n(\xi_i, \xi_j)$ becomes independent of the outside $B(\xi_i; 5R_2/8) \subset B(z; R_2)$ due to the annulus buffer region $\{u: 4^{-1}R_2 < |z-u| < R_2\}$. 30

2.7 Path division procedure described in Proposition 2.19. Here $k = 4$. Note that $z_i \in U_i$ and $z_{i+1} \in V_i$ for $i = 1, 2, 3$. Shortest path is depicted as a smooth curve for illustration purpose only and it is actually piece-wise geodesic. . . 34

3.1 An example when μ and ν are Gaussian distributions in $\Omega = \mathbb{R}$. Radon-Nikodym derivatives with respect to μ are plotted on the right. 54

3.2 Curve γ runs in M. Given a vector field Y in M, the tangent vectors $Y_{\gamma(t)} \in T_{\gamma(t)}M$ and $Y_{\gamma(0)} \in T_{\gamma(0)}M$ are in different vector spaces, and one cannot be subtracted from the other unless we define identifications between the tangent fibers. 58

3.3 In this example, P_θ and $P_{\theta'}$ are Gaussian distributions in $\Omega = \mathbb{R}$ with means at $x = 0$ and $x = 2$. Suppose we have a measure derivative $Y_{\theta'}P_{\theta'}$ around $x = 2$. See the figure on the left. The derivative amplitude is relatively large in the view of P_θ while it is relatively small in the view of $P_{\theta'}$. On the right we compare α-parallel transports for $\alpha = 0, 1$. 60

3.4 α-divergence measures the curvature of the curve from $1 = dP_\theta/dP_\theta$ to $dP_{\theta'}/dP_\theta$ in $L^{1/\alpha}(P_\theta)$. 61

3.5 An example of $M = [0,1] \times I$. Top horizontal line at $s = 0$ is equal to P_t, and bottom horizontal line at $s = 1$ is equal to Q_t. Each vertical line maps to a geodesic curve in the unit sphere of $L^2(P_t)$. 64

3.6 Vertical lines in (a) represent the geodesics used in M. In triangulation approximation, vertical lines are replaced by diagonal lines. See red lines in (b). The red lines on the right side of $t = 0$ represents J_Q where P_t is fixed at P_0 and Q_t proceeds. The red lines on the left side of $t = 0$ represents J_P where Q_t is fixed at Q_0 and P_t proceeds. 69

3.7 Illustrations of $\inf_{t\in I}\|f(x) - g(t)\|$. Red lines denote the nearest neighbor pairs for $f(x)$. Except for case Figure 3.7d, $d_C(g \to f)$ depends on some small portions of f. 74

3.8 Plot of numerical integral J (- - -), triangulation approximation J_T (——), and square root surface energy $\sqrt{E}$ (——) against σ^2. 76

3.9 Plot of numerical integral J (- - -), triangulation approximation J_T (——), and square root surface energy $\sqrt{E}$ (——) against μ. In (a), J is difficult to notice since $\sqrt{E}$ overlaps. The difference of J and $\sqrt{E}$ is shown in (b). . . . 76
3.10 Confusion matrix of PGM data. 5-fold cross-validation. Average over 100 runs. . . . 78
3.11 Sample skeleton images from MSR-Action3D data. . . . 79
3.12 Confusion matrix of MSR-Action3D data. 5-fold cross-validation. Average over 100 runs. . . . 80

4.1 Quantum states are ψ_1 and ψ_2. Any partition operator may applied to the states. . . . 87
4.2 Examples of a stereographic projection. The normalized state variables ψ_1, ψ_2, ψ_3 on $\mathbb{S}^m$ are mapped onto $\xi_1, \xi_2, \xi_3 \in \mathbb{R}^m$, respectively. The focal point is $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^{m+1}$. When ψ_i approaches and is close to the focal point e_1 for some i, the state variable ψ_i is replaced by $-\psi_i$. . . . 91
4.3 Three circles data. (a) is the raw form of the data. (b) is Laplacian Eigenmaps with three eigenvectors. (c) is the Gaussian heat kernel matrix used as the weight matrix for the Laplacian Eigenmaps result in (b). Unlike Figure 4.4b, the cluster structure is not obvious from (c). . . . 95
4.4 Three-dimensional states, $H = \mathbb{R}^3$. Quantum states and association probability matrix of the three circles data. Blue arrows for the first 65 points (inner circle in Figure 4.3a), red arrows for the second 65 points (middle circle in Figure 4.3a), and teal arrows for the last 65 points (outer circle in Figure 4.3a). Unlike Figure 4.3b, the data points from different circles are well separated in terms of their state vectors. (a) shows that the state vectors from different circles are nearly orthogonal. (b) confirms the near orthogonality of the state vectors by the block-diagonal structure of the probability matrix. The diagonal blocks have the association probability $p_{ij} \approx 1$ while the off-diagonal blocks have $p_{ij} \approx 0$. . . . 96
4.5 Two-dimensional states, $H = \mathbb{R}^2$. Quantum states and association probability matrix of the three circles data. Blue arrows for the first 65 points, red arrows for the second 65 points, and teal arrows for the last 65 points. . . . 96

4.6 Cluster analysis comparison for iris dataset from UCI repository. The employed dissimilarity measure is the Euclidean metric. For IRIS dataset, the quantum state optimization performs slightly better than the conventional centroid-based k-means. The data is projected onto the plane by MDS for visualization purpose. 98
4.7 Cluster analysis comparison for Wisconsin diagnostic breast cancer dataset from UCI repository. The employed dissimilarity measure is the Euclidean metric. The quantum state optimization significantly improved the cluster analysis result. The data is projected onto the plane by MDS for visualization purpose. 98
4.8 Quantum cluster analysis results for two-moons data. (a) and (b) are the optimized quantum states and partition result based on Euclidean distance. The Rand index score is 0.5047. (c) and (d) are the optimized quantum states and partition result based on super-additive shortest path distance. The Rand index score is 1, which indicates the perfect partition. 99

Chapter 1
Introduction

Nonlinear data analysis made remarkable progress in the last decade. At the heart of the nonlinear analysis techniques is the local neighborhood structures, and differential geometry arises as a natural foundation to analyze and understand the practice. When random data from a high dimensional space is sampled, a crucial task is to find a data representation that reveals intrinsic lower dimensional structure in the data. This thesis explores data representations using differential geometries that specifically account for the random nature of the data.

Traditional approaches of representations for high-dimensional data are based on linear models. For example, principal component analysis assumes the data variation is concentrated in some linear subspaces. Nonlinear data analysis extends the idea and assumes the data lies in some curved non-flat lower dimensional structure, and more geometric ideas and concepts appear in machine learning studies. ISOMAP (Tenenbaum, de Silva, and Langford 2000) models the data in isometrically embedded Riemannian manifolds, and uses the shortest paths in neighborhood graphs to estimate the geodesic distances. Laplacian Eigenmaps (Belkin and Niyogi 2003) and diffusion maps (Coifman and Lafon 2006) deploy diffusion processes in manifolds to represent the data by eigenfunctions of the generating operators.

A main focus of this thesis is to extend the geometric ideas introduced in previous researches, and integrate statistical analysis and geometric analysis in a unified framework. When sample points are assumed to follow some unknown probability distribution, one of the main challenges is to design a metric or proximity measure which incorporates the statistical aspects of the underlying distribution. Theoretically or ideally one wants to make decisions based on statistical arguments such as Neyman-Pearson lemma or posterior probabilities. On the other hand, practical procedures based on finite samples rely on geometric tools such as nearest neighbor structures or inner products. Machine learning theories should find a balance in these two aspects. k nearest-neighbor classifier serves as

a good example. The justification of the classifier is provided by a statistical theory and its consistency condition is based on density estimation arguments. On the other hand, the algorithm is distance-based, and its behavior for finite samples depends on geometric information of the samples.

Machine learning researchers have developed algorithms for classification and clustering using combinations of deterministic and probabilistic tools of analysis. Deterministic approaches exploit geometric properties of the data, e.g., smoothness of the data manifold using inter-feature distances in ISOMAP, while probabilistic tools exploit statistical models of the data, e.g., assuming that data is generated from an unknown mixture of probability densities. Few approaches combine the strengths of statistical and geometric approaches to design machine learning algorithms. For example, the Gaussian mixture model assumes that the data is generated by some mixture probability distributions, and the optimal parameters are estimated by algorithms like expectation-maximization. This clustering algorithm is purely based on a statistical model and does not explicitly incorporate geometry into the design of the algorithm. On the other hand, the k-means clustering algorithm is motivated by a purely geometric argument yet it is equivalent to the Gaussian mixture model under some conditions. Afterwards, some geometric interpretations are provided such as a variation of the k-means with soft assignments. Thus the two approaches are not mutually exclusive but rather are dual to each other. For instance, Laplacian Eigenmaps put its emphasis on geometric aspects and the problem is formulated as a gradient minimization. On the other hand, diffusion maps put its emphasis on probabilistic aspects and the method analyzes random walks over the samples.

One of the main purposes of this thesis is to pioneer and investigate examples which consolidate both approaches. For such purpose, this thesis studies conformally deformed geometry in Chapter 2 to illustrate how Riemannian geometry may be used to combine both geometry and statistics. In Chapter 3 the thesis studies the space of probability measures with infinite-dimensional geometry, and connects information theory, functional analysis, and differential geometry. In Chapter 4, we propose a cluster analysis framework to define a random walk based on manifold metrics.

1.1 Short introductions on topics and contributions

1.1.1 Power-weighted shortest paths and conformal deformations

Geodesic curves are essential geometric objects in Riemannian geometry. Their lengths are direct extensions of Euclidean distances, hence geodesic distance estimation in manifold

learning is as fundamental as Euclidean distance computation is in Euclidean geometry. ISOMAP (Tenenbaum *et al.* 2000; Bernstein, de Silva, Langford, and Tenenbaum 2000) shows that the shortest path lengths through random sample points converge to geodesic distances when the sample points are in an isometrically embedded manifold. However, as noted in Costa and Hero (2004), the shortest path lengths do not reflect the probability distribution of the sample.

In Chapter 2, we investigate power-weighted shortest paths through random points in Riemannian manifolds. The main difference from ISOMAP is that each graph edge weight is raised to power $p > 1$. The contribution of the chapter is to prove that the power-weighted shortest path lengths completely converge to a new Riemannian distance on the manifold conformally deformed by the underlying probability density function.

This convergence result is an extension of the Beardwood-Halton-Hammersley theorem (Beardwood, Halton, and Hammersley 1959; Yukich 1998) in Euclidean random graph theory. We discuss how these conformal deformations can be used in machine learning, and how the theory provides a new interpretation of spectral methods including anisotropic diffusion maps (Coifman and Lafon 2006).

1.1.2 Information curve comparison

The relationship between probability and geometry may be studied from a different perspective. The collections of the probability measures over a fixed measurable space forms a space of measures. This measure space is called a statistical manifold, and its connection to Riemannian theory has been observed for a long time through Fisher information.

In this statistical manifold the concept of entropy and divergence play many important roles by means of dissimilarity measures between certain classes of probability distributions (Cover and Thomas 2006). For example, Shannon entropy (Shannon 1948) and Kullback-Leibler divergence (Kullback and Leibler 1951) have many crucial applications in coding theories and estimation theories. Rényi entropy (Rényi 1961) generalizes Shannon entropy and it is used in e.g., vector quantizations and binary detection problems. Tsallis entropy (Tsallis 1988) is another type of information divergence which is useful in thermodynamics of non-extensive systems.

Information divergence is essentially a distance-like quantity between measures, and entropy is information divergence from some reference measure. Consider now that the probability measures are parameterized. For example, suppose that we have probability measures which vary over time. Information divergence can tell us how close and how far two chosen *snap shots* of the time-varying probability measure are. It does not, however, tells

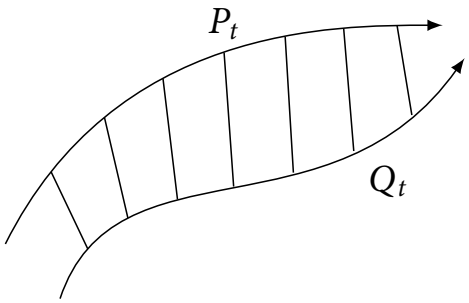


Figure 1.1: Given flows of the distributions $\{P_t\}$, $\{Q_t\}$, a surface is built in the statistical manifold.

us how far the overall parameterization is from some probability measure, or more generally, from another parameterization. The simplest way to tackle this problem would be to add up the divergence quantity over time. This approach potentially suffers from inadequate parameterization. To illustrate, suppose that we have a continuous-time Markov system, e.g., thermodynamic system, and that we have two initial probability distributions P_0, Q_0 of the system. The information divergence between the initial distributions measures a certain dissimilarity but it does not take the Markov system into consideration. A better approach is to measure the dissimilarity between the induced flow of the distribution $\{P_t\}$, $\{Q_t\}$ from P_0, Q_0, respectively, as in Rached, Alajaji, and Campbell (2001) and Rached, Alajaji, and Campbell (2004).

There exist various approaches to the above mentioned method. We propose in Chapter 3 an approach that defines a distance-like quantity that is invariant over reparameterizations and is geometrically appealing. Invariance over reparameterizations appears naturally in geometry. For example, the length of a curve is defined in a way that it does not depend on the parameterization but only on the image of the parameterization. Our idea is essentially to consider information geometry, and create a two-dimensional parameterization, or a surface, which has the one-dimensional parameterizations as boundary. Then we may use the Riemannian structure to calculate the surface area induced from two parameterizations. See Figure 1.1 for a visualization.

1.1.3 Quantum cluster analysis

As outlined above, Chapter 2 and Chapter 3 study dissimilarities within random samples. In Chapter 4, we attempt to establish a connection between the dissimilarities and spectral methods in a cluster analysis framework. Specifically, we propose an algorithm to transform dissimilarities into similarities.

Cluster analysis or data clustering is to partition data into subsets so that the data points in each subset share some common properties. Spectral methods are strongly related to spectral clustering (Shi and Malik 2000; Ng, Jordan, and Weiss 2002; von Luxburg 2007), and it is not surprising that a cluster analysis framework is used to establish a connection with spectral methods.

Cluster analysis holds practical importance as well. Due to the technological advances of computer systems, the volume of data is constantly increasing and the dimensionality of data is also increasing. However, more variables and more features in data do not directly translate into more information unless adequate processing and interpretations of the data follow. For instance, the internet is flooded with a huge amount of information but human users may hardly benefit from it without proper knowledge of addresses or uses of internet search engines. Cluster analysis aims to find subsets of data so that the user may discover new structures in the data and to split the problem to adopt divide-and-conquer approaches.

In Chapter 4, we propose *quantum cluster framework* motivated by quantum mechanics, and quantum state optimization based on the k-means. This model of cluster analysis uses quantum states, i.e., linear subspaces of some Hilbert space, which is called the state space. We discuss why quantum states model and its use of projective geometry would be beneficial to cluster analysis.

The quantum state optimization transforms dissimilarities into similarities. This is a dimensionality reduction method based on gradient descent optimization in spheres. We discuss and explore the issues that appear when a dimensionality reduction method targets bounded spaces such as spheres or projective spaces rather than unbounded Euclidean spaces. Another important aspect of the quantum state optimization is that it learns a transition system over the sample points. The optimization establishes a similarity learning like semidefinite embedding (Weinberger, Sha, and Saul 2004) and complements the dissimilarity-based methods presented in the other chapters.

1.2 Presentations, posters, and publications

Sung Jin Hwang, Steven B. Damelin, and Alfred O. Hero III (2012). "Shortest path through random points." arXiv: 1202.0045 [math.PR].

Sung Jin Hwang, Steven B. Damelin, and Alfred O. Hero III (2012). "Shortest path for high-dimensional data representation." Poster presentation at: *2012 SIAM Annual Meeting*. (Minneapolis, Minnesota, USA. July 9–13, 2012.)
URL: http://meetings.siam.org/sess/dsp_talk.cfm?p=51935.

Sung Jin Hwang, Steven B. Damelin, and Alfred O. Hero III (2011). "Surface energy for information geometric curve comparison." Poster presentation at: *The Fifth Michigan Student Symposium for Interdisciplinary Statistical Sciences*. (Ann Arbor, Michigan, USA. April 8, 2011.) URL: http://sitemaker.umich.edu/mssiss/mssiss_2011.

Sung Jin Hwang, Steven B. Damelin, and Alfred O. Hero (2010). "Comparing information geometric curves." Talk at: *13th International Conference on Approximation Theory*. (San Antonio, Texas, USA. March 7–10, 2010.)
URL: http://www.math.vanderbilt.edu/~at13/minisymposia.html.

Sung Jin Hwang and Alfred Hero (2009). "Geometric optimization in probability density space." Poster presentation at: *Engineering Graduate Symposium*. (Ann Arbor, Michigan, USA. November 13, 2009.)
URL: http://www.eecs.umich.edu/eecs/graduate/symposium.html.

Chapter 2
Shortest Paths and Conformal Deformations

2.1 Introduction

The shortest path problem is of interest both in theory and in applications since it naturally arises in combinatorial optimization problems, such as optimal routing in communication networks. It also draws attentions for practical reasons, since efficient algorithms—Dijkstra's, Bellman-Ford, or Floyd-Warshall—exist to solve the problem (Cormen, Leiserson, Rivest, and Stein 2009). In this chapter, we are interested in the shortest paths over random sample points in Euclidean and Riemannian spaces.

Many graph structures over Euclidean sample points have been studied in the context of the Beardwood-Halton-Hammersley (BHH) theorem and its extensions. The BHH theorem states that law of large numbers holds for certain spanning graphs over random samples. Such graph structures include the travelling salesman path (Steele 1981), the minimal spanning tree (Yukich 2000), and the minimal matching graphs (Rhee 1993). For thorough details, we refer the reader to Steele (1997) and Yukich (1998). The theorem applies to graphs that span all of the points in the random sample. This chapter establishes a BHH-type theorem for power-weighted shortest paths between any two points.

In the last few years, the BHH theorem for spanning graphs such as the minimal spanning tree (MST) or the traveling salesman path (TSP) has been extended to Riemannian case, e.g., Costa and Hero (2004) extended the MST in the context of entropy and intrinsic dimensionality estimation. More general non-Euclidean extensions have been established by Penrose and Yukich (2011). This chapter extends the BHH theorem in a different direction, i.e., the power-weighted shortest path between points in a Riemannian manifold.

The shortest path length convergence results are provided in several contexts. Theorem 2.1 claims the convergence in compact manifolds, and Section 2.3 provides the proofs

in detail. Several extensions of Theorem 2.1 are presented in Section 2.4. We introduce the concept of super-additive shortest paths which generalizes the power-weighted shortest paths. The super-additive shortest paths provide the convergence results for sample points in embedded manifolds as in ISOMAP (Tenenbaum *et al.* 2000), and for sample points in statistical manifolds as in FINE (Carter, Raich, and Hero 2009a). We also extend the results into the domains of complete manifolds rather than compact manifolds.

2.2 Main result

Let (M, g_1) be a smooth compact d-dimensional Riemannian manifold without boundary. Always assume that $d > 1$. The use of the subscript for g_1 will become clear shortly.

Consider a probability distribution Pr over Borel subsets of M. Assume that the distribution has a smooth Lebesgue probability density function f with respect to g_1. Let $X_1, X_2, \ldots$ denote an i.i.d. sequence drawn from this density.

For $p > 1$, called the power parameter, define a new conformal Riemannian metric $g_p = f^{2(1-p)/d} g_1$. That is, if Z_x and W_x are two tangent vectors at a point $x \in M$, then $g_p(Z_x, W_x) = f(x)^{2(1-p)/d} g_1(Z_x, W_x)$.

The main result of this chapter, stated as Theorem 2.1, establishes an asymptotic limit of the lengths of the shortest paths through finite subsets of points $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ as $n \to \infty$. If $x, y \in M$, then $L_n(x, y)$ denotes the shortest path length from x to y through $\mathcal{X}_n \cup \{x, y\}$. Here the edge weight between two points u and v is $\mathrm{dist}_1(u, v)^p$ where dist_1 denotes the Riemannian distance under g_1. The power weighted graph (PWG) is defined as the complete graph over $\mathcal{X} \cup \{x, y\}$.

2.2.1 Convergence of the length of the shortest path

The following states the main result of this chapter.

Theorem 2.1. *Assume that* $\inf_M f > 0$. *Let* $x, y \in M$. *Then*

$$\lim_{n\to\infty} n^{(p-1)/d} L_n(x, y) = C(d, p)\, \mathrm{dist}_p(x, y) \quad \text{c.c.}, \tag{2.1}$$

where dist_p *denotes the Riemannian distance under* g_p, *and c.c. stands for complete convergence.*[1] $C(d, p)$ *is an explicit positive constant given in Lemma 2.8 that only depends on* d *and* p.

[1] Recall that a sequence of random variables, say Y_n, converges completely (c.c.) to another random variable, say Z, if for every $\varepsilon > 0$, the infinite sum $\sum_n \Pr\{|Y_n - Z| > \varepsilon\}$ is finite. Complete convergence implies almost sure convergence by the Borel-Cantelli lemma.

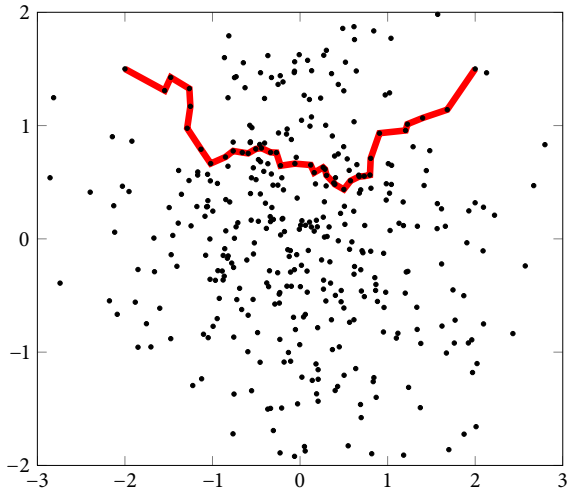


Figure 2.1: The power-weighted shortest path runs from $(-2, 1.5)$ to $(2, 1.5)$ through 400 Gaussian sample points. The path is bent toward the central region where the density is high.

Since the Riemannian metric g_p is g_1 scaled by $f^{2(1-p)/d}$, and this scaling is inversely proportional to the probability density function f, the theorem says that the density f has the effect of shortening or lengthening paths that pass through high density regions or low density regions, respectively. See Figure 2.1 for an example.

The induced distance dist_p has desirable properties for applications such as clustering and dimensionality reduction. In such applications the distance, called dissimilarity, between feature points has a central role. There are many properties that a useful dissimilarity should satisfy (Belkin and Niyogi 2003; Coifman and Lafon 2006). One of them is the density mode separation property: if one must pass through low density region in order to move from x to y, then the dissimilarity between x and y must be large. An electrical circuits analogy is that if the vertices and edges of the complete graph are circuit nodes and meshes, respectively, then more current would flow along edges with high conductance between the endpoints. As conductance increases with free electron density f the analogy is complete.

Another important implication of Theorem 2.1 is that $L_n(x, y)$ shrinks to zero as n tends to infinity so that every edge in L_n must have small length. This property makes the shortest paths in the PWG favorable for manifold analysis. Suppose that M is isometrically embedded in some Euclidean space. The ratio between the Riemannian distance under the intrinsic Riemannian metric and the Euclidean distance in the ambient space uniformly approaches one as either distance tends to zero (since M is compact). Therefore the ratio between the shortest path lengths in both distances converges to one as n tends to infinity, and Theorem 2.1 is still valid when graph edge lengths are Euclidean distances in the ambient

space instead of Riemannian distances.

It is important for certain applications to note that graph edge lengths need not be Euclidean distance but any smooth dissimilarity will do as long as its difference from the Riemannian distance becomes negligible in small neighborhoods. A notable application is the use of Bregman divergence in statistical manifolds (Amari and Nagaoka 2000; Banerjee, Merugu, Dhillon, and Ghosh 2005). Information divergence such as Kullback-Leibler or Bregman divergence violates the triangle inequality, and this makes geometric interpretation difficult except for a few special cases. Theorem 2.1 gives the insight that the sum of the Bregman divergences over shortest paths can be viewed as an estimate of Fisher information distance deformed by the prior distribution. This intuition is proved in Section 2.4.

From the traveling salesman paths to the shortest paths

The convergence result established in this chapter differs from the previous BHH-type theorems in two ways. The first difference is that the graph considered in our theorem, the shortest path, is not a regular graph.[2] All the graphs that have been considered so far are either regular or almost regular. For example, the nearest-neighbor graph and the minimal matching graph are regular. The TSP and the MST are regular except for a single node. The second difference is that the shortest path has fixed anchor points, hence it is not translation-invariant. This is in contrast to BHH theory developed in Penrose and Yukich (2003) and Penrose and Yukich (2011) where Euclidean functionals are generalized to locally stable functionals while translation-invariance requirement is maintained.

The essence of the BHH theorem is that the law of large numbers (LLN) holds for certain graph lengths over random sample. The BHH theorem and its extensions, may be used to estimate information of the probability distributions and the underlying spaces of the data (Hero, Ma, Michel, and Gorman 2002; Costa and Hero 2004; Hero 2007). The information provided such as entropy of the probability measure, is averaged over the space since the theorem is a variation of the LLN applied for certain graphs. On the contrary, many applications such as classification or clustering, require quantitative relationships between points rather than average information over the whole sample. Therefore, one may ask if such a convergence theorem exists for graphs other than spanning ones, such as shortest path.

Some observations of the previous works already suggest that a similar result holds for the shortest path case. The first observation comes from ISOMAP (Tenenbaum *et al.* 2000),

[2]Recall that a directed graph is regular if every node has the same number of incident edges. The MST is usually considered to be undirected, but it may be transformed into a directed graph by, e.g., the Prim's algorithm.

which demonstrates that the shortest path length is an estimator of Riemannian distance. The second observation is that when we begin to adopt Riemannian geometry, the PWG reflects a conformal deformation in Riemannian metric. To explain the second observation, note that the BHH theorem holds for subgraphs in power-weighted graph lengths (Steele 1988; Yukich 2000). The theorem states that when normalized properly, the power-weighted TSP or MST length converges to some constant multiplied by

$$\int f(x)^{(d-p)/d}\, dx \tag{2.2}$$

where f is a compactly supported probability density function, d is the dimension of the Euclidean space, and the integral is over $\mathbb{R}^d$. Suppose that the standard Riemannian metric is deformed and the volume form dx is replaced with $f^{1-p}dx$. Then, the new probability density function is f^p instead of f. If the new deformed Riemannian metric is used to compute the graph length, without the power-weight effect, the graph lengths converge to some constant multiplied by

$$\int \left(f(x)^p\right)^{(d-1)/d} f(x)^{1-p}\, dx = \int f(x)^{(d-p)/d}\, dx, \tag{2.3}$$

and the same integral is obtained. Roughly speaking, the PWG effect is equivalent to Riemannian metric deformation by f^{1-p}, as long as graph length is concerned. These observations lead us to conjecture that the power-weighted shortest path length converges to Riemannian distance deformed by f^{1-p}, and Theorem 2.1 claims it to be true.

2.3 Proofs

First we introduce notations and conventions for the proofs. Some notations are already defined in the previous section, but since the proofs work in Euclidean spaces until the very last part, it will be convenient to re-introduce notations for Euclidean cases.

The i.i.d. samples $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ will be assumed to be in $\mathbb{R}^d$, $d > 1$ until Section 2.3.7. For convenience let $\mathcal{X}_n = \{X_1, X_2, \ldots, X_n\}$. We assume that the reference metric g_1 in the previous section is the Euclidean metric so that the power-weighted edge weight between points $u, v \in \mathbb{R}^d$ is $|u - v|^p$, where $|\cdot|$ denotes the Euclidean norm.

If $x \in \mathbb{R}^d$ and $r > 0$, then $B(x; r)$ will denote the open ball in $\mathbb{R}^d$ of radius r, centered at x. When we write $(x_1, x_2) \in \mathbb{R}^d$, the convention will be that $x_1 \in \mathbb{R}$ and $x_2 \in \mathbb{R}^{d-1}$.

The number of nodes in the shortest paths is important in the proofs. We abuse the notation and $|L_n| = |L_n(x, y)|$ will denote the number of nodes in the shortest path corresponding to $L_n = L_n(x, y)$. Since the shortest path length is never negative, there will be no

ambiguity.

The Poisson point process will play a central role in the proofs. We adopt the definitions and developments introduced in Baddeley (2007). If λ is positive then $\mathcal{H}_\lambda$ denotes the homogeneous Poisson point process in $\mathbb{R}^d$ of intensity λ.

The proofs involve establishing asymptotic properties of the shortest paths through the points in $\mathcal{H}_\lambda$. To avoid ambiguity, define $\mathcal{L}_\lambda(x, y)$ to be the shortest path length from x to y in $\mathcal{H}_\lambda$. Often x and y will have the form $(s, 0)$ and $(t, 0)$ for some $s, t \in \mathbb{R}$. In that case, we write $\mathcal{L}_\lambda(s, t)$ instead of $\mathcal{L}_\lambda(x, y)$.

We often restrict shortest paths to neighborhoods of straight line segments between end points. Define $\mathcal{L}_\lambda(x, y; R)$ for $0 < R \leq \infty$ to be the shortest path length in the PWG from x to y within the region $\bigcup_u B(u; R)$ where u is on the straight line segment between x and y. We define $\mathcal{L}_\lambda(s, t; R)$ similarly.

A brief outline of the proof is as follows. First note that the theory is restricted to the case that f is a uniform, or locally uniform, density function until Section 2.3.7.

Corollary 2.9 establishes mean convergence of the shortest path lengths for the case of a homogeneous Poisson point process supported in a finite ball. The mean convergence result is de-Poissonized to i.i.d. samples in Proposition 2.12. To aid this procedure, Lemma 2.11 uses a percolation argument (Lemma 2.4) to bound the number of nodes in the shortest path. Lemma 2.11 also plays an important role in Proposition 2.13 where Talagrand's concentration inequality shows complete convergence over sets of uniform sample points. The theory is generalized to non-uniform densities in Theorem 2.16, which applies Proposition 2.13 multiple times to show that complete convergence holds in a ball where the density is locally uniform. Theorem 2.16 shows that there exists buffer zone which isolates the local shortest paths from the probability distribution outside the buffer zone.

In Section 2.3.7, the locally uniform density condition is relaxed to a smooth density condition. Lemma 2.18 shows that the normalized shortest path lengths measured by Riemannian distance are close to the lengths measured by Euclidean distance in sufficiently small normal charts. Proposition 2.19 shows that if the local convergence Lemma 2.18 holds in some finite open cover, then global shortest path length L_n converges completely by near sub- and super-additivity of L_n. Proposition 2.19 implies Theorem 2.1.

Remark 2.2. The shortest path length $\mathcal{L}_\lambda$ satisfies two important properties that will be used in the proofs. Firstly, $\mathcal{L}_\lambda(x, y; R)$ is monotonically non-increasing in both λ and R. Adding more paths from x to y cannot increase the minimum path length. Secondly, $E\mathcal{L}_\lambda(x, y; R) = \lambda^{-p/d} E\mathcal{L}_1(\lambda^{1/d}x, \lambda^{1/d}y; \lambda^{1/d}R)$. This follows from the facts that a homogeneous Poisson point process $\mathcal{H}_\lambda$ may be scaled by a factor $a > 0$, and yield

another Poisson point process $\mathcal{H}_{a^{-d}\lambda}$, and that $\mathcal{L}_\lambda(x, y; R)$ is a sum of Euclidean distances raised to a p-th power.

Remark 2.3. In general, the endpoints x, y will be indexed by the number n of points, or the mean number λ of points when $\mathcal{X}_n$ are points of a Poisson process. Most of the lemmas and propositions include conditions of the form $\liminf_n n^\alpha |x_n - y_n| = +\infty$ for some $\alpha > 0$, to prevent the end points from approaching each other too fast. Such conditions are essential to the lemmas and propositions. To see why they are essential, note that the normalization $n^{(p-1)/d}$ in (2.1) increases in n, unlike in the standard BHH theorem normalization $n^{(p-d)/d}$ for $0 < p < d$ or the usual LLN normalization n^{-1}. Theorem 2.1, like many other graph-related theorems, approximates arbitrary shortest paths by the shortest paths between some finite discrete points. Since the normalization increases in n, the accuracy of the path approximation must improve as n increases to keep the same level of error in the overall path lengths. Therefore the points must be spread in the space more densely as n increases. On the other hand, if the points are too close to each other then the convergence in (2.1) will not hold since the nearest neighbor distances shrink at rate $n^{-1/d}$. To summarize, the proofs must be careful to assure that the convergence holds for point pairs close enough but at the same time, not too close.

2.3.1 Percolation lemma

We start with a few lemmas which will be useful in the main proofs.

Consider the point $0 \in \mathbb{R}^d$ and a homogeneous Poisson point process $\mathcal{H}_\lambda$. Recall that the probability that the k-th nearest neighbor distance from 0 to any point in $\mathcal{H}_\lambda$ to be less than or equal to $u > 0$ is proportional to u^d. The following lemma shows that the rate becomes exponential when considering multiple neighborhoods jointly.

Lemma 2.4. *Let π be a graph path in a homogeneous Poisson point process $\mathcal{H}_\lambda$ starting at $0 \in \mathbb{R}^d$. Suppose that π has power-weighted path length at most $c_0\lambda^{(1-p)/d}$ and has at least $c_1\lambda^{1/d}$ nodes for some positive c_0, c_1. Then there exists a constant $\rho_0 > 0$ such that if $c_1 > \rho_0 c_0$, the probability that such path π exists is exponentially small in $c_1\lambda^{1/d}$.*

Proof. The strategy of this proof is similar to that of Meester and Roy (1996, Theorem 6.1).

We first define a Galton-Watson process $\mathbb{X}_n$. Let $\mathbb{X}_0 = \{x_0 = 0 \in \mathbb{R}^d\}$ be the ancestor of the family, and associate the parameter $r_0 > 0$. Then define the offsprings $\mathbb{X}_1(r_0)$ to be $\mathcal{H}_\lambda \cap B(x_0; r_0^{1/p})$. $\mathbb{X}_1(r_0)$ is the set of points in $\mathcal{H}_\lambda$ that may be reached from x_0 with exactly single edge in the PWG, path length at most r_0. Note that $E|\mathbb{X}_1(r_0)| = \lambda V_d r_0^{d/p}$ where $|\mathbb{X}_1(r_0)|$ denote the cardinality of $\mathbb{X}_1(r_0)$, and V_d denotes the volume of $B(0;1)$.

For each offspring $x_{1,k} \in \mathbb{X}_1(r_0)$, we associate the parameter $r_{1,k} = r_0 - |x_{1,k} - x_0|^p$. Then $\mathcal{H}_\lambda$ in the union of $B(x_{1,k}; r_{1,k}^{1/p}) - \{x_{1,k}, x_0\}$ over k is the set of points that may be reached from x_0 with exactly two edges, while the power-weighted path length is at most r_0. Note that x_0 is discarded to prevent loops. Define $\mathbb{X}_2(r_0)$ to be the collection of all the second generation offsprings, then

$$\left|\mathbb{X}_2(r_0)\right| = \Big| \bigcup_{x_{1,k} \in \mathbb{X}_1} \mathcal{H}_\lambda \cap \left(B(x_{1,k}; r_{1,k}^{1/p}) - \{x_{1,k}, x_0\}\right)\Big| \tag{2.4}$$

$$\leq \sum_{x_{1,k} \in \mathbb{X}_1} \left|\mathcal{H}_\lambda \cap \left(B(x_{1,k}; r_{1,k}^{1/p}) - \{x_{1,k}\}\right)\right|. \tag{2.5}$$

Define recursively the n-th generation offsprings $\mathbb{X}_n(r_0)$. Then $\mathbb{X}_n(r_0)$ is the set of all the points that may be reached in n hops from the ancestor x_0 within path length r_0. See Figure 2.2. Apply the Campbell-Mecke formula recursively, (Baddeley 2007, Theorem 3.2, p. 48)

$$E\left|\mathbb{X}_n(r_0)\right| \leq \lambda \int_{B(x; r_0^{1/p})} E\left|\mathbb{X}_{n-1}\left(r_0 - |x - x_0|^p\right)\right| dx \tag{2.6}$$

$$= \left(\lambda V_d r_0^{d/p}\right)^n \frac{\Gamma(1 + d/p)^n}{\Gamma(1 + nd/p)}. \tag{2.7}$$

If a path starting at x has more than $n = c_1 \lambda^{1/d}$ nodes and has path length less than $r_0 = c_0 \lambda^{(1-p)/d}$, then the n-th generation set $\mathbb{X}_n(r_0)$ must not be empty. The survival probability is bounded above using the Markov's inequality, and by the Stirling's approximation

$$\log E\left|\mathbb{X}_n(r_0)\right| \leq n \log\left(V_d \Gamma\left(1 + \frac{d}{p}\right)\left(\frac{c_0}{c_1} \cdot \frac{pe}{d}\right)^{d/p}\right) + o(n) \tag{2.8}$$

as $n \to \infty$. The claim follows since, if the ratio c_1/c_0 is sufficiently large, then the logarithm becomes negative. In that case, the survival probability is exponentially small in $n = c_1 \lambda^{1/d}$. □

2.3.2 Path refinement

Let $x, y \in \mathbb{R}^d$. Define a function $\beta_{x,y} \colon \mathbb{R}^d \to \mathbb{R}$,

$$\beta_{x,y}(u) = |x - u|^p + |y - u|^p - |x - y|^p. \tag{2.9}$$

If $\beta_{x,y}(u) < 0$, then the path length $L(x \to u \to y)$ is less than $L(x \to y)$, and hence if x, y, u are nodes in the PWG, then $x \to y$ cannot appear in any shortest path, as it may be replaced

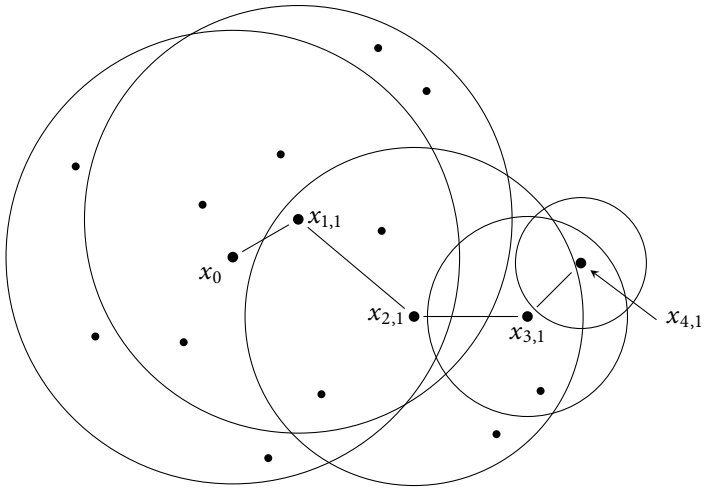


Figure 2.2: A run through the family tree generated by $\mathbb{X}_n$ with $p = 2$. The point x_0 is the ancestor with parameter $r_0 = 9$. This means that all the runs through the family tree are paths with power-weighted length less than $r_0^{1/p} = 3$. Here $x_{1,1} \in \mathbb{X}_1$ is among the first generations since it is within $B(x_0; r_0^{1/p})$, and $x_{2,1} \in \mathbb{X}_2$ is among the second generations since it is within the balls centered at the first generation offsprings, e.g., $x_{1,1}$. This particular run ends at $x_{4,1}$ as there is no point in the vicinity. In this example, the power-weighted path length is $\sqrt{1^2 + 2^2 + 1.5^2 + 1^2} = \sqrt{8.25} < 3$. Note that $x_{2,1}$ is also in the ball centered at x_0, so it is also a first generation offspring. Some other runs through the family tree will have the point $x_{2,1}$ as a first generation offspring.

with $x \to u \to y$.

Define the set

$$\Theta(x, y) = \left\{u \in \mathbb{R}^d : \beta_{x,y}(u) < 0\right\}. \tag{2.10}$$

It is clear that the function β is invariant to rotations, and that its sign is invariant to the scaling. Then there exists $\theta_0 > 0$ such that the volume of $\Theta(x, y)$ is

$$|\Theta_{x,y}| = \theta_0 |x - y|^d. \tag{2.11}$$

Lemma 2.5. *Consider the collection of points z such that $\mathcal{H}_\lambda$ is empty in $\Theta(0, z)$. Define $\xi = \sup|z|$. Similarly define ξ_b for $b > 0$ where both z and $\mathcal{H}_\lambda$ are restricted to the tubular region $T_b = \bigcup_u B((u, 0); b)$ over $u \in \mathbb{R}$. Then for $p > 1$,*

$$\lim_{b \to \infty} E\xi_b^p = E\xi^p \tag{2.12}$$

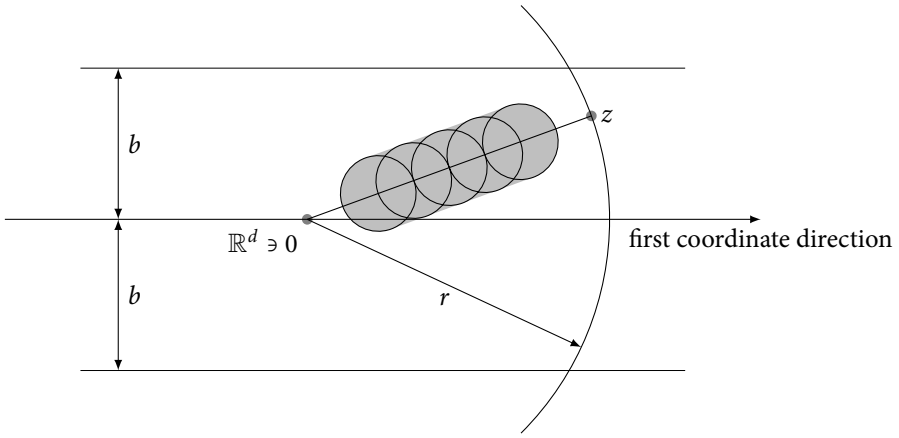


Figure 2.3: T_b is the region between the upper and the lower horizontal lines. The shaded region indicates $U = \bigcup_t B(tz; \delta b)$. Because $\delta < 1/4$ and $t \le 3/4$, no point in U may have its second component norm greater than b. Hence $U \subset T_b$. Inside U there is a cylinder of radius δb and length $2^{-1}r$, so the volume of U is at least $V_{d-1}(\delta b)^{d-1}2^{-1}r$.

and

$$\lim_{\lambda^{1/d}b \to \infty} \frac{E\xi_b^p}{\lambda^{(1-p)/d} b} = 0. \tag{2.13}$$

Proof. If $z \in \mathbb{R}^d$, then $\Theta(0, z) \subset \Theta(0, tz)$ for all $t \ge 1$. Therefore if $\xi > r$ then there exists z with $|z| = r$ such that $\Theta(0, z)$ is empty. The unit sphere---the boundary of $B(0;1)$---is compact, so a finite open cover $V_1, \ldots, V_n$ of the unit sphere and open subsets $U_1, \ldots, U_n$ of $\mathbb{R}^d$ may be selected such that $\beta_{0,z}(y) < 0$ for all $y \in U_k$, $z \in V_k$, for each $k = 1, \ldots, n$. Then for $\xi > r$, $\mathcal{H}_\lambda$ must contain no point in at least one of $rU_1, \ldots, rU_n$. Therefore $\Pr\{\xi > r\}$ is bounded above by $n \exp\bigl(-\lambda A r^d\bigr)$, where A is the minimum volume of $U_1, \ldots, U_n$.

For ξ_b, suppose that $|z| = r > 0$. There exists $0 < \delta < 1/4$ such that if $V = B(z; \delta b)$ and $U = \bigcup_t B(tz; \delta b)$ for $1/4 \le t \le 3/4$, then $\beta_{0,z}(y) < 0$ for all $y \in U$, $z \in V$. Simple calculation shows that we may choose δ that works for all r. The intersection of a sphere and T_b is relatively compact, hence there exists finite open cover $V_1, \ldots, V_m$ and corresponding $U_1, \ldots, U_m$. By the same argument for the ξ, if $r > b$, $\Pr\{\xi_b > r\} \le m \exp\bigl(-\lambda(\delta b)^{d-1}2^{-1}r\bigr)$. After an integration,

$$E\xi_b^p \le \frac{n\Gamma(1 + p/d)}{(\lambda A)^{p/d}} + \frac{m2^p\Gamma(1 + p)}{\lambda^p(\delta b)^{p(d-1)}}. \tag{2.14}$$

(2.13) follows by a direct substitution. The other claim (2.12) follows from an application of the dominated convergence theorem. □

The next lemma states similar result for the i.i.d. case. As before, let $z \in \mathbb{R}^d$, $R_2 > 0$. Assume that the density f is uniform in $B(z; R_2)$ with $f(u) = f_0 > 0$ for $u \in B(z; R_2)$.

Lemma 2.6. *Fix n and $0 < \alpha < 1$. Define the event $E_F(i, j)$ for each pair $1 \le i, j \le n$ such that $E_F(i, j)$ does* not *occur if and only if (i) both X_i and X_j are in $B(z; R_2)$, (ii) $|X_i - X_j| > (nf_0)^{(\alpha-1)/d}$, and (iii) the shortest path from X_i to X_j contains no sample point X_k other than X_i and X_j. Let $E_F = \bigcap_{i,j} E_F(i, j)$. Then*

$$\limsup_{n\to\infty} \frac{1}{(nf_0)^\alpha} \log\bigl(1 - \Pr(E_F)\bigr) \le -\theta_1 \tag{2.15}$$

for some constant $\theta_1 > 0$ which depends only on d and p.

Proof. If X_1 and X_2 are in $B(z; R_2)$, then it is not difficult to show that a certain proportion of $\Theta(X_1, X_2)$ intersects with $B(z; R_2)$, i.e., there exists $\theta_1 > 0$ which depends only on the shape of Θ such that the volume of the intersection is at least $\theta_1 |X_1 - X_2|^d$. Suppose that $E_F(1, 2)$ does not occur. Then the shortest path from X_1 to X_2 contains no sample point other than X_1 and X_2, and the intersection of $\Theta(X_1, X_2)$ and $B(z; R_2)$ must not contain any of $X_3, X_4, \ldots, X_n$. Since it is assumed that $|X_1 - X_2| > (nf_0)^{(\alpha-1)/d}$, the probability that E_F does not occur is at most $(1 - f_0\theta_1(nf_0)^{\alpha-1})^{n-2}$. Since there are $n(n-1)/2 \le n^2$ pairs of sample points, $1 - \Pr(E_F) \le n^2(1 - \theta_1 f_0^\alpha n^{\alpha-1})^{n-2}$ and the claim follows. □

2.3.3 Mean convergence for Poisson point processes

We begin the main proof with Poisson point processes. Before we proceed, one should note that it is *not* obvious $\mathcal{L}_\lambda(x, y)$ for $x, y \in \mathbb{R}^d$ is well defined. Since $\mathcal{H}_\lambda$ contains infinitely many points, the set of all paths from x to y is an infinite set. Thus the shortest path from x to y may not exist. The following proposition shows that this pathological behavior happens with probability zero.

Proposition 2.7. *For every $x, y \in \mathbb{R}^d$, the random variables $\mathcal{L}_\lambda(x, y)$ is well-defined with probability one.*

Proof. The point process $\mathcal{H}_\lambda$ is locally finite almost surely. Therefore $\mathcal{L}_\lambda(x, y; b)$ is well defined for any $0 < b < \infty$. Let $r_1 = \mathcal{L}_\lambda(x, y; 1)$.

Suppose a path from x to y has path length $r < r_1$ and it is not contained in $B(x; b)$, $b > 0$. Then from x, there exists a path π that ends at some point outside $B(x; b)$ and has path length less than r_1. Assume that π has n edges with Euclidean edge lengths $a_1, a_2, \ldots, a_n > 0$.

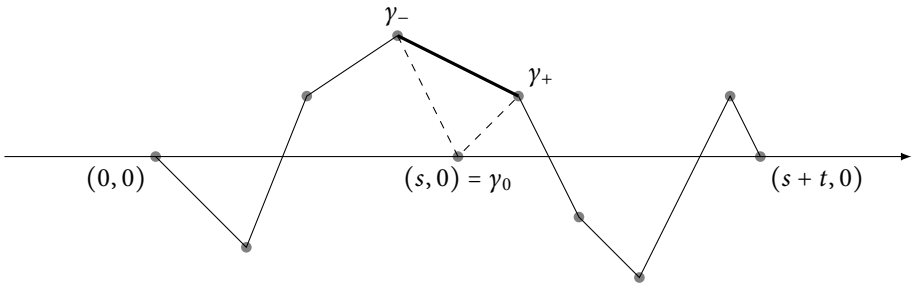


Figure 2.4: An illustration of the path paste procedure. A new path from $(0,0)$ to $(s+t,0)$ is created by removing $(s,0) = \gamma_0$ and joining γ_- and γ_+. Only the end points are fixed points in the new path.

By the triangle inequality and the Hölder's inequality,

$$b^p \le \Big(\sum_i a_i\Big)^p \le \Big(\sum_i a_i^p\Big) n^{p-1} < n^{p-1} r_1. \tag{2.16}$$

Hence $n > (r_1^{-1} b^p)^{1/(p-1)}$. By Lemma 2.4, for b large enough, the probability that such path π exists is exponentially small in $b^{p/(p-1)}$. By the Borel-Cantelli lemma, with probability one there exists $b^* > 0$ such that no path from x has path length less than r_1 and reaches some point outside $B(x; b^*)$. It follows that $\mathcal{L}_\lambda(x, y)$ is determined by all the paths in $B(x; b^*)$, and the number of paths in $B(x; b^*)$ is finite with probability one. □

Lemma 2.8. *The limit*

$$C(d,p) = \lim_{t\to\infty} \frac{1}{t} E\mathcal{L}_1(0,t) \tag{2.17}$$

exists. Furthermore if positive b_t satisfies $b_t \to \infty$ as $t \to \infty$, then

$$\lim_{t\to\infty} \frac{1}{t} E\mathcal{L}_1(0,t;b_t) = C(d,p). \tag{2.18}$$

This lemma defines the convergence value $C(d,p)$. The main body of the subsequent proofs will be devoted to show that this simple lemma on the mean convergence for Poisson point processes generalizes to complete convergence in n for a set of i.i.d. samples.

Proof of Lemma 2.8. Let $s, t, b > 0$. Consider the shortest path for $\mathcal{L}_1(0, s; b)$, and let γ_- denote the node that directly connects to $(s, 0)$. Similarly consider the path for $\mathcal{L}_1(s, s+t; b)$, and let γ_+ denote the node that directly connects to $(s, 0)$. Therefore γ_- and γ_+ are Poisson sample points incident to $(s, 0)$. Remove $\gamma_0 = (s, 0)$ in the paths, and join the nodes γ_- and γ_+ so that we have a new path connecting $(0, 0)$ and $(s + t, 0)$, and this new path length is

an upper bound of $\mathcal{L}_1(0, s+t; b)$, i.e.,

$$\mathcal{L}_1(0, s+t; b) \le \mathcal{L}_1(0, s; b) + \mathcal{L}_1(s, s+t; b) + (2^{p-1}-1)(|\gamma_0 - \gamma_-|^p + |\gamma_+ - \gamma_0|^p), \tag{2.19}$$

by the convex property of the power function for $p \ge 1$. See Figure 2.4 for an illustration. Let $\xi_b(s)$ be the maximum distance from $(s, 0)$ to the points v where the shortest path from $(s, 0)$ to v in the tubular region $\bigcup_u B((u, 0); b)$, $u \in \mathbb{R}$, is the direct path. Lemma 2.5 establishes properties of this variable. Both $|\gamma_0 - \gamma_-|$ and $|\gamma_+ - \gamma_0|$ are bounded above by $\xi_b(s)$, and

$$\mathcal{L}_1(0, s+t; b) \le \mathcal{L}_1(0, s; b) + \mathcal{L}_1(s, s+t; b) + 2^{p-1}(\xi_b(s)^p + \xi_b(s)^p). \tag{2.20}$$

Add $2^{p-1}(\xi_b(0)^p + \xi_b(s+t)^p)$ to the both sides of (2.20). By Lemma 2.5, p-th moment $E\xi_b^p = E\xi_b(0)^p$ is finite. Since the distribution of $\mathcal{H}_\lambda$ is invariant under translations,

$$E\mathcal{L}_1(0, s+t; b) + 2^p E\xi_b^p \le E\mathcal{L}_1(0, s; b) + E\mathcal{L}_1(0, t; b) + 2^{p+1} E\xi_b^p \tag{2.21}$$

Thus $E\mathcal{L}_1(0, t; b) + 2^p E\xi_b^p$ is sub-additive in t. Note that $\mathcal{L}_1(0, t; b) \le t^p$, and $t \mapsto t^p$ is Lipschitz in compact intervals. Therefore $E\mathcal{L}_1(0, t; b)$ is continuous for $t \ge 0$. A standard proof for the Fekete's lemma (for example, see Steele 1997, Lemma 1.2.1) may be easily adapted to continuous sub-additive functions, and since $E\xi_b^p$ is finite, the limit

$$C(d, p; b) = \lim_{t\to\infty} \frac{1}{t} E\mathcal{L}_1(0, t; b) = \inf_{t>0} \frac{E\mathcal{L}_1(0, t; b) + 2^p E\xi_b^p}{t} < \infty \tag{2.22}$$

exists. Substitute b with ∞, and the first part of the claim is proved.

It remains to prove the second part of the claim. Fix $\varepsilon > 0$. From (2.22) there exists $t > 0$ such that

$$\frac{E\mathcal{L}_1(0, t) + 2^p E\xi_\infty^p}{t} < C(d, p) + \varepsilon. \tag{2.23}$$

Since $\mathcal{L}_1(0, t; b)$ monotonically decreases in b and converges to $\mathcal{L}_1(0, t)$ almost surely,

$$C(d, p; b) \le \frac{E\mathcal{L}_1(0, t; b) + 2^p E\xi_b^p}{t} < \frac{E\mathcal{L}_1(0, t) + 2^p E\xi_\infty^p}{t} + \varepsilon < C(d, p) + 2\varepsilon \tag{2.24}$$

when b is large enough, by the monotone convergence theorem and Lemma 2.5. This implies that $C(d, p; b)$ monotonically converges to $C(d, p; \infty) = C(d, p)$ as $b \to \infty$ since ε

is arbitrary. On the other hand, when $b_t \to \infty$ as $t \to \infty$,

$$C(d,p) \le \lim_{t\to\infty} \frac{1}{t} E\mathcal{L}_1(0,t;b_t) \le \lim_{t\to\infty} \frac{1}{t} E\mathcal{L}_1(0,t;b) = C(d,p;b) \tag{2.25}$$

for all $b > 0$. □

The previous lemma is for a fixed point process $\mathcal{H}_1$ and a moving pair of end points $(0,0)$, $(t,0)$. The next step is to restate the lemma for a fixed pair of end points and an increasing point process $\mathcal{H}_\lambda$.

Let $z \in \mathbb{R}^d$, $0 < R_1 < R_2$. Let $v = v(z,R_1,R_2)\colon \mathbb{R}^d \to \mathbb{R}_{\ge 0}$,

$$v(u) = \begin{cases} 1 & \text{when } u \in B(z;R_2), \\ 0 & \text{when } u \notin B(z;R_2). \end{cases} \tag{2.26}$$

v is an intensity function for Poisson point process, and let $\mathcal{P}_{\lambda v}$ denote the non-homogeneous Poisson point process with intensity λv. Define $\mathcal{L}_{\lambda v}$ and related quantities for $\mathcal{P}_{\lambda v}$ as $\mathcal{L}_\lambda$ is defined for $\mathcal{H}_\lambda$.

As mentioned earlier, points $x = x_\lambda$ and $y = y_\lambda$ are parameterized by λ and are not fixed. These assumptions will remain until we carry the results to i.i.d. cases in Lemma 2.11.

Corollary 2.9. *Assume that $x_\lambda, y_\lambda \in B(z;R_1)$ for all $\lambda > 0$. If*

$$\liminf_{\lambda\to\infty} \lambda^{1/d}|x_\lambda - y_\lambda| = +\infty, \tag{2.27}$$

then

$$\lim_{\lambda\to\infty} \frac{E\mathcal{L}_{\lambda v}(x_\lambda, y_\lambda)}{\lambda^{(1-p)/d}|x_\lambda - y_\lambda|} = C(d,p). \tag{2.28}$$

Furthermore, if positive b_λ satisfies $b_\lambda \to 0$ and $\lambda^{1/d} b_\lambda \to \infty$ as $\lambda \to \infty$, then

$$\lim_{\lambda\to\infty} \frac{E\mathcal{L}_{\lambda v}(x_\lambda, y_\lambda; b_\lambda)}{\lambda^{(1-p)/d}|x_\lambda - y_\lambda|} = C(d,p). \tag{2.29}$$

This proposition hints the local nature of the shortest paths. If the endpoints x_λ, y_λ are in $B(z;R_1)$ and are away from the boundary of $B(z;R_2)$, then the shortest path lengths in $\mathcal{P}_{\lambda v}$ converge as if they were in $\mathcal{H}_\lambda$.

Proof of Corollary 2.9. The probability space may be configured so that $\mathcal{H}_\lambda$ dominates $\mathcal{P}_{\lambda v}$ since $\lambda v(u) \le \lambda$ for all $u \in \mathbb{R}^d$. Then $\mathcal{L}_\lambda(x_\lambda, y_\lambda) \le \mathcal{L}_{\lambda v}(x_\lambda, y_\lambda)$ as every path in $\mathcal{P}_{\lambda v}$ is also a path in $\mathcal{H}_\lambda$.

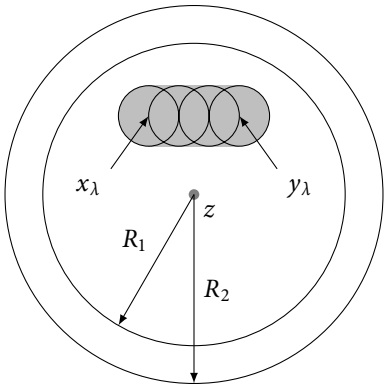


Figure 2.5: $\mathcal{L}_{\lambda v}\big(x_\lambda, y_\lambda; \lambda^{(\alpha-1)/d}|x_\lambda - y_\lambda|\big)$ is the shortest path that is contained in the shaded region. The radii of the circles inside the shaded region are $\lambda^{(\alpha-1)/d}|x_\lambda - y_\lambda|$. If λ is large enough so that $\lambda^{(\alpha-1)/d}|x_\lambda - y_\lambda| \le 2R_1\lambda^{(\alpha-1)/d} \le R_2 - R_1$, then the shaded region must be contained in $B(z; R_2)$ where the intensity is uniform.

Also for $0 < \alpha < 1$, we have $\mathcal{L}_{\lambda v}(x_\lambda, y_\lambda) \le \mathcal{L}_{\lambda v}\big(x_\lambda, y_\lambda; \lambda^{(\alpha-1)/d}|x_\lambda - y_\lambda|^\alpha\big)$ as restriction always increases the minimum path length. Note that the last upper bound equals to $\mathcal{L}_\lambda\big(x_\lambda, y_\lambda; \lambda^{(\alpha-1)/d}|x_\lambda - y_\lambda|^\alpha\big)$ when λ is large enough since $x_\lambda, y_\lambda \in B(z; R_1)$ and the region searched in the length functional will eventually be contained within $B(z; R_2)$ where the intensity is uniform, and it becomes independent of the point process outside $B(z; R_2)$. See Figure 2.5. To summarize, we have

$$\mathcal{L}_\lambda(x_\lambda, y_\lambda) \le \mathcal{L}_{\lambda v}(x_\lambda, y_\lambda) \le \mathcal{L}_\lambda\big(x_\lambda, y_\lambda; \lambda^{(\alpha-1)/d}|x_\lambda - y_\lambda|^\alpha\big). \tag{2.30}$$

We will show that both the upper and the lower bounds have the same limit.

By the invariance of $\mathcal{H}_\lambda$ under translations and rotations, without loss of generality we may set $x_\lambda = (0,0)$ and $y_\lambda = (t_\lambda, 0)$ where $t_\lambda = |x_\lambda - y_\lambda|$, hence $E\mathcal{L}_\lambda(x_\lambda, y_\lambda) = E\mathcal{L}_\lambda(0, t_\lambda)$. Then use the scale property of $\mathcal{H}_\lambda$,

$$\lim_{\lambda\to\infty} \frac{E\mathcal{L}_\lambda(0, t_\lambda)}{\lambda^{(1-p)/d} t_\lambda} = \lim_{\lambda\to\infty} \frac{E\mathcal{L}_1(0, \lambda^{1/d} t_\lambda)}{\lambda^{1/d} t_\lambda} = \lim_{t\to\infty} \frac{E\mathcal{L}_1(0,t)}{t} \tag{2.31}$$

and

$$\lim_{\lambda\to\infty} \frac{E\mathcal{L}_\lambda\big(0, t_\lambda; \lambda^{(\alpha-1)/d} t_\lambda^\alpha\big)}{\lambda^{(1-p)/d} t_\lambda} = \lim_{\lambda\to\infty} \frac{E\mathcal{L}_1\big(0, \lambda^{1/d} t_\lambda; (\lambda^{1/d} t_\lambda)^\alpha\big)}{\lambda^{1/d} t_\lambda} = \lim_{t\to\infty} \frac{E\mathcal{L}_1\big(0, t; t^\alpha\big)}{t} \tag{2.32}$$

when the limits exist. From Lemma 2.8 both limits in (2.31) and (2.32) exist and are equal to $C(d,p)$, and the claim follows from (2.30).

For $\mathcal{L}_{\lambda v}(x_\lambda, y_\lambda; b_\lambda)$ where $b_\lambda \to 0$, note that it is equal to $\mathcal{L}_\lambda(x_\lambda, y_\lambda; b_\lambda)$ for sufficiently large λ. Its convergence can be proved similarly, from the second claim of Lemma 2.8. □

2.3.4 Shortest path size

In order to apply the concentration of measure later in the chapter, we must establish that the shortest paths may not consist of too many sample points.

Lemma 2.10. *Let $\varepsilon > 0$ be fixed, and $\alpha = (d + 2p - 1)^{-1}$. If*

$$\liminf_{\lambda\to\infty} \lambda^{\alpha}|x_\lambda - y_\lambda| = +\infty, \tag{2.33}$$

then

$$\limsup_{\lambda\to\infty} \frac{1}{\lambda^{\alpha}|x_\lambda - y_\lambda|} \log \Pr\left\{ \frac{\mathcal{L}_{\lambda\nu}(x_\lambda, y_\lambda)}{\lambda^{(1-p)/d}|x_\lambda - y_\lambda|} \geq C(d,p) + \varepsilon \right\} \leq -\min\left\{ \frac{\theta_0}{2R_1}, \frac{1}{8}\left(\frac{\varepsilon}{2^p}\right)^2 \right\} \tag{2.34}$$

where $\theta_0 > 0$ is the constant defined in (2.11).

The inequality (2.34) states that the probability $\Pr\{\cdot\}$ quantity decays no slower than $\exp(-c\lambda^{\alpha}|x_\lambda - y_\lambda|)$, for some positive constant c, as λ increases to infinity. Note that the condition (2.33) is stronger than the condition (2.27), and Corollary 2.9 may be applied in the proof.

Proof. Fix λ. From the inequalities (2.30) it is sufficient to prove (2.34) holds for $\mathcal{L}_\lambda(x_\lambda, y_\lambda; b_\lambda)$ instead of $\mathcal{L}_{\lambda\nu}(x_\lambda, y_\lambda)$ for positive b_λ satisfying $b_\lambda \to 0$ as $\lambda \to \infty$. Let us choose and fix $b = b_\lambda$ for the fixed λ.

As in the procedure described by (2.20), $\mathcal{L}_\lambda(0, 2b; b)$ may be bounded above by $\mathcal{L}_\lambda(0, b; b) + \mathcal{L}_\lambda(b, 2b; b) + (2^{p-1} - 1)(Z_1^p + Y_0^p)$, where Z_k and Y_k are the first and the last link distances in $\mathcal{L}_\lambda(kb, (k+1)b; b)$, respectively. See Figure 2.4.

Note that the shortest path for $\mathcal{L}_\lambda(kb, kb+b; b)$ is not the direct path $(kb, 0) \to (kb+b, 0)$ with high probability. If it were the direct path, then $\mathcal{H}_\lambda$ is empty in $\Theta((kb, 0), (kb + b, 0))$, where Θ is defined in (2.10), and it happens with probability $\exp(-\lambda\theta_0 b^d)$, where $\theta_0 > 0$ is defined in (2.11). If none of the shortest paths for $\mathcal{L}_\lambda(kb, (k+1)b; b)$ is a direct path, then this paste procedure may be repeated,

$$\mathcal{L}_\lambda(0, mb; b) \leq \sum_{k=0}^{m-1} \Big(\mathcal{L}_\lambda\big(kb, (k+1)b; b\big) + (2^{p-1} - 1)(Z_k^p + Y_k^p) \Big), \tag{2.35}$$

with probability at least $1 - m\exp(-\lambda\theta_0 b^d)$.

If k, l are integers and $l - k \geq 3$, then $\mathcal{L}_\lambda(kb, (k+1)b; b)$ and $\mathcal{L}_\lambda(lb, (l+1)b; b)$ become mutually independent, and so are Z_k and Z_l, as well as Y_k and Y_l. Then the sum in (2.35) may split into $K \geq 3$ sums of independent variables, and each sum has at most $K^{-1}m$ summands. Note that each summand is almost surely bounded since $Z_k^p + Y_k^p \leq \mathcal{L}_\lambda(kb, kb + b; b) \leq b^p$.

Then apply the Azuma's inequality for $K = 4$ separate sequences,

$$\Pr\left\{\frac{\mathcal{L}_\lambda(0, mb; b)}{\lambda^{(1-p)/d} mb} \geq \mu_b + \varepsilon\right\} \leq me^{-\lambda\theta_0 b^d} + 4\exp\left(-\frac{(m-3)\varepsilon^2}{2^{1+2p}(\lambda^{1/d}b)^{2(p-1)}}\right), \tag{2.36}$$

where μ_b is the expectation $E\mathcal{L}_\lambda(0, b; b) + (2^{p-1} - 1)(EZ_0^p + EY_0^p)$ divided by $\lambda^{(1-p)/d} b$.

Set $m = \lfloor \lambda^{(1-\alpha)/d} |x_\lambda - y_\lambda| \rfloor$ and $mb = |x_\lambda - y_\lambda|$. By the definition, both Z_k and Y_k are bounded above by ξ_b in Lemma 2.5, and $\lambda^{(p-1)/d} b^{-1} E\xi_b^p$ shrinks to zero when $\lambda^{1/d} b \geq \lambda^{\alpha/d} \to \infty$. Apply Corollary 2.9 and Lemma 2.5 to see that μ_b converges to $C(d, p)$ as $\lambda \to \infty$. Then (2.36) becomes

$$\begin{aligned}\Pr\left\{\frac{\mathcal{L}_\lambda(x_\lambda, y_\lambda; b)}{\lambda^{(1-p)/d}|x-y|} \geq C(d,p) + 2\varepsilon\right\} &\leq \lambda^{(1-\alpha)/d}|x_\lambda - y_\lambda| e^{-\theta_0 \lambda^\alpha} \\ &\quad + 4\exp\left(-\frac{\lambda^\alpha |x_\lambda - y_\lambda| \varepsilon^2}{2^{1+2p}}\left(1 + o\left(\frac{1}{\lambda^\alpha |x_\lambda - y_\lambda|}\right)\right)\right)\end{aligned} \tag{2.37}$$

as λ and $\lambda^\alpha |x_\lambda - y_\lambda|$ tends to infinity. The claim follows since $|x_\lambda - y_\lambda|$ is bounded above by $2R_1$. □

Now we switch to the case of i.i.d. sequences. Define $L_n(s, t; R)$ for $0 < s < t$, $R > 0$ as in the case of the Poisson process. Let $z \in \mathbb{R}^d$, $0 < R_1 < R_2$. Assume that the probability density f is uniformly supported in $B(z; R_2)$. That is, $f(u) = f_0 > 0$ for all $u \in B(z; R_2)$ and $f(u) = 0$ for all $u \notin B(z; R_2)$. Let x_n and y_n be sequences of points in $B(z; R_1)$. These assumptions will remain in force until Theorem 2.16.

Recall that $|L_n(x_n, y_n)|$ is the number of nodes in the shortest path from x_n to y_n corresponding to $L_n(x_n, y_n)$.

Lemma 2.11. *Let $\alpha = (d + 2p - 1)^{-1}$, and $\rho_0 > 0$ be the constant introduced in Lemma 2.4. If $C_1 > C(d, p)\rho_0$ and $\liminf_n (nf_0)^\alpha |x_n - y_n| = +\infty$, then*

$$\limsup_{n\to\infty} \frac{1}{(nf_0)^\alpha |x_n - y_n|} \log \Pr\left\{\frac{|L_n(x_n, y_n)|}{(nf_0)^{1/d}|x_n - y_n|} > C_1\right\} < 0. \tag{2.38}$$

Proof. Let N be a Poisson variable with mean na, $a > 1$. Suppose that $N \geq n$. Then a path in $\mathcal{X}_n$ is also a path in $\mathcal{X}_N = \mathcal{P}_{naf} \subset \mathcal{H}_{naf_0}$ since $f(u) \leq f_0$ for all $u \in \mathbb{R}^d$. Choose $\varepsilon > 0$ sufficiently small so that $C_1 > (C(d, p) + 2\varepsilon)\rho_0$. Suppose that

$$L_n(x_n, y_n) \leq (C(d,p) + 2\varepsilon)(naf_0)^{(1-p)/d}|x_n - y_n| \tag{2.39}$$

and that the number of nodes in the path is at least $C_1(naf_0)^{1/d}|x_n - y_n|$. Then by Lemma 2.4,

the probability that such path exists decays exponentially in $(naf_0)^{1/d}|x_n - y_n|$ since $C_1 > (C(d,p) + 2\varepsilon)\rho_0$.

Let N' be a Poisson variable with mean na', $a' < 1$. Suppose that $N' \le n$. Then $L_n(x_n, y_n) \le L_N(x_n, y_n) = \mathcal{L}_{na'f}(x_n, y_n)$, and the probability that

$$\frac{L_n(x_n, y_n)}{(na'f_0)^{(1-p)/d}|x_n - y_n|} > C(d,p) + \varepsilon \tag{2.40}$$

is exponentially small in $(na'f_0)^{\alpha}|x_n - y_n|$ by Lemma 2.10.

Choose a and a' so that

$$\big(C(d,p) + \varepsilon\big)(a')^{(1-p)/d} < \big(C(d,p) + 2\varepsilon\big)a^{(1-p)/d} \tag{2.41}$$

so that the assumption (2.39) is satisfied when (2.40) is false. In addition, both $\Pr\{N < n\}$ and $\Pr\{N' > n\}$ are exponentially small in n by the Chernoff bound. Then the claim is proved since the slowest probability decay is determined by Lemma 2.10. □

2.3.5 Mean convergence in i.i.d. cases

We show that the mean convergence result in Corollary 2.9 holds for the i.i.d. case as well.

Proposition 2.12. *Let* $\alpha = (d + 2p)^{-1}$. *If* $\liminf_n n^{\alpha}|x_n - y_n| = +\infty$, *then*

$$\lim_{n\to\infty} \frac{EL_n(x_n, y_n)}{(nf_0)^{(1-p)/d}|x_n - y_n|} = C(d,p). \tag{2.42}$$

Proof. Let us fix x_n and y_n so that L_k denotes $L_k(x_n, y_n)$ for all $k \ge 0$. Let $C_1 > 0$ and suppose that the number of nodes $|L_n|$ in the shortest path is less than $C_1(nf_0)^{1/d}|x_n - y_n|$. Suppose that the event E_F in Lemma 2.6 occurred and all the shortest path link distances are at most $(nf_0)^{(\alpha-1)/d}$. When a sample point in $\mathcal{X}_n$ is discarded, L_{n-1} remains the same as L_n if the discarded sample point were not a node in L_n. Furthermore since E_F occurred, L_{n-1} and L_n may differ at most by $2^p(nf_0)^{(\alpha-1)p/d}$. Therefore

$$EL_{n-1} - EL_n \le \frac{C_1(nf_0)^{1/d}|x_n - y_n|}{n} \cdot 2^p(nf_0)^{(\alpha-1)p/d} + h_n EL_0, \tag{2.43}$$

where h_n denotes the probability that either $|L_n| > C_1(nf_0)^{1/d}|x_n - y_n|$, or the event E_F does not occur. EL_0 in the last term is chosen because $EL_k \le EL_0$ for all $k > 0$. By conditioning

the Poisson process, the difference between EL_n and $E\mathcal{L}_{nf}$ is at most

$$EL_0 \Pr\{N < 2^{-1}n\} + \sum_{k \geq 2^{-1}n} |EL_n - EL_k| \Pr\{N = k\}, \tag{2.44}$$

where N is a Poisson random variable with mean n. When $N < 2^{-1}n$, the mean difference is simply bounded by EL_0. Note that the first term on the right of (2.43) has monotonic decrease for fixed x_n and y_n. Therefore if $k \geq 2^{-1}n$,

$$|EL_n - EL_k| \leq 2^p C_1 |x_n - y_n| |n - k| \left(\frac{n}{2}\right)^{-1} \left(\frac{nf_0}{2}\right)^{\frac{1+p(\alpha-1)}{d}} + EL_0 \sum_{l > 2^{-1}n} h_l, \tag{2.45}$$

and since $E|N - n| \leq \sqrt{n}$ and $EL_0 = |x_n - y_n|^p$,

$$\frac{|EL_n - E\mathcal{L}_{nf}|}{(nf_0)^{(1-p)/d}|x_n - y_n|} \leq O\big((nf_0)^{\alpha p/d} n^{-1/2}\big) + \frac{\Pr\{N < 2^{-1}n\} + \sum h_l}{(nf_0)^{(1-p)/d}|x_n - y_n|^{1-p}} \tag{2.46}$$

where the summation $\sum h_l$ is still for $l > 2^{-1}n$. The first term decays to zero since $\alpha < d/(2p)$. The second term also decays to zero since while the denominator has at most polynomial decay, the numerator has exponential decay by the Chernoff bound, Lemma 2.11,[3] and Lemma 2.6. Thus the claim follows by Corollary 2.9. □

2.3.6 Concentration of measure

Now we are ready to apply Talagrand's inequality to L_n. This concentration inequality provides a high-probability lower bound on L_n, and gives a complete convergence result that will be used in the sequel.

Proposition 2.13. *Let* $\alpha = (d + 2p)^{-1}$. *If* $\liminf_n n^\alpha |x_n - y_n| = +\infty$, *then*

$$\limsup_{n \to \infty} \frac{1}{(nf_0)^\alpha |x_n - y_n|} \log \Pr\left\{ \left| \frac{L_n(x_n, y_n)}{(nf_0)^{(1-p)/d}|x_n - y_n|} - C(d,p) \right| > \varepsilon \right\} < 0. \tag{2.47}$$

Proof. This proof basically follows the proof outline of Yukich (2000, Theorem 4.1) and Talagrand (1995, Section 7.1). Let

- E_S be the event that $|L_n(x_n, y_n)| \leq C_1 (nf_0)^{1/d} |x_n - y_n|$ for some $C_1 > C(d,p)\rho_0$, (Lemma 2.11)
- E_F be the event that all the shortest path link distances are at most $(nf_0)^{(\alpha-1)/d}$, (Lemma 2.6)

[3] Note the slight difference in constant definition of α from Lemma 2.11.

- E_C be the event that at every point $u \in B(z; R_2)$, at least one of the sample points is in $B(u; (nf_0)^{(\alpha-1)/d})$.

These events occur with high probability so that their occurrence does not hinder the convergence rate stated in this proposition. As a check, both $\Pr(E_S)$ and $\Pr(E_F)$ approaches 1 exponentially fast in $(nf_0)^\alpha |x_n - y_n|$ by Lemma 2.11 and Lemma 2.6, respectively, since $|x_n - y_n| \le 2R_1$. The probability $\Pr(E_C)$ may be shown to approach 1 exponentially fast as well by a similar proof to Lemma 2.6.

For every $a > 0$, define $E(a)$ to be the event that $L_n = L_n(x_n, y_n) \ge a$. Let ω and η be outcomes in the probability space. Assume that $\omega \in E_S \cap E_F$, and that $\eta \in E(a) \cap E_C$. Let $\pi^*(\omega)$ be the shortest path $L_n(x_n, y_n)$ between x_n and y_n through realized sample points $X_1(\omega), \dots, X_n(\omega)$. If $\pi^*(\omega)$ is sequence

$$x_n = \pi_0(\omega) \to \pi_1(\omega) \to \cdots \to \pi_k(\omega) = y_n, \tag{2.48}$$

where $k = |L_n(\omega)|$, then build the path $\pi(\eta)$ from x_n to y_n through $X_1(\eta)$,…,$X_n(\eta)$ as follows. For each $i \in \{1, \dots, k-1\}$, let j denote the index where $X_j(\omega) = \pi_i(\omega)$. If $X_j(\omega) = X_j(\eta)$, then set $\pi_i(\eta) = \pi_i(\omega)$. Otherwise, since it was assumed that $\eta \in E_C$, there exists some $X_l(\eta) \in B(\pi_i(\omega); (nf_0)^{(\alpha-1)/d})$. Set $\pi_i(\eta) = X_l(\eta)$. Then it follows that $|\pi_i(\eta) - \pi_i(\omega)| \le (nf_0)^{(\alpha-1)/d}$ for all $i = 1, \dots, k$.

Let I be the set of indices i where $\pi_i(\omega) \ne \pi_i(\eta)$. Then $L(\pi(\eta)) \le L(\pi^*(\omega)) + 2|I|3^p (nf_0)^{(\alpha-1)p/d}$ since $\omega \in E_F$. On the other hand, $\eta \in E(a)$. Therefore

$$L(\pi^*(\omega)) = L_n(\omega) \ge a - 2|I|3^p (nf_0)^{(\alpha-1)p/d}. \tag{2.49}$$

Let $d_c(\omega; E(a) \cap E_C)$ be the convex distance of ω to $E(a) \cap E_C$ (See Talagrand 1995, Section 4.1). By Talagrand (1995, Lemma 4.1.2), there exists $\eta \in E(a) \cap E_C$ such that $|I| \le d_c(\omega; E(a) \cap E_C)\sqrt{|L_n(\omega)|}$, hence

$$L_n(\omega) \ge a - 2 \cdot 3^p \cdot d_c(\omega; E(a) \cap E_C)\sqrt{|L_n(\omega)|}(nf_0)^{(\alpha-1)p/d}. \tag{2.50}$$

In particular, if $L_n(\omega) \le a - u$ for $u > 0$,

$$\begin{aligned} d_c(\omega; E(a) \cap E_C(\alpha)) &\ge \frac{u}{2 \cdot 3^p \cdot \sqrt{|L_n(\omega)|}} (nf_0)^{(1-\alpha)p/d} \\ &\ge \frac{u}{2 \cdot 3^p \cdot \sqrt{C_1 (nf_0)^{1/d} |x_n - y_n|}} (nf_0)^{(1-\alpha)p/d} \end{aligned} \tag{2.51}$$

since $\omega \in E_S$.

Let M_n be a median of L_n. If $a = M_n$, then by (2.51) and Talagrand (1995, Theorem 4.1.1), for $u > 0$,

$$\Pr\{L_n \le M_n - u\} \le 3\exp\left(-\frac{C_2 u^2}{|x_n - y_n|}(nf_0)^{\frac{2p(1-\alpha)-1}{d}}\right) + \big(1 - \Pr(E_F)\big) + \big(1 - \Pr(E_S)\big) \tag{2.52}$$

where $C_2 = (2^4 3^{2p} C_1)^{-1}$, since $\Pr(E_C)$ approaches 1 as $n \to \infty$ and $\Pr(E(M_n) \cap E_C(\alpha)) \ge 3^{-1}$ for sufficiently large n. For the upper bound part, let $a = M_n + u$. Then repeat a similar procedure,

$$\Pr\{L_n \ge M_n + u\} \le 3\exp\left(-\frac{C_2 u^2}{|x_n - y_n|}(nf_0)^{\frac{2p(1-\alpha)-1}{d}}\right) + \big(1 - \Pr(E_C)\big) \tag{2.53}$$

for sufficiently large n since both $\Pr(E_F)$ and $\Pr(E_S)$ converge to 1 as $n \to \infty$. Therefore

$$\Pr\left\{\frac{|L_n - M_n|}{(nf_0)^{(1-p)/d}|x_n - y_n|} > u\right\} \le 3\exp\Big(-C_2 u^2 |x_n - y_n|(nf_0)^{\alpha}\Big) + h_n, \tag{2.54}$$

where $h_n = \big(1 - \Pr(E_C)\big) + \big(1 - \Pr(E_F)\big) + \big(1 - \Pr(E_S)\big)$.

Now we show that the median and the mean are close. By the Jensen's inequality, $|EL_n - M_n| \le E|L_n - M_n|$, and if (2.54) is integrated with respect to u,

$$\frac{|EL_n - M_n|}{(nf_0)^{(1-p)/d}|x_n - y_n|} \le 3\sqrt{\frac{\pi}{C_2|x_n - y_n|(nf_0)^{\alpha}}} + \Big((nf_0)^{1/d}|x_n - y_n|\Big)^{p-1} h_n. \tag{2.55}$$

Note that the range of the integral is restricted since $L_n \le |x_n - y_n|^p$. All events E_C, E_F, and E_S have their probability approach to 1 exponentially fast. Therefore

$$\lim_{n\to\infty} \frac{|EL_n - M_n|}{(nf_0)^{(p-1)/d}|x_n - y_n|} = 0. \tag{2.56}$$

By Proposition 2.12, if n is large enough then

$$\Pr\left\{\left|\frac{L_n}{(nf_0)^{(p-1)/d}|x_n - y_n|} - C(d,p)\right| > \varepsilon\right\} \le \Pr\left\{\frac{|L_n - M_n|}{(nf_0)^{(p-1)/d}|x_n - y_n|} > \frac{\varepsilon}{2}\right\}. \tag{2.57}$$

Then the claim follows from (2.54). □

Remark 2.14. At this point, the value of (2.47) is worth discussion. From (2.54), the limit supremum value is determined by the constant $C_2 = (2^4 3^{2p} C_1)^{-1}$ and $h_n = \big(1 - \Pr(E_C)\big) + \big(1 - \Pr(E_F)\big) + \big(1 - \Pr(E_S)\big)$.

From Lemma 2.6, the log limit supremum of $1 - \Pr(E_F)$ is $-\theta_1 < 0$, dependent only on d and p. From a similar proof of $1 - \Pr(E_C)$, one may conclude that the log limit supremum of $1 - \Pr(E_C)$ is also some value, say $-\theta_2 < 0$, dependent only on d and p.

Now the event E_S. From Lemma 2.11 which in turn is dependent on Lemma 2.10, the log limit supremum is

$$-\min\left\{\frac{\theta_0}{2R_1}, \frac{1}{8}\left(\frac{\delta}{2^p}\right)^2\right\} \quad \text{where} \quad 2\delta < \frac{C_1}{\rho_0} - C(d,p). \tag{2.58}$$

To summarize, the value of (2.47) is bounded above by

$$-\min\left\{\frac{\theta_0}{2R_1}, \theta_1, \theta_2, \frac{1}{8}\left(\frac{\delta}{2^p}\right)^2, C_2\varepsilon^2 = \frac{\varepsilon^2}{2^4 3^{2p} C_1}\right\}. \tag{2.59}$$

In reader will see in the rest of the proofs that R_1 will not increase indefinitely and it may be bounded above. Other values C_2 and δ are determined by $C_1 > C(d,p)\rho_0$. Note that ρ_0 from Lemma 2.4 depends only on d and p.

Therefore the value of (2.47) is some strictly negative value that depends only on d, p, and ε. When ε is sufficiently small, the log limit supremum is $-C_2\varepsilon^2$. Otherwise it is some saturated value dependent only d and p.

To make the dependencies clear, define $\varsigma(d,p), \bar{\omega}(d,p) > 0$ and function $\varsigma(d,p;\varepsilon)$ of $\varepsilon > 0$ such that (2.47) becomes

$$\limsup_{n\to\infty} \frac{1}{(nf_0)^{\alpha}|x_n - y_n|} \log \Pr\left\{\left|\frac{L_n(x_n,y_n)}{(nf_0)^{(1-p)/d}|x_n - y_n|} - C(d,p)\right| > \varepsilon\right\} \le -\varsigma(d,p;\varepsilon) < 0$$

and

$$\varsigma(d,p;\varepsilon) = \begin{cases} \varsigma(d,p)\varepsilon^2 & \text{when } \varepsilon < \bar{\omega}(d,p), \\ \varsigma(d,p)\bar{\omega}(d,p)^2 & \text{otherwise.} \end{cases} \tag{2.60}$$

We next relax the uniformity condition on the probability density, and instead assume that it is only locally uniform. That is, the density f is still uniform in $B(z;R_2)$ but may have probability mass outside of $B(z;R_2)$. Therefore $f(u)$ may be arbitrary non-negative function over $u \notin B(z;R_2)$ as long as it integrates to one over the entire space.

The following lemma helps to decouple the local probability distribution from outer region.

Lemma 2.15. *Let $x \in B(z;R_2)$. Denote by $L_n(x;r)$, $r > 0$ the minimum power-weighted path length over all shortest paths from x to all boundary points of $B(x;r)$, i.e., $L_n(x;r) =$*

$\inf L_n(x,u)$ *over all* $u \in \partial B(x;r)$. *Choose* r *so that* $B(x;r) \subset B(z;R_2)$. *Then*

$$\limsup_{n\to\infty} \frac{1}{r\,(nf_0)^{\alpha}} \log \Pr\left\{ \left| \frac{L_n(x;r)}{(nf_0)^{(1-p)/d} r} - C(d,p) \right| > \varepsilon \right\} \le -\varsigma(d,p;\varepsilon). \tag{2.61}$$

Proof. First note that the boundary of $B(x;r)$ may be covered with open balls of radii $(nf_0)^{(\alpha-1)/d}$, and the number of cover elements m may be chosen less than $(2nf_0r)^d$. Let $v_1, v_2, \ldots, v_m$ be the centers of the cover elements. If the event E_F in Lemma 2.6 occurs and

$$\left| \frac{L_n(x,v_k)}{(nf_0)^{(1-p)/d} r} - C(d,p) \right| \le \frac{\varepsilon}{2}, \tag{2.62}$$

for all $k = 1, \ldots, m$, then $|(nf_0)^{(p-1)/d} r^{-1} L_n(x,u) - C(d,p)| < \varepsilon$ for all u on the boundary of $B(x;r)$ for sufficiently large n. Note that $L_n(x;r)$ is not affected by the probability distribution outside $B(x;r) \subset B(z;R_2)$ since the boundary of $B(x;r)$ disconnects the interior and the exterior. If the shortest path to the boundary were to reach any point outside $B(x;r)$, the path must have already passed through the boundary, which is a contradiction. Therefore Proposition 2.13 may be applied to $v_1, v_2, \ldots, v_m$, and (2.61) follows since $m \le (2nf_0r)^d$ is of polynomial order. □

Theorem 2.16. *Let* $0 < b < 1$ *and* $c > 0$ *be constants. Let* $\alpha = (d+2p)^{-1}$. *Then*

$$\limsup_{n\to\infty} \frac{1}{(nf_0)^{(1-b)\alpha}} \log \Pr\left\{ \sup_{x,y} \left| \frac{L_n(x,y)}{(nf_0)^{(1-p)/d}|x-y|} - C(d,p) \right| > \varepsilon \right\} \le -\varsigma(d,p;\varepsilon) \tag{2.63}$$

where the supremum is taken over x *and* y *such that* $x, y \in B(z;4^{-1}R_2)$ *and* $(nf_0)^{b\alpha}|x-y| \ge c$.

The condition of $(nf_0)^{b\alpha}|x-y| \ge c$ is to enforce that $(nf_0)^{\alpha} \inf|x-y|$ has polynomial order and to prevent sub-polynomial, e.g., logarithmic, growth. Note that we no longer have sequences x_n and y_n but rather we have a supremum over some subset points x, y. This change in the formulation will turn out to be useful later when we adapt the result to open covers of compact regions.

Proof of Theorem 2.16. Recall the definition of $L(x;r)$ from Lemma 2.15. Let $\xi_i, \xi_j \in B(z;4^{-1}R_2)$. If

$$\frac{L_n(\xi_i;5R_2/8)}{(nf_0)^{(1-p)/d}(5R_2/8)} \ge C(d,p) - \delta \tag{2.64}$$

holds where $\delta > 0$ is chosen so that $(C(d,p)+\varepsilon)(R_2/2) < (C(d,p)-\delta)(5R_2/8)$, then

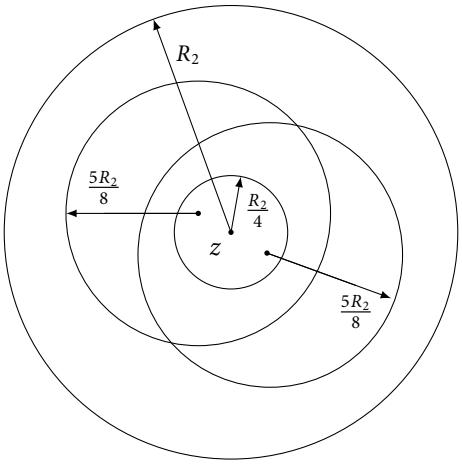


Figure 2.6: All ξ_i's are within $B(z; 4^{-1}R_2)$. Therefore all $B(\xi_i; 5R_2/8)$ are contained in $B(z; R_2)$. Since $|\xi_i - \xi_j| \le 2^{-1}R_2$, with high probability we have $L_n(\xi_i, \xi_j) < L_n(\xi_i; 5R_2/8)$, and $L_n(\xi_i, \xi_j)$ becomes independent of the outside $B(\xi_i; 5R_2/8) \subset B(z; R_2)$ due to the annulus buffer region $\{u: 4^{-1}R_2 < |z - u| < R_2\}$.

whether

$$\frac{L_n(\xi_i, \xi_j)}{(nf_0)^{(1-p)/d}|\xi_i - \xi_j|} \le C(d, p) + \varepsilon \tag{2.65}$$

becomes independent of the sample points—hence the probability distribution—outside $B(z; R_2)$, since $|x - y| \le (R_2/2)$ and $B(x; (5R_2/8)) \subset B(z; R_2)$. See Figure 2.6.

Let $x, y \in B(z; 4^{-1}R_2)$. Again suppose that the event E_F occurs, and let $\{B(\xi_i; (nf_0)^{(\alpha-1)/d}), 1 \le i \le m\}$ be an open cover of $B(z; 4^{-1}R_2)$ with $m \le (nf_0R_2)^d$. Then there exists ξ_i, ξ_j such that $|x - \xi_i| < (nf_0)^{(\alpha-1)/d}$ and $|y - \xi_j| < (nf_0)^{(\alpha-1)/d}$, hence

$$\big||x - y| - |\xi_i - \xi_j|\big| < 2(nf_0)^{(\alpha-1)/d} \tag{2.66}$$

and

$$|L_n(x, y) - L_n(\xi_i, \xi_j)| \le 2^{p+1}(nf_0)^{(\alpha-1)p/d}. \tag{2.67}$$

Since $|x - y| \ge c(nf_0)^{-\alpha b}$, it follows that if n is sufficiently large,

$$\left|\frac{L_n(x, y)}{|x - y|(nf_0)^{(1-p)/d}} - \frac{L_n(\xi_i, \xi_j)}{|\xi_i - \xi_j|(nf_0)^{(1-p)/d}}\right| < \frac{\varepsilon}{2}, \tag{2.68}$$

and

$$\Pr\left\{\sup_{x,y}\left|\frac{L_n(x,y)}{(nf_0)^{(1-p)/d}|x-y|}-C(d,p)\right|>\varepsilon\right\} \tag{2.69}$$

$$\le \sum_{i,j}\left\{\left|\frac{L_n(\xi_i,\xi_j)}{(nf_0)^{(1-p)/d}|\xi_i-\xi_j|}-C(d,p)\right|>\frac{\varepsilon}{2}\right\} \tag{2.70}$$

$$+\sum_{i=1}^{m}\Pr\left\{\frac{L_n\big(\xi_i;(5R_2/8)\big)}{(nf_0)^{(1-p)/d}(5R_2/8)}<C(d,p)-\delta\right\}, \tag{2.71}$$

where the first sum is over i and j that $|\xi_i-\xi_j|\ge c(nf_0)^{-\alpha b}-2(nf_0)^{(\alpha-1)/d}$. Note that the probability bound in Proposition 2.13 accounts for E_F as well. The claim follows from Proposition 2.13 and Lemma 2.15 as $m\le(nf_0R_2)^d$ is of polynomial order. □

Now we discard the local uniform probability density condition, i.e., f restricted to $B(z;R_2)$ is no longer uniform.

Corollary 2.17. *Let $f_\circ>0$. Let α,b,c be defined as in Theorem 2.16. Suppose that $f(u)\ge f_\circ$ for all $u\in B(z;R_2)$. Then*

$$\limsup_{n\to\infty}\frac{1}{(nf_\circ)^{(1-b)\alpha}}\log\Pr\left\{\sup_{x,y}\frac{L_n(x,y)}{(nf_\circ)^{(1-p)/d}|x-y|}<C(d,p)+\varepsilon\right\}\le-\varsigma(d,p;\varepsilon)$$

where the supremum is taken over x and y satisfying $x,y\in B(z;4^{-1}R_2)$ and $(nf_\circ)^{b\alpha}|x-y|\ge c$. Similarly, let $f^\circ>0$. Suppose that $f(u)\le f^\circ$ for all $u\in B(z;R_2)$ and $\int_{B(z;R_2)}f^\circ\le 1$. Then

$$\limsup_{n\to\infty}\frac{1}{(nf^\circ)^{(1-b)\alpha}}\log\Pr\left\{\sup_{x,y}\frac{L_n(x,y)}{(nf^\circ)^{(1-p)/d}|x-y|}>C(d,p)-\varepsilon\right\}\le-\varsigma(d,p;\varepsilon)$$

where the supremum is taken over x and y satisfying $x,y\in B(z;4^{-1}R_2)$ and $(nf^\circ)^{b\alpha}|x-y|\ge c$.

Proof. If a sample point X_i is in $B(z;R_2)$, then discard the point with probability $f(X_i)^{-1}f_\circ$. Let $L_\circ(x,y)$ denote the shortest path length through the filtered sample. Note that $L_n(x,y)\le L_\circ(x,y)$ and the filtered sample has uniform density $f_\circ$ in $B(z;R_2)$. The first claim holds by the inequality

$$\frac{L_n(x,y)}{(nf_\circ)^{(1-p)/d}|x-y|}\le\frac{L_\circ(x,y)}{(nf_\circ)^{(1-p)/d}|x-y|}<C(d,p)+\varepsilon \tag{2.72}$$

and Theorem 2.16. Repeat a similar procedure for the second claim with f°. □

2.3.7 Convergence in Riemannian manifolds

Finally we move to the case when the probability distribution is supported in a Riemannian manifold. Let (M, g_1) be a (smooth) d-dimensional complete connected Riemannian manifold without boundary, where g_1 is the base metric. Thus the edge weights of the PWG are geodesic distances under g_1. From now on, the shortest path $L_n(x, y)$ for $x, y \in M$ is always based on g_1.

Note that the probability density function is dependent on the Riemannian metric. Let f denote the probability density function under g_1. It will always be assumed that f is smooth and $\inf_M f > 0$. For $p \geq 1$, define a conformal family of metrics $g_p = f^{2(1-p)/d} g_1$, that is if $x \in M$ and $u, v \in T_x M$, then $g_p(u, v) = f(x)^{2(1-p)/d} g_1(u, v)$. Let $\mathrm{dist}_p(x, y)$ denote the geodesic distance between $x, y \in M$ under the metric g_p, $p \geq 1$.

Lemma 2.18. *Let $z \in M$. Let $0 < b < 1$ and $c > 0$ be constants. Then for every fixed $\varepsilon > 0$, there exists $R > 0$ such that*

(2.73)
$$\limsup_{n\to\infty} \frac{1}{(nf(z))^{(1-b)/(d+2p)}} \log \Pr\left\{ \sup_{x,y} \left| \frac{L_n(x, y)}{n^{(1-p)/d} \mathrm{dist}_p(x, y)} - C(d, p) \right| > \varepsilon \right\} \leq -\varsigma(d, p; \varepsilon)$$

where the supremum is taken over $x, y \in M$ such that $\mathrm{dist}_1(x, z) < R$, $\mathrm{dist}_1(y, z) < R$, and $\mathrm{dist}_1(x, y) \geq c(nf(z))^{-b/(d+2p)}$.

Proof. Let $\varphi \colon U \subset M \to V \subset \mathbb{R}^d$ be a normal chart map of z where $\varphi(z) = 0$. By the properties of the normal chart, the push-forward $\varphi_* g_1$ of the metric g_1 over φ is Euclidean at $0 = \varphi(z)$. Choose $\delta > 0$ such that

$$\frac{C(d, p) + 2^{-1}\varepsilon}{C(d, p) + \varepsilon} < \left(\frac{1-\delta}{1+\delta}\right)^p \quad \text{and} \quad \frac{C(d, p) - \varepsilon}{C(d, p) - 2^{-1}\varepsilon} < \left(\frac{1-\delta}{1+\delta}\right)^p. \tag{2.74}$$

By continuity there exists $R > 0$ such that for all $x \neq y \in M$ satisfying the conditions $\mathrm{dist}_1(x, z) < 4R$ and $\mathrm{dist}_1(y, z) < 4R$, we have

$$1 - \delta < \frac{\mathrm{dist}_1(x, y)}{|\varphi(x) - \varphi(y)|} < 1 + \delta \tag{2.75}$$

and

$$(1 - \delta)^d \sup_R f < f(z) < (1 + \delta)^d \inf_R f \tag{2.76}$$

where $\sup_R f = \sup_{\mathrm{dist}_1(x,z)<4R} f(x)$, and $\inf_R f = \inf_{\mathrm{dist}_1(x,z)<4R} f(x)$. Shrink U if necessary so that $U = \{u \in M \colon \mathrm{dist}_1(u, z) < 4R\}$.

Let $L_n(\varphi(x), \varphi(y))$ denote the shortest path between $\varphi(x), \varphi(y) \in \mathbb{R}^d$ in Euclidean

metric. Apply Corollary 2.17 for the shortest paths inside V so that

$$C(d,p) - \frac{\varepsilon}{2} \le \frac{L_n(\varphi(x),\varphi(y))}{(nf(z))^{(1-p)/d}|\varphi(x)-\varphi(y)|} \le C(d,p) + \frac{\varepsilon}{2} \tag{2.77}$$

holds with high probability. Then by (2.74), (2.75), and (2.76)

$$\big(C(d,p)-\varepsilon\big)(\inf_R f)^{(1-p)/d} \le \frac{L_n(x,y)}{n^{(1-p)/d}\operatorname{dist}_1(x,y)} \le \big(C(d,p)+\varepsilon\big)(\sup_R f)^{(1-p)/d} \tag{2.78}$$

when δ is sufficiently small.

Note that $\operatorname{dist}_p(x,y) \le \operatorname{dist}_1(x,y)(\inf_R f)^{(1-p)/d}$, and if the minimal geodesic curve from x to y under g_p is contained in U, then

$$\operatorname{dist}_p(x,y) \ge \operatorname{dist}_1(x,y)(\sup_R f)^{(1-p)/d}. \tag{2.79}$$

If a (piece-wise) smooth curve from x exits outside U, then the curve length under g_p must be at least $(3R)(\sup_R f)^{(1-p)/d}$ since $\operatorname{dist}_1(x,z) < R$ and by the definition of U. Therefore if δ is small enough, (2.79) holds for all x and y since $\operatorname{dist}_1(x,y) < 2R$ and by the assumption (2.76).[4]

Combining the above results, we obtain that

$$\big(C(d,p)-\varepsilon\big)\operatorname{dist}_p(x,y) \le n^{(p-1)/d}L_n(x,y) \le \big(C(d,p)+\varepsilon\big)\operatorname{dist}_p(x,y) \tag{2.80}$$

holds with high probability. □

The main result of this chapter is finally proved using Lemma 2.18 applied to an open cover of the manifold.

Proposition 2.19. *Assume that M is compact. Let $0 < b < 1$ and $c > 0$ be constants. Then for every fixed $\varepsilon > 0$,*

$$\limsup_{n\to\infty}(n\inf f)^{(b-1)/(d+2p)} \Pr\left\{\sup_{x,y}\left|\frac{L_n(x,y)}{n^{(1-p)/d}\operatorname{dist}_p(x,y)} - C(d,p)\right| > \varepsilon\right\} < -\varsigma(d,p;\varepsilon)$$

where the supremum is taken over x and y such that $\operatorname{dist}_1(x,y) \ge c(n\inf f)^{-b/(d+2p)}$.

Proof. The crux of the proof is that the shortest path length L_n has near sub- and super-additivity with high probability. We will show that if Lemma 2.18 holds in open cover

[4] The exact condition for δ is that $\big((1-\delta)/(1+\delta)\big)^{p-1} > 2/3$, and the condition is not dependent on the choice of R. Therefore there is no cycle trap.

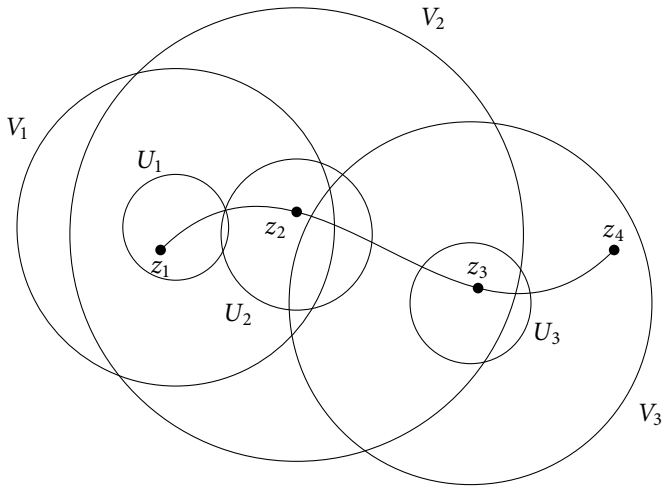


Figure 2.7: Path division procedure described in Proposition 2.19. Here $k = 4$. Note that $z_i \in U_i$ and $z_{i+1} \in V_i$ for $i = 1, 2, 3$. Shortest path is depicted as a smooth curve for illustration purpose only and it is actually piece-wise geodesic.

elements, then the local convergences may be assembled together to yield global convergence of the curve length.

For each $\xi_i \in M$, we may associate positive $R_i > 0$ such that Lemma 2.18 holds within the region $V_i = \{v \in M : \mathrm{dist}_1(v, \xi_i) < 3R_i\}$. Let $U_i = \{v \in M : \mathrm{dist}_1(v, \xi_i) < R_i\}$. Since M is assumed to be compact, there is finite $m > 0$, $\{\xi_i \in M\}_{i=1}^m$, and corresponding $\{R_i > 0\}_{i=1}^m$ such that corresponding $\{U_i\}$ is a finite open cover of M.

Suppose that $x, y \in M$ and assume to the contrary that

$$n^{(p-1)/d} L_n(x, y) < (C(d, p) - \varepsilon) \, \mathrm{dist}_p(x, y). \tag{2.81}$$

Reorder the indices if necessary so that $x \in U_1$. Define $z_1 = x$. If $L_n(x, y)$ ever exits V_1, then a point $z_2 \in V_1$ on the shortest path may be chosen such that $z_2 \notin U_1$. Then $\mathrm{dist}_1(z_1, z_2) \geq R_i$. Reorder the indices of the open cover again if necessary so that z_2 is in U_2. Repeat the procedure until $L_n(x, y)$ ends in, say V_k. Set $z_{k+1} = y$. Then points $x = z_1, z_2, \ldots, z_k, z_{k+1} = y$ satisfy $z_i, z_{i+1} \in V_i$ for $i = 1, 2, \ldots, k$, and $\mathrm{dist}_1(z_i, z_{i+1}) \geq R_i \geq R$ for $i = 1, 2, \ldots, k-1$, where $R = \min_i R_i$. The last link distance $\mathrm{dist}_1(z_k, z_{k+1})$ may be less than R. However, note that $z_{k-1} \in U_{k-1}$ and $y = z_{k+1} \notin V_{k-1}$ by the definition, hence $\mathrm{dist}_1(z_{k-1}, z_{k+1}) > 2R_{k-1} \geq 2R$. Therefore z_k may be adjusted so that $\mathrm{dist}_1(z_k, z_{k+1}) \geq R$ as well, and it is easily checked that z_k is still in V_k. See Figure 2.7 for illustration.

If

$$(C(d,p)-\varepsilon)\,\mathrm{dist}_p(z_i,z_{i+1}) \le n^{(p-1)/d} L_n(z_i,z_{i+1}) \tag{2.82}$$

for all $i = 1, 2, \ldots, k$, then by the triangle inequality and the property of the power function that $\alpha^p + \beta^p \le (\alpha+\beta)^p$ for $\alpha, \beta \ge 0$,

$$(C(d,p)-\varepsilon)\,\mathrm{dist}_p(x,y) \le (C(d,p)-\varepsilon)\sum_{i=1}^{k}\mathrm{dist}_p(z_i,z_{i+1}) \tag{2.83}$$

$$\le \sum_{i=1}^{k} n^{(p-1)/d} L_n(z_i,z_{i+1}) \tag{2.84}$$

$$\le n^{(p-1)/d} L_n(x,y) < (C(d,p)-\varepsilon)\,\mathrm{dist}_p(x,y) \tag{2.85}$$

and a contradiction is encountered. Therefore (2.82) must not hold for some pair z_i and z_{i+1}. Since m is finite and Lemma 2.18 should hold in $V_1, V_2, \ldots, V_m$,

$$\limsup_{n\to\infty}(n \inf f)^{(b-1)/(d+2p)} \Pr\left\{\inf_{x,y}\frac{L_n(x,y)}{n^{(1-p)/d}\,\mathrm{dist}_p(x,y)} < C(d,p)-\varepsilon\right\} < 0. \tag{2.86}$$

For the upper bound, we follow a similar strategy to Bernstein *et al.* (2000). If $z_1 = x$, $z_{k+1} = y$, and z_i are points on the minimal geodesic curve from x to y under g_p, then $\mathrm{dist}_p(x,y) = \sum_{i=1}^{k}\mathrm{dist}_p(z_i,z_{i+1})$. It has been argued above that the points may be chosen and indices of the open cover may be rearranged such that $z_i, z_{i+1} \in V_i$ and $\mathrm{dist}_1(z_i,z_{i+1}) \ge R$ for all $i = 1, 2, \ldots, k$. Note that Lemma 2.18 depends on Proposition 2.13, and Proposition 2.13 includes the event E_F in Lemma 2.6. Therefore each paste procedure may incur additional cost of at most $2^p n^{p(\alpha-1)/d}$ so that

$$L_n(x,y) \le \sum_{i=1}^{k} L_n(z_i,z_{i+1}) + k2^p n^{(\alpha-1)p/d} \tag{2.87}$$

where $\alpha = (d+2p)^{-1}$. Therefore if Lemma 2.18 holds in $V_1, V_2, \ldots, V_m$, then

$$n^{(p-1)/d} L_n(x,y) \le \left(C(d,p)+\frac{\varepsilon}{2}\right)\left(\mathrm{dist}_p(x,y) + k2^p n^{(\alpha p-1)/d}\right), \tag{2.88}$$

and if n is large enough, $n^{(p-1)/d} L_n(x,y) \le (C(d,p)+\varepsilon)\,\mathrm{dist}_p(x,y)$ since $n^{(\alpha p-1)/d} n^{\alpha b}$ shrinks to zero as $n \to \infty$. Therefore the claim is proved by applications of Lemma 2.18 to $V_1, V_2, \ldots, V_m$. □

Remark 2.20. Proposition 2.19 implies that

$$\lim_{n\to\infty} \frac{L_n(x,y)}{n^{(1-p)/d}\,\mathrm{dist}_p(x,y)} = C(d,p) \quad \text{c.c.} \tag{2.89}$$

for every pair $x, y \in M$, and it proves Theorem 2.1. Proposition 2.19 in fact implies a stronger result. For instance, let $\mathcal{Z}_n$ be a sequence of finite subsets of M. If $|\mathcal{Z}_n|$ grows as a polynomial in n and if

$$\liminf_{n\to\infty} n^{b/(d+2p)} \min_{x,y\in\mathcal{Z}_n} \mathrm{dist}_1(x,y) \geq c > 0, \tag{2.90}$$

then

$$\sum_{n\geq 1} \Pr\left\{ \sup_{x,y\in\mathcal{Z}_n} \left| \frac{L_n(x,y)}{n^{(1-p)/d}\,\mathrm{dist}_p(x,y)} - C(d,p) \right| > \varepsilon \right\} < \infty. \tag{2.91}$$

2.4 Extensions

Some limitations of Theorem 2.1 and Proposition 2.19 may be removed and further extended. This section discusses a few of them.

2.4.1 Super-additive dissimilarity

Many learning methods work with neighborhood graphs to search local geometric information (Tenenbaum *et al.* 2000; Roweis and Saul 2000; Donoho and Grimes 2003; Belkin and Niyogi 2003; Weinberger *et al.* 2004). Power-weighted shortest path, however, need not begin with neighborhood graphs but finds the neighborhood structure automatically. However, the absence of the neighborhood graph may slow down the convergence rate. Suppose that p is close to 1. Then the shortest paths are almost straight lines, and a large number of sample points are required to converge. See Lemma 2.6 to find that the constant θ_1 is near zero when p is nearly 1. See Chapelle and Zien (2005) for similar issues in shortest path approaches.

Finding the neighborhoods is equivalent to pruning large edges in the complete graph. If $x, y \in M$ pair is far away and should be pruned from the graph, such pair should be assigned a large weight—larger than $|x-y|^p$—so that the direct path $x \to y$ is removed from consideration in early step. For the purpose, we introduce super-additive dissimilarities.

Let M be a manifold. Let $h\colon M \times M \to \mathbb{R}$ be a continuous function. We say h is *super-additive of order* $p > 1$ if

SPA1 $h(x,y) \geq 0$ and $h(x,x) = 0$ for $x, y \in M$,

SPA2 for every open neighborhood U of $x \in M$, there exists some $\varepsilon > 0$ such that $h(x, y) \geq \varepsilon$ for all $y \notin U$,

SPA3 for every $\delta > 0$ and $x \in M$ there exists an open neighborhood U of x such that for every $y, z \in U$,

$$1 - \delta < \frac{h(y, z)}{\mathrm{dist}_1(y, z)^p} < 1 + \delta.$$

Let $L_n(x, y; h)$ denote the shortest path length from x to y but measured by h instead of dist_1^p. Condition SPA2 prevents some pathological paths. It also implies that h induces the topology of M. Condition SPA3 determines the decay rate of $L_n(x, y; h)$.

Super-additive dissimilarities may be defined with super-additive functions in $\mathbb{R}_{\geq 0}$. If $\tilde{h}: \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that $h(0) = 0$ and $h(r + s) > h(r) + h(s)$ for all $r, s > 0$, then $h(x, y) = \tilde{h}(\mathrm{dist}_1(x, y))$ is a super-additive dissimilarity. For example, $\tilde{h}(r) = 2(\cosh r - 1) = r^2 + O(r^4)$ induces a super-additive dissimilarity of order 2.

Remark 2.21. Super-additive dissimilarity is closely related to super-additive Euclidean functionals. Let $R \subset \mathbb{R}^d$ be a d-dimensional rectangle and $F \subset R$ be a finite subset. If A^p maps (F, R) into reals and satisfy $A^p(\varnothing, R) = 0$,

$$A^p(F + y, R + y) = A^p(F, R) \text{ for all } y \in \mathbb{R}^d, \tag{2.92}$$

$$A^p(\alpha F, \alpha R) = \alpha^p A^p(F, R) \text{ for all } \alpha > 0, \tag{2.93}$$

and

$$A^p(F, R) \geq A^p(F \cap R_1, R_1) + A^p(F \cap R_2, R_2) \tag{2.94}$$

whenever R is partitioned into R_1, R_2, then A^p is a super-additive Euclidean functional of order p (Steele 1981; Steele 1988; Yukich 1998, Chapter 3). The condition (2.93) is often called homogeneity of order p, and SPA3 above is a local relaxation of the homogeneity requirement. Lemma 2.22 implies that SPA3 implies (2.94) with large number of sample points.

Lemma 2.22. *Let $x, y \in M$. For every $\varepsilon > 0$ there exists a finite sequence $x = z_0, z_1, \ldots, z_k = y$ in M such that*

$$\sum_{i=0}^{k-1} h(z_i, z_{i+1}) < \varepsilon. \tag{2.95}$$

Proof. Let $x \in M$. Let U be the neighborhood described by SPA3. Shrink U if necessary so that U is path-connected. Let $y \in U$ and $\gamma: [0, 1] \to M$ be a curve in U such that $\gamma(0) = x$,

$\gamma(1) = y$. Let $0 = t_0 < t_1 < \cdots < t_k = 1$ be a finite partition of $[0,1]$. Then

$$\sum_{i=0}^{k-1} h(\gamma(t_i), \gamma(t_{i+1})) < (1+\delta) \sum_{i=0}^{k-1} \mathrm{dist}_1(\gamma(t_i), \gamma(t_{i+1}))^p \tag{2.96}$$

and the right side converges to zero as the partition refines.[5]

Now suppose y need not be in U. Choose a curve $\gamma\colon [0,1] \to M$ as before. Since γ is compact, there exists a finite open cover $U_1, \ldots, U_k$ with property SPA3. Repeat the procedure above in each element. □

Lemma 2.6 holds for h in place of power-weighted distance.

Corollary 2.23. *Fix n and $0 < \alpha < 1$. Let $U \subset M$ satisfying SPA3. Define the event $E_U(h; i, j)$ for each pair $1 \le i, j \le n$ such that $E_U(h; i, j)$ does* not *occur if and only if (i) both X_i and X_j are in U, (ii) $h(X_i, X_j) > (nf_0)^{(\alpha-1)/dp}$, and (iii) the h-shortest path from X_i to X_j contains no sample point X_k other than X_i and X_j. Let $E_U(h) = \bigcap_{i,j} E_U(h; i, j)$. Then*

$$\limsup_{n\to\infty} \frac{1}{(n \inf_U f)^\alpha} \log\bigl(1 - \Pr(E_U(h))\bigr) \le -\theta_1' \tag{2.97}$$

for some constant $\theta_1' > 0$ which depends only on d and p.

Proof. Since U satisfies the property SPA3, the proof is the same as Lemma 2.6. □

Lemma 2.24. *Let $z \in M$. Let $0 < b < 1$ and $c > 0$ be constants. Then for every fixed $\varepsilon > 0$, there exists $R > 0$ such that*

$$\limsup_{n\to\infty} \frac{1}{(nf(z))^{(1-b)/(d+2p)}} \log \Pr\left\{ \sup_{x,y} \left| \frac{L_n(x, y; h)}{n^{(1-p)/d} \, \mathrm{dist}_p(x, y)} - C(d, p) \right| > \varepsilon \right\} \le -\varsigma(d, p; \varepsilon)$$

where the supremum is taken over $x, y \in M$ such that $\mathrm{dist}_1(x, z) < R$, $\mathrm{dist}_1(y, z) < R$, and $\mathrm{dist}_1(x, y) \ge c(nf(z))^{-b/(d+2p)}$.

Recall that $\varsigma(d, p; \varepsilon)$ was defined in (2.60).

Proof. The proof is almost the same as Lemma 2.18. Replace the Euclidean distance with h, and repeat the proof in the opposite direction, i.e., prove the convergence of $L_n(x, y; h)$ from $L_n(x, y)$ as Lemma 2.18 did for $L_n(x, y)$ from $L_n(\varphi(x), \varphi(y))$. □

[5] Take net limit as in Riemann integral.

Theorem 2.25. *Assume that M is compact. Let $0 < b < 1$ and $c > 0$ be constants. Then for every fixed $\varepsilon > 0$,*

$$\limsup_{n\to\infty} (n \inf f)^{(b-1)/(d+2p)} \log \Pr\left\{ \sup_{x,y} \left| \frac{L_n(x,y;h)}{n^{(1-p)/d} \operatorname{dist}_p(x,y)} - C(d,p) \right| > \varepsilon \right\} < -\varsigma(d,p;\varepsilon)$$

where the supremum is taken over x and y such that $\operatorname{dist}_1(x,y) \geq c(n \inf f)^{-b/(d+2p)}$.

The proof is the same as Proposition 2.19 except that we use Lemma 2.24 instead of Lemma 2.18.

Embedded compact manifolds

An important application of Theorem 2.25 is the embedded compact manifold cases. Tenenbaum *et al.* (2000) and Bernstein *et al.* (2000) showed the shortest path length convergence with $p = 1$ when compact manifolds are embedded in Euclidean spaces. The following extends these results.

Proposition 2.26. *Suppose the compact Riemannian manifold M is embedded in Banach space V. Assume that the embedding is smooth and is a Banach space isometry in every tangent fiber. Let h be the norm of V. Then h^p is super-additive in M of order $p > 1$, and $L_n(x,y;h)$ converges as in Theorem 2.25.*

Proof. M is compact hence is finite-dimensional, the second fundamental form $I\!I$ at each point is a continuous bilinear operator. For immersion theory in Banach manifolds, see Lang (1999, Chapter XIV). For every $x \in M$, lift $I\!I$ from $T_xM \otimes T_xM$ to $\mathbb{R}^d \otimes \mathbb{R}^d$, i.e.,

$$\mathbb{R}^d \otimes \mathbb{R}^d \longrightarrow T_xM \otimes T_xM \longrightarrow T_xV \cong V \tag{2.98}$$

so that we have a collection $\mathcal{T}$ of continuous linear maps from $\mathbb{R}^d \otimes \mathbb{R}^d$ to V.[6] Since M is compact and $I\!I$ is smooth, for every $u \in \mathbb{R}^d \otimes \mathbb{R}^d$ its set of images under the linear maps in $\mathcal{T}$ is compact, hence bounded. By a corollary of the Banach-Steinhaus theorem (Rudin 1991, Chapter 2), the linear maps in $\mathcal{T}$ is equicontinuous and there exists a global bound $K > 0$ such that

$$|I\!I(X,Y)| \leq K|X||Y| \tag{2.99}$$

for all tangent vectors X, Y of M.

[6] Choose a frame bundle section of M.

Note that the difference of distances in M and V is determined by $I\!I$. If γ is a curve in M,

$$\ddot{\gamma} = \gamma'' + I\!I(\gamma', \gamma') \tag{2.100}$$

where $\ddot{\gamma}$ and γ'' are the acceleration of γ in V and M, respectively (O'Neill 1983, Chapter 4). Therefore the ratio of geodesic distance in M to norm distance in V is uniformly bounded for close enough pairs. Therefore h satisfies condition SPA3. SPA2 follows from the fact M is compact and embedded. SPA1 is easy to verify. □

Information geometry

See Chapter 3 for parameter spaces and information geometry.

Let M be a compact manifold without boundary. Let $F\colon M \times M \to \mathbb{R}$ be a smooth nonnegative function such that $F(x, x) = 0$ for all $x \in M$. Then $dF(x, x) = 0$. Let $F_x(y) = F(x, y)$ and X, Y be vector fields in M. Define $g_1(X, Y)(x) = XYF_x(x)$. Then it is symmetric and positive semidefinite since

$$0 = [X, Y] \cdot F_x(x) = XYF_x(x) - YXF_x(x). \tag{2.101}$$

Assume that g_1 is strictly positive definite everywhere. Then (M, g_1) is a Riemannian manifold. If F were an information divergence, say Kullback-Leibler divergence, then g_1 is the Fisher information (Amari and Nagaoka 2000, Chapter 3).

By a lemma of Morse there exists a local coordinate system $u = (u_1, \ldots, u_d)$ in a neighborhood U of x such that $u(x) = 0$ and (Milnor 1963, Lemma 2.2)

$$F(x, u) = F_x(u) = u_1^2 + \cdots + u_d^2. \tag{2.102}$$

In particular, the local expression of g_1 is the identity at $u(x) = 0$. Since F is smooth there exists a neighborhood $V \subset U$ such that g_1 deviates from the identity at most by $\varepsilon > 0$ in V, and

$$(1 - \varepsilon)^2 < \frac{\mathrm{dist}_1(x, u)^2}{F(x, u)} < (1 + \varepsilon)^2. \tag{2.103}$$

Since both dist_1 and F are continuous in $M \times M$, there exists some neighborhood W of x such that for all $u, v \in W$,

$$1 - \delta < \frac{\mathrm{dist}_1(u, v)^2}{F(u, v)} < 1 + \delta \tag{2.104}$$

when ε was chosen sufficiently small. Therefore F satisfies the condition SPA3.

Proposition 2.27. *Let M be a compact manifold parameter space. Let ζ be an embedding parameterization, i.e., a smooth embedding map of M into the set of probability measures in some measurable space. Suppose we have a smooth prior probability density function f with respect to the Fisher information in M. Let $L_n(x, y; F)$ denote the shortest path length in F through sampled parameters from f from a parameter $x \in M$ to another parameter $y \in M$. If $\inf_M f > 0$ then Theorem 2.25 holds with $h = F$ and $p = 2$.*

2.4.2 Non-compact complete manifolds

Consider a non-compact but complete manifold M. From Proposition 2.19 one may observe that $L_n(x, y)$ should be close to the minimizing geodesic curve from x to y under g_p. To prove the convergence, however, every neighborhood of x should be inspected until all possible paths are eliminated, i.e., one needs to search in large enough region U such that $L_n(x, u) > L_n(x, y)$ for every u on the boundary of U. It makes Proposition 2.19 difficult to prove in non-compact manifolds since the convergence of Lemma 2.18 should hold in infinitely many neighborhoods at once.

Let $\lambda > 0$ be a constant satisfying

$$1 + \lambda > \frac{C(d, p) + \varepsilon}{C(d, p) - \varepsilon}. \tag{2.105}$$

Proposition 2.28. *Let (M, g_1) be a complete Riemannian manifold. Assume that $f(u) > 0$ for all $u \in M$. Let $K \subset M$ be compact and $f_K = \inf_K f > 0$. Let $0 < b < 1$ and $c > 0$ be constant. Then for every fixed $\varepsilon > 0$,*

$$\limsup_{n\to\infty} (n f_K)^{(b-1)/(d+2p)} \log \Pr\left\{ \sup_{x,y} \left| \frac{L_n(x, y)}{n^{(1-p)/d} \operatorname{dist}_p(x, y)} - C(d, p) \right| > \varepsilon \right\} < -\varsigma(d, p; \varepsilon)$$

where the supremum is taken over x and y in the same connected interior of K such that

(i) $\operatorname{dist}_1(x, y) \geq c(n f_K)^{-b/(d+2p)}$, *and*

(ii) for every $u \notin K$ both $\operatorname{dist}_p(x, u)$ *and* $\operatorname{dist}_p(y, u)$ *are greater than* $(1 + \lambda) \operatorname{dist}_p(x, y)$.

Proof. Since K is compact, there exists a finite open cover with properties in Lemma 2.18. Then by the arguments in the proof of Proposition 2.19 applied to some finite open cover of K, with high probability,

$$n^{(p-1)/d} L_n^K(x, u) > (C(d, p) - \varepsilon) \operatorname{dist}_p(x, u) \tag{2.106}$$

for all x in the interior of K and for all u on the boundary of K. L_n^K denotes the shortest path length when sample points are restricted to K. Since $L_n^K \geq L_n$, any shortest paths that exits K from x should have length greater than

$$(C(d,p)-\varepsilon)\operatorname{dist}_p(x,u) \geq (C(d,p)-\varepsilon)(1+\lambda)\operatorname{dist}_p(x,y) \tag{2.107}$$

$$> (C(d,p)+\varepsilon)\operatorname{dist}_p(x,y) \tag{2.108}$$

for every x, y in consideration by the choice of λ in (2.105). □

Corollary 2.29. *Let $x, y \in M$. Then for some $f_0 > 0$,*

$$\limsup_{n\to\infty} \frac{1}{(nf_0)^{1/(d+2p)}} \log \Pr\left\{ \left| \frac{L_n(x,y)}{n^{(1-p)/d}\operatorname{dist}_p(x,y)} - C(d,p) \right| > \varepsilon \right\} < -\varsigma(d,p;\varepsilon). \tag{2.109}$$

Proof. A bounded and closed subset in M is compact by Hopf and Rinow theorem (do Carmo 1992, Chapter 7). Choose $K = K_x \cup K_y$,

$$K_x = \{u \in M : \operatorname{dist}_p(x,u) \leq (1+\lambda)\operatorname{dist}_p(x,y)\}, \tag{2.110}$$

$$K_y = \{u \in M : \operatorname{dist}_p(y,u) \leq (1+\lambda)\operatorname{dist}_p(x,y)\} \tag{2.111}$$

for Proposition 2.28. □

2.5 Future works

We describe some preliminary results which needs further study in the future.

2.5.1 Manifolds with boundary

Consider the case when a compact manifold M has boundary.

Go back to Euclidean spaces and let $(u_1, \ldots, u_d)$ denote the coordinates of $\mathbb{R}^d$. Consider the half Poisson point process $\mathcal{H}_\lambda^+$ in $\mathbb{R}^d$ where no point exists in the region $u_d < 0$. Let $\mathcal{L}_\lambda^+(0,t)$ denote the power-weighted shortest path length from $(0,0)$ to $(t,0) \in \mathbb{R}\times\mathbb{R}^{d-1} \cong \mathbb{R}^d$ for $t > 0$.

Lemma 2.30.

$$\lim_{t\to\infty} \frac{1}{t} E\mathcal{L}_1^+(0,t) = C_b(d,p) \tag{2.112}$$

for some constant $C_b(d,p) > 0$ which depends only on d and p.

The proof is the same as Lemma 2.8. The new constant $C_b(d,p)$ plays the role of $C(d,p)$ when the paths run on the boundary. Obviously there is inequality $C(d,p) \le C_b(d,p)$ but at this point we do not know whether the inequality is strict or not. We conjecture that $C_b(d,p) = C(d,p)$, since if otherwise discontinuities of $L_n(x,y)$ between the boundary and the interior is introduced but $L_n(x,y)$ is continuous for all finite n.

Proofs in Section 2.3 can be reproduced for boundary shortest paths till Theorem 2.16. In Poisson processes, replace $\mathcal{H}_\lambda$ with $\mathcal{H}_\lambda^+$ and keep x_λ, y_λ on the hyperplane $u_d = 0$. In binomial processes, replace the uniform distribution in $B(z;R_2)$ with the uniform distribution in the upper hemisphere $\{u_d \ge 0\} \cap B(z;R_2)$, and keep x_n, y_n on the hyperplane $u_d = 0$.

Suppose that $C_b(d,p) = C(d,p)$. In that case, the same proofs may be repeated to the end of Section 2.3 and we may conclude that Proposition 2.19 holds with compact manifolds with boundary.

A possible complication occurs if $C_b(d,p) > C(d,p)$. If the inequality were strict, then the metric g_p must be scaled by $\big(C_b(d,p)/C(d,p)\big)^2$ in the boundary so that dist_p reflects the difference in the constants. Since shortest paths are continuous, the scaling should be applied to places near the boundary so that continuity holds between $C(d,p)$ and $C_b(d,p)$. In that case, we may conclude that with exponentially high probability in the sense of Proposition 2.19,

$$C(d,p) - \varepsilon < \frac{L_n(x,y)}{(n \inf f)^{(1-p)/d} \, \mathrm{dist}_p(x,y)} < C_b(d,p) + \varepsilon. \tag{2.113}$$

2.5.2 Conformal deformations in anisotropic diffusion maps

In this subsection, we discuss the anisotropic diffusion map (Coifman and Lafon 2006) and the conformal deformations. Let (M, g_1) denote a compact Riemannian manifold.

Define the diffusion kernel in $\mathbb{R}^d$, $x, y \in \mathbb{R}^d$,

$$\kappa_t(x,y) = \exp\left(-\frac{|x-y|^2}{4t}\right) \tag{2.114}$$

where $t > 0$ is the kernel width. For $\alpha > 0$, anisotropic diffusion map builds a Markov chain over the sample points with transition probability

$$p_{ij}(t) = b_i(t) \frac{\kappa_t(X_i, X_j)}{\big(\sum_k \kappa_t(X_i, X_k)\big)^\alpha \big(\sum_k \kappa_t(X_j, X_k)\big)^\alpha} \tag{2.115}$$

where $b_i(t) > 0$ is a scalar such that $\sum_j p_{ij} = 1$ for $i = 1, 2, \ldots, n$.

When $t \to 0$, the backward infinitesimal generator $\mathcal{H}_b^{(\alpha)}$ of the anisotropic diffusion map

is (Nadler, Lafon, Coifman, and Kevrekidis 2006, Section 4)

$$\mathcal{H}_b^{(\alpha)}\psi = \Delta_1\psi + 2(1-\alpha)\,\mathrm{d}\,\psi(\nabla_1 \log f), \tag{2.116}$$

where $\psi: M \to \mathbb{R}$ is an arbitrary smooth function, d denotes the exterior derivative, Δ_1 and ∇_1 are the Laplace-Beltrami and the gradient operator with respect to the base metric g_1, respectively. See Morita (2001) for more details. The infinitesimal generator in the limit is Δ_1 when $\alpha = 1$. For Laplacian Eigenmaps, the graph Laplacian converges $f\mathcal{H}_b^{(0)}$ (Belkin and Niyogi 2008, Theorem 5.1).

Recall the definition $g_p = f^{2(1-p)/d} g_1$. If Δ_p denotes the Laplace-Beltrami operator under g_p where p is allowed to be any real value, (Besse 1987, Theorem 1.159)

$$\Delta_p\psi = f^{2(p-1)/d}\left(\Delta_1\psi + \frac{(d-2)(1-p)}{d}\,\mathrm{d}\,\psi(\nabla_1 \log f)\right). \tag{2.117}$$

Then for $d > 2$, (2.116) and (2.117) are equivalent when

$$p = 1 - \frac{2d(1-\alpha)}{d-2}. \tag{2.118}$$

Therefore the backward generator of the anisotropic diffusion map is conformal to the generator of the standard Wiener process under g_p. The same argument may be repeated for the forward infinitesimal generator $\mathcal{H}_f^{(\alpha)}$ in Nadler *et al.* (2006), and they may be summarized

$$\mathcal{H}_b^{(\alpha)}\psi = f^{4(1-\alpha)/(d-2)}\Delta_p\psi = f^{2(1-p)/d}\Delta_p\psi, \tag{2.119}$$

$$\mathcal{H}_f^{(\alpha)}\psi = f^{1-2\alpha}\mathcal{H}_b^{(\alpha)}(f^{2\alpha-1}\psi) = f^{-p}\Delta_p(f^{2\alpha-1}\psi), \tag{2.120}$$

and the two infinitesimal generators are similar to each other. In other words,

$$\mathcal{H}_b^{(\alpha)}\colon \psi \xmapsto{\Delta_p} \Delta_p\psi \xmapsto{f^{2(1-p)/d}} f^{2(1-p)/d}\Delta_p\psi, \tag{2.121}$$

$$\mathcal{H}_f^{(\alpha)}\colon \psi \xmapsto{f^{2\alpha-1}} f^{2\alpha-1}\psi \xmapsto{\mathcal{H}_b^{(\alpha)}} \mathcal{H}_b^{(\alpha)}(f^{2\alpha-1}\psi) \xmapsto{f^{1-2\alpha}} f^{1-2\alpha}\mathcal{H}_b^{(\alpha)}(f^{2\alpha-1}\psi). \tag{2.122}$$

Note that if the metric deformation were uniform over the space, i.e., if f were a constant function, then $\Delta_p\psi = f^{2(p-1)/d}\psi$. In the context, the multiplicative factor $f^{2(1-p)/d}$ after Δ_p in $\mathcal{H}_b^{(\alpha)}$ may be interpreted as a point-wise normalization.

Note that from (2.118), the effective p for anisotropic diffusion map is less than one, while the power-weighted shortest paths restrict $p > 1$. Therefore these two methods are, in some sense, dual to each other, and works in separate regimes: the power-weighted shortest

paths for g_p, $p > 1$ and the anisotropic diffusion maps for g_p, $p < 1$.

Chapter 3
Information Geometric Curves

3.1 Introduction

This chapter considers curves in information geometry. Time-parametrized probability distributions appear frequently in time-series analysis. In information theory, probability distributions are measured their discrepancy by information divergence such as Kullback-Leibler divergence. Information divergence is pairwise dissimilarity and it does not extend naturally when probability distributions are time-parameterized.

We use information geometry, a differential geometric framework, to the space of probability distributions and parameter spaces. And time-parameterized distributions become curves in the location parameter space. We propose a novel measure of dissimilarity between the curves in the information geometry, the surface area of the minimal surface that contains the two time-parameterized probability distributions. The surface area measure is stable under re-parameterizations in time since curves in geometry are considered equivalent when their images are the same.

In information geometry, the Fisher information, a positive definite bilinear form, forms a Riemannian metric in the parameter space. We examine the immersion maps between the parameter space and L^p spaces behind the theories of information geometry. We establish relations between the surface area measure and common curve comparison measures such Hausdorff distance and Chamfer distance.

3.2 Preliminary

This introductory section provides basic definitions and discussions. Throughout the chapter, Ω denotes a measurable space and all measures are assumed to be supported in Ω.

3.2.1 Information divergence

Let μ and ν be probability measures. Lebesgue-Radon-Nikodym theorem states that ν is the sum of two positive measures ν_μ and $\nu_\perp$ where ν_μ is absolutely continuous with respect to μ and $\nu_\perp$ is mutually singular to μ (Rudin 1986, Theorem 6.10). Furthermore there exists a unique function $d\nu/d\mu$ which is integrable with respect to μ and

$$\int \frac{d\nu}{d\mu} 1_E d\mu = \nu_\mu(E) \tag{3.1}$$

for every measurable $E \subset \Omega$, and 1_E is the indicator function.

Many studies in information theory assume all probability measures of interest are absolutely continuous to each other. However, non-absolutely continuous probability measures appear both in theory and in applications. For an example in theory, uniform distributions in unit intervals are not absolutely continuous to each other unless they are identical to each other. For an example in application, bag-of-words approach is common in document retrieval problems (Dhillon and Modha 2001; Teh, Jordan, Beal, and Blei 2006). This model represents documents by multinomial probability measures of select words, and of course different documents have different vocabularies in general. See Bag-of-Words data from UCI repository (Frank and Asuncion 2010). Another example arises in density estimation problems of high dimensional data. If the number of sample points is few relative to the dimensionality, probability density estimation can no longer assume a common support of the measures. For instance, the data may be supported in some embedded manifolds (Tenenbaum *et al.* 2000). See Schölkopf, Platt, Shawe-Taylor, Smola, and Williamson (2001) for support estimation problem in high dimensions. Also see Ledoit and Wolf (2004) and Chen, Wiesel, and Hero (2011) for degenerate covariance problems.

Example 3.1. Let $\Omega = [0,1]$, μ be the uniform distribution in $[0,1]$, and ν be the standard Gaussian distribution in $\mathbb{R}$. Then ν is not absolutely continuous with respect to μ. ν_μ is the truncated Gaussian distribution in $[0,1]$, $\nu_\perp$ is the truncated Gaussian distribution outside $[0,1]$,

$$\frac{d\nu}{d\mu}(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, & x \in [0,1], \\ 0, & x \notin [0,1]. \end{cases} \tag{3.2}$$

On the other hand, μ is absolutely continuous with respect to ν,

$$\frac{d\mu}{d\nu}(x) = \begin{cases} \sqrt{2\pi}e^{\frac{1}{2}x^2}, & x \in [0,1], \\ 0, & x \notin [0,1]. \end{cases} \tag{3.3}$$

In information theory, *Kullback-Leibler (KL) divergence* D_{KL} is a common dissimilarity measure, and the dissimilarity of ν with respect to μ is (Kullback and Leibler 1951; Cover and Thomas 2006)

$$D_{KL}(\mu \parallel \nu) = \int \frac{d\nu}{d\mu} \log \frac{d\nu}{d\mu} d\mu = \int \log \frac{d\nu}{d\mu} d\nu. \tag{3.4}$$

The base of the logarithm is arbitrary but fixed.

KL divergence derives from Shannon entropy which is defined from several postulates motivated in communication theory (Shannon 1948; Cover and Thomas 2006, Chapter 2). Define

$$I_\alpha(\mu \parallel \nu) = \int \left(\frac{d\nu}{d\mu} \right)^\alpha d\mu. \tag{3.5}$$

Rényi (1961) showed that a class of information divergences emerges if one of the postulates is generalized. This class of information divergences, indexed by $\alpha > 0$, $\alpha \neq 1$ is called Rényi's divergence

$$D_{R,\alpha}(\mu \parallel \nu) = \frac{1}{\alpha - 1} \log \left(\frac{\int (d\nu/d\mu)^\alpha \, d\mu}{I_1(\mu \parallel \nu} \right). \tag{3.6}$$

Another generalization of KL divergence is α-divergence for $\alpha > 0$, $\alpha \neq 1$,[1]

$$D_\alpha(\mu \parallel \nu) = \frac{1}{\alpha(1-\alpha)} \left(I(\mu \parallel \nu) - \int \left(\frac{d\nu}{d\mu} \right)^\alpha d\mu \right). \tag{3.7}$$

f-divergence

The main idea behind information divergence is to measure dissimilarity by the deviation of $d\nu/d\mu$ from $d\mu/d\mu = 1$. KL divergence (3.4) and α divergence (3.7) fall into the class of f-divergence (Ali and Silvey 1966; Csiszár 1967; Csiszár and Shields 2004; Cover and Thomas 2006). For a (continuous) convex function $f: [0, \infty) \to \mathbb{R}$ with $f(1) = 0$, the

[1] The definition differs from others in the literature, e.g., see Amari and Nagaoka (2000). Also D_α is sometimes parameterized in a different way from (3.6). The rationale of this choice will become clear in subsequent discussions.

f-divergence D_f is defined by

$$D_f(\mu \parallel \nu) = \int f\left(\frac{d\nu}{d\mu}\right) d\mu. \tag{3.8}$$

For KL divergence, $f(u) = u \log u$, and for α divergence,[2]

$$f(u) = \frac{1}{\alpha(1-\alpha)}(u - u^\alpha). \tag{3.9}$$

D_f is non-negative by Jensen's inequality when ν is absolutely continuous with respect to μ by the condition $f(1) = 0$. If ν is not absolutely continuous, the problem becomes complicated. Ali and Silvey (1966) defines generalized expectation to compensate the singular part of ν and $D_f(\mu \parallel \nu)$ is assigned infinite value or some correction term is introduced based on the behavior of $f(t)$ when $t \to \infty$.

Our approach is to ignore the singular part, and define *f-variation*

$$\mathcal{D}_f(\mu \parallel \nu) = \int f\left(\frac{d\nu}{d\mu}\right) d\mu - f\left(\int \frac{d\nu}{d\mu}\, d\mu\right). \tag{3.10}$$

A drawback of $\mathcal{D}_f$ is that it cannot measure singular differences. If ν is purely singular to μ and $d\nu/d\mu = 0$ almost everywhere, then $\mathcal{D}_f(\mu \parallel \nu) = 0$.

Let us define f-variation versions of divergences,

$$\mathcal{D}_{KL}(\mu \parallel \nu) = \int \frac{d\nu}{d\mu} \log \frac{d\nu}{d\mu}\, d\mu - I_1(\mu \parallel \nu) \log I_1(\mu \parallel \nu), \tag{3.11}$$

$$\mathcal{D}_\alpha(\mu \parallel \nu) = I_1(\mu \parallel \nu)^\alpha - \int \left(\frac{d\nu}{d\mu}\right)^\alpha d\mu, \tag{3.12}$$

where $I_1(\mu \parallel \nu)$ is defined in (3.5). For KL variation, still $f(u) = u \log u$. For α variation, $f(u) = -u^\alpha$.

Kullback-Leibler divergence

The definition of D_{KL} varies in literature. Another common definition different from (3.4) is

$$D_{KL*}(\mu \parallel \nu) = -\int \log \frac{d\nu}{d\mu} d\mu. \tag{3.13}$$

[2]There are many different ways that f has been defined for α divergence. For example, Zhang (2004) used $f(u) = (1 - \alpha + \alpha u + u^\alpha)/(\alpha(1-\alpha))$.

Two definitions (3.4) and (3.13) agree when μ and ν are absolutely continuous with respect to each other.

Suppose ν is *not* absolutely continuous with respect to μ, i.e., $I_1(\mu \parallel \nu) < 1$. Then by Jensen's inequality,

$$D_{KL}(\mu \parallel \nu) \geq I_1(\mu \parallel \nu) \log I_1(\mu \parallel \nu) \tag{3.14}$$

and the equality holds when $d\nu/d\mu$ is constant μ-almost everywhere. Hence a drawback of definition (3.4) is that it may become negative. This is fixed in (3.11).

Definition (3.13) does not have the problem of being negative. This follows from another application of Jensen's inequality. However, there is a trade-off. Suppose that there exists some measurable $E \subset \Omega$ such that $\mu(E) > 0$ and $\nu_\mu(E) = 0$, i.e., μ is not absolutely continuous with respect to ν_μ. Then the value of $D_{KL*}(\mu \parallel \nu)$ is infinite.

One may prefer (3.13) to (3.4) due to its non-negativity. However, one should note that Rényi's divergence and α-divergence are closely connected to (3.4). Assume that the integral

$$\int \left(\frac{d\nu}{d\mu}\right)^{\alpha} \log\left(\frac{d\nu}{d\mu}\right) d\mu \tag{3.15}$$

is finite for $\alpha = 1 \pm \varepsilon$, $0 < \varepsilon < 1$. Note that if $1 - \varepsilon < \alpha < 1 + \varepsilon$,

$$\left(\frac{d\nu}{d\mu}\right)^{\alpha} \leq \left(\frac{d\nu}{d\mu}\right)^{1-\varepsilon} + \left(\frac{d\nu}{d\mu}\right)^{1+\varepsilon}. \tag{3.16}$$

Then by Lang (1993, Lemma 2.2, p. 226),

$$\lim_{\alpha \to 1} \frac{I_1(\mu \parallel \nu) - I_\alpha(\mu \parallel \nu)}{1 - \alpha} = D_{KL}(\mu \parallel \nu). \tag{3.17}$$

Note that the limit is (3.4) not (3.13). From definition (3.7)

$$\lim_{\alpha \to 1} D_\alpha(\mu \parallel \nu) = \lim_{\alpha \to 1} \frac{I_1(\mu \parallel \nu) - I_\alpha(\mu \parallel \nu)}{\alpha(1 - \alpha)} = D_{KL}(\mu \parallel \nu). \tag{3.18}$$

Similar steps show that

$$\lim_{\alpha \to 1} \frac{1}{1 - \alpha} \mathcal{D}_\alpha(\mu \parallel \nu) = \mathcal{D}_{KL}(\mu \parallel \nu). \tag{3.19}$$

The same holds for Rényi divergence,

$$\lim_{\alpha \to 1} D_{R,\alpha}(\mu \parallel \nu) = \frac{D_{KL}(\mu \parallel \nu)}{I_1(\mu \parallel \nu)} \tag{3.20}$$

when $I_1(\mu \parallel \nu) > 0$.

Shannon entropy

Suppose that the measurable space Ω is discrete and finite, hence a probability measure μ may be expressed by probability masses p_i for $i = 1, 2, \ldots, n$, $\sum_i p_i = 1$. Shannon entropy is defined as (Shannon 1948)

$$H(\mu) = -K \sum_{i=1}^{n} p_i \log p_i \tag{3.21}$$

where K is a positive constant.

When Ω is continuous, the Shannon entropy can be defined through *differential entropy* when the probability measure is Lebesgue continuous (Cover and Thomas 2006, Chapter 8). Let $\Omega = \mathbb{R}^d$ for now and let m denote the Lebesgue measure. If μ is absolutely continuous with respect to m, the differential entropy of μ is

$$H_m(\mu) = -\int \frac{d\mu}{dm} \log \frac{d\mu}{dm} \, dm \tag{3.22}$$

If $K = 1$ in (3.21) then Shannon entropy H is differential entropy H_m with m being the counting measure.

KL divergence—often called relative entropy—is a special case of differential entropy where the base measure m is always some probability measure, i.e.,

$$D_{KL}(\mu \parallel \nu) = -H_\mu(\nu). \tag{3.23}$$

The formulation of KL divergence has several advantages over differential entropy. For example, the maximum entropy distribution is well defined. In $\mathbb{R}^d$, the probability distribution that maximizes H_m with prescribed covariance is Gaussian distribution (Cover and Thomas 2006, Theorem 8.6.5). However, this upper bound does not extend to general measurable space Ω since the definitions of Gaussian distribution and covariance are not well defined outside the Euclidean spaces. Futhermore, there is no maximum entropy distribution without the covariance constraint. On the other hand, if the base measure is another probability measure, i.e., a finite measure where constant functions are integrable,

then the maximum entropy distribution exists and it is the base measure itself, possibly with a normalization. Therefore we may pick a distribution that is *most random* in a view.

Remark 3.2. Radon-Nikodym derivative $dv/d\mu$ is a real measurable function. Therefore $dv/d\mu$ defines a real-valued non-negative random variable $V_v = V_v^\mu$, the *variation* of v with respect to μ, under the probability law μ. Note that $d\mu/d\mu = 1$ almost everywhere, and the it becomes how to measure the difference of V_v from constant function 1. Then

$$I_\alpha(\mu \parallel v) = E_\mu|V_v|^\alpha, \tag{3.24}$$

$$I_1(\mu \parallel v) = E_\mu|V_v| = \|v_\mu\|, \tag{3.25}$$

$$D_{KL}(\mu \parallel v) = E_\mu(V_v \log V_v) = -H_\mu(V_v), \tag{3.26}$$

$$D_{R,\alpha}(\mu \parallel v) = \frac{1}{\alpha - 1} \log \frac{E_\mu|V_v|^\alpha}{E_\mu|V_v|}, \tag{3.27}$$

$$D_\alpha(\mu \parallel v) = \frac{1}{\alpha(1-\alpha)}(E_\mu|V_v| - E_\mu|V_v|^\alpha), \tag{3.28}$$

where E_μ denotes expectation under μ. The right side for $D_{R,\alpha}$ is the definition of Rényi entropy of V_v (Rényi 1961).

3.2.2 L^p space

This section discusses Banach space and L^p space theory to be used in this chapter.

The space of (finite) signed measures $\mathcal{M}$ is a Banach space with total variation norm,

$$\|\mu\| = \sup \sum_{i=1}^{n} |\mu(E_i)| \quad \text{for } \mu \in \mathcal{M} \tag{3.29}$$

where the supremum runs over all finite partitions $\{E_1, \ldots, E_n\}$ into measurable subsets.

Define its *tangent bundle* $T\mathcal{M} = \mathcal{M} \times \mathcal{M}$. Each subspace

$$T_\mu\mathcal{M} = \{(\mu, \omega) \in T\mathcal{M} : \omega \in \mathcal{M}\} \tag{3.30}$$

is called the *(tangent) fiber* at μ, and each element (μ, ω) is called a *tangent vector* at μ. Note that a fiber is a vector space. A tangent vector (μ, ω) represents the directional derivative at μ in the direction of ω, i.e., if $f : \mathcal{M} \to \mathbb{R}$ is a differentiable function, then the tangent vector acts on f as

$$(\mu, \omega) \cdot f = \lim_{r \to 0} \frac{f(\mu + r\omega) - f(\mu)}{r}. \tag{3.31}$$

Often ω is also called a tangent vector. A vector field $X\colon \mathcal{M} \to T\mathcal{M}$ is a differentiable function such that $X_\mu \in T_\mu\mathcal{M}$. The subscript notation is standard in differential geometry. For more details and the definitions in manifolds, see Morita (2001, Chapter 1) or Lee (2003, Chapter 3).

Fix a probability measure $\mu \in \mathcal{M}$. Define a collection of real-valued measurable functions

$$L^p = L^p(\mu) = \left\{ f\colon \int |f|^p \, d\mu < \infty \right\} \tag{3.32}$$

for $p > 0$. L^p is a Banach space for $1 \le p < \infty$ (Rudin 1986, Chapter 3).

A probability measure μ induces a Banach space projection by the Radon-Nikodym derivative, and also induces a sequence of mappings $\nu \in \mathcal{M} \mapsto \nu_\mu \mapsto d\nu/d\mu \in L^1(\mu)$. Recall that ν_μ is the absolutely continuous part of ν with respect to μ. This sequence extends for $0 < \alpha \le 1$,

$$\nu \longmapsto \nu_\mu \longmapsto \frac{d\nu}{d\mu} \longmapsto \left(\frac{d\nu}{d\mu}\right)^\alpha \in L^{1/\alpha}(\mu). \tag{3.33}$$

Since μ is a probability measure, every $L^{1/\alpha}(\mu)$-function f is also integrable, i.e., is a $L^1(\mu)$-function, and the identity map $f \in L^{1/\alpha}(\mu) \mapsto f \in L^1(\mu)$ is linear and injective, but not necessarily continuous. In particular it is never continuous when $\alpha = 0$ and $f \in L^\infty(\mu)$.

Even though the identity map may be discontinuous, the spaces $L^{1/\alpha}(\mu)$, $0 < \alpha \le 1$ are homeomorphic to each other. Let $0 < \alpha \le 1$ and let $f \in L^1(\mu)$. Define $T_\alpha\colon L^1(\mu) \to L^{1/\alpha}(\mu)$,

$$T_\alpha f(x) = \begin{cases} |f(x)|^{\alpha-1} f(x), & \text{if } |f(x)| \neq 0, \\ 0, & \text{if } |f(x)| = 0. \end{cases} \tag{3.34}$$

A direct computation shows that $\|T_\alpha f\| = \|f\|^\alpha$, hence T_α is continuous. Define the inverse map $T_\alpha^{-1}\colon L^{1/\alpha}(\mu) \to L^1(\mu)$,

$$T_\alpha^{-1} g(x) = |g(x)|^{(1-\alpha)/\alpha} g(x). \tag{3.35}$$

The inverse relationship $\|T_\alpha^{-1} g\| = \|g\|^{1/\alpha}$ shows that T_α^{-1} is also continuous.

Note that the transformation T_α, $\alpha = 2^{-1}$ maps $L^1(\mu)$-functions into $L^2(\mu)$-functions, which is a Hilbert space. Of course, the map is not linear thus one cannot endow an inner product structure to $L^1(\mu)$. However, it is possible to endow a symmetric bilinear form in each tangent fiber of $\mathcal{M}$ in $T\mathcal{M}$.

Assume $\nu \in \mathcal{M}$ is a probability measure. Let (ν, ω) be a tangent vector at ν. Note that ω

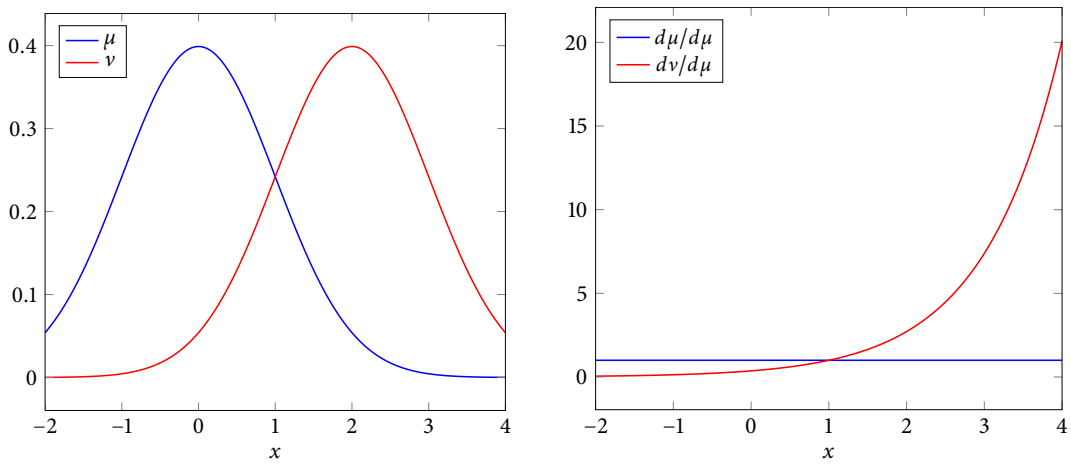


Figure 3.1: An example when μ and ν are Gaussian distributions in $\Omega = \mathbb{R}$. Radon-Nikodym derivatives with respect to μ are plotted on the right.

is a signed measure in $\mathcal{M}$. Assume that

$$\int \left|\frac{d\omega}{d\nu}\right|^{1/\alpha} d\nu < \infty, \tag{3.36}$$

i.e., $d\omega/d\nu$ is in $L^{1/\alpha}(\nu)$. This is also called the α^{-1} *variation* of ω (Leonard and Sundaresan 1974). Define $T_{\alpha *}\colon \mathcal{M} \to L^{1/\alpha}(\mu)$ by for every $x \in \Omega$,

$$T_{\alpha *}\omega(x) = \begin{cases} \left|\frac{d\nu}{d\mu}(x)\right|^{\alpha} \frac{d\omega}{d\nu}(x), & \text{if } \frac{d\nu}{d\mu}(x) \neq 0, \\ 0, & \text{if } \frac{d\nu}{d\mu}(x) = 0. \end{cases} \tag{3.37}$$

Figure 3.1 has an example for $d\nu/d\mu$. To see it is indeed in $L^{1/\alpha}(\mu)$,

$$\begin{aligned} \int |T_{\alpha *}\omega|^{1/\alpha} \, d\mu &= \int \left|\frac{d\omega}{d\nu_\mu}\right|^{1/\alpha} \left|\frac{d\nu}{d\mu}\right| d\mu \\ &= \int \left|\frac{d\omega}{d\nu_\mu}\right|^{1/\alpha} d\nu_\mu < \infty \end{aligned} \tag{3.38}$$

by the assumption (3.36) and $d\nu/d\mu$ being nonnegative. Recall that ν_μ is the absolutely continuous part of ν with respect to μ. ν_μ appears instead of ν since $T_{\alpha *}\omega$ is defined to be zero where $d\nu/d\mu = 0$. If $\nu = \mu$, or ν were absolutely continuous respect to μ, then ν_μ may be replaced with ν.

Note that $T_{\alpha *}$ is linear in each fiber but not injective since $T_{\alpha *}\omega = T_{\alpha *}\eta$ if two tangent

measures ω and η different only for x where $d\nu/d\mu = 0$. Measure theoretically speaking, the singular part of ω with respect to μ and ν is thrown away. By Hölder's inequality, if ω and η are two tangent measures at ν, $d\omega/d\nu \in L^{1/\alpha}(\nu)$, $d\eta/d\nu \in L_{1-\alpha}(\nu)$, then the integral

$$\int (T_{\alpha *}\omega)(T_{(1-\alpha)*}\nu)\, d\mu = \int \frac{d\omega}{d\nu_\mu}\frac{d\eta}{d\nu_\mu}\, d\nu_\mu \tag{3.39}$$

defines a bilinear form. If $\alpha = 2^{-1}$, then it is a symmetric bilinear form.

We define the positive semidefinite bilinear form

$$\langle \omega, \eta \rangle = \int \frac{d\omega}{d\nu_\mu}\frac{d\eta}{d\nu_\mu}\, d\nu_\mu \tag{3.40}$$

for all ω, η satisfy $d\omega/d\nu, d\eta/d\nu$ in $L^2(\nu)$.

3.3 Information geometry

This section introduces the theory of information geometry. Amari and Nagaoka (2000) and Kass and Vos (1997) already provided introductions of fundamentals and discussions of many related topics in information geometry. This section differs in the that (i) we avoid the use of a common base measure such as Lebesgue measure, and (ii) we use L^p spaces to formalize the mathematical concepts. An introduction of Banach space manifolds may be found in Lang (1999).

3.3.1 Parameterization

The space of signed measures $\mathcal{M}$ is too large and inadequate for the techniques in differential geometry. For instance, suppose that $\Omega = [0,1]$ and μ is uniform distribution in $[0,1]$. Fix $0 < \varepsilon < 1$. If $H \in \mathcal{M}$,

$$\frac{dH}{d\mu}(x) = \begin{cases} \varepsilon^{-1}, & \text{if } x < \varepsilon^2, \\ 0, & \text{otherwise,} \end{cases} \tag{3.41}$$

then $\mu + H$ is not a positive measure any more even though $\|H\| = \varepsilon$, an arbitrarily small value. It makes certain differentiable conditions difficult. For example, the sequence (3.33) is not smooth in general. The observation above leads to the idea of parameterization.

A *parameter space* M is a finite-dimensional connected manifold with *parameterization map* $P\colon M \to \mathcal{M}$ into probability measures in $\mathcal{M}$. We denote $P(\theta)$, $\theta \in M$ by P_θ. Parameterized probability measures satisfy following condition.

PAR1 The map $P: M \to \mathcal{M}$ is continuous.

PAR2 For every $E \subset \Omega$, the map $\theta \in M \mapsto P_\theta(E)$ is infinitely many times differentiable.

Example 3.3 (Amari and Nagaoka (2000, Chapter 2))**.** Consider Gaussian distributions in $\Omega = \mathbb{R}$. Since a Gaussian distribution over $\mathbb{R}$ is uniquely characterized by its mean and variance, a possible parameterization is $M = \mathbb{R} \times (0, \infty)$,

$$P: (\mu, \sigma^2) \in M \longmapsto N(\mu, \sigma^2). \tag{3.42}$$

Remark 3.4. The map $\theta \in M \mapsto P_\theta(E)$ is required to be of $C^\infty(M)$-class for each $E \subset \Omega$ but $P: M \to \mathcal{M}$ is not required to be differentiable since the latter condition may exclude some simple examples. For example, let $\Omega = \mathbb{R}$, $M = \mathbb{R}$, and P maps $\theta \in \mathbb{R}$ to the uniform distribution in $[\theta, \theta + 1]$. Then P is not differentiable in total variation norm. See Hamilton (1982, Part 1) for more examples.

3.3.2 Tangent bundle of parameter space

In addition to PAR1 and PAR2, parameterized measures are assumed to satisfy the following conditions.

PAR3 For every tangent vector $X_\theta \in T_\theta M$, the collection of derivatives $\{X_\theta \cdot P_\theta(E): E \subset \Omega\}$ form a signed measure, i.e., for every disjoint $E_1, E_2, \cdots \subset \Omega$, $E = \bigcup_i E_i$,

$$X_\theta \cdot P_\theta(E) = \sum_{i=1}^{\infty} X_\theta \cdot P_\theta(E_i). \tag{3.43}$$

$X_\theta P_\theta$ will denote the signed measure (3.43), and define $XP: \theta \in M \mapsto X_\theta P_\theta \in \mathcal{M}$.

PAR4 For every $\theta \in M$ there exists an open neighborhood $U \subset M$ of θ such that if $\theta' \in U$ and $X_{\theta'} \in T_{\theta'} M$ then

$$\frac{dX_{\theta'} P_{\theta'}}{dP_\theta} = 0 \quad \text{if and only if} \quad X_{\theta'} = 0. \tag{3.44}$$

PAR5 In the neighborhood U above, every differential measure $X_{\theta'} P_{\theta'}$ satisfy the condition (3.36)

$$\sup \sum_i \left| \frac{X_{\theta'} P_{\theta'}(E_i)}{P_{\theta'}(E_i)} \right|^{1/\alpha} P_{\theta'}(E_i) < \infty \tag{3.45}$$

where the supremum is taken over all finite partitions $\{E_i\}$ of Ω.

PAR4 implies that there is no ambiguity in the tangent directions when $U \subset M$ is mapped into $L^1(P_\theta)$ by Radon-Nikodym derivative.

Proposition 3.5. *Signed measure $X_\theta P$ is absolutely continuous with respect to P_θ. In particular*

$$\int \frac{dX_\theta P_\theta}{dP_\theta}\, dP_\theta = 0 \quad \textit{and} \quad \frac{dX_\theta P_\theta}{dP_\theta}\frac{dP_\theta}{dP_{\theta'}} = \frac{dX_\theta P_\theta}{dP_{\theta'}} \tag{3.46}$$

for any $\theta' \in M$.

Proof. Assume to the contrary that there exists $E \subset \Omega$ where $X_\theta P_\theta(E) \neq 0$ and $P_\theta(E) = 0$. In other words, a $C^\infty(M)$-function $\theta \mapsto P_\theta(E)$ is zero and it has a non-zero partial derivative at θ. Then at some point θ' near θ, $P_{\theta'}(E)$ should map to a negative value. However, a probability measure must not take a negative value. Therefore $X_\theta P_\theta(E) = 0$ for all E satisfying $P_\theta(E) = 0$. □

Define a positive definite bilinear form by (3.40) in each tangent fiber $T_\theta M$,

$$\langle X_\theta, Y_\theta \rangle = \int \frac{dX_\theta P_\theta}{dP_\theta}\frac{dY_\theta P_\theta}{dP_\theta}\, dP_\theta \tag{3.47}$$

for $\theta \in M$ and $X_\theta, Y_\theta \in T_\theta M$. The form is strictly positive definite by PAR4. Note that (3.47) is equal to the Fisher information. If M were an open subset of $\mathbb{R}^d$, $\theta = (\theta_1, \ldots, \theta_d)$, $X_\theta = \partial/\partial\theta_i$, $Y_\theta = \partial/\partial\theta_j$, and if P_θ were Lebesgue-continuous so that p_θ is a probability density function with respect to the Lebesgue measure, then we have the familiar expression

$$\left\langle \frac{\partial \log p_\theta}{\partial \theta_i}, \frac{\partial \log p_\theta}{\partial \theta_j} \right\rangle = E_\theta\left[\frac{\partial \log p_\theta}{\partial \theta_i}\frac{\partial \log p_\theta}{\partial \theta_j}\right]. \tag{3.48}$$

3.3.3 Parallel transport

Tangent fibers have inner products by the Fisher information (3.47). The next step is to construct connections between the fibers. This allows us to measure how curved the space M is, and we will see the relationship between α connection (Amari and Nagaoka 2000, Chapter 3) and $L^{1/\alpha}(P_\theta)$.

Let X, Y be vector fields in M. Consider a curve $\gamma\colon [-1,1] \to M$ such that $\gamma(0) = \theta \in M$ and $\gamma'(0) = X_\theta \in T_\theta M$. It is *not* possible in general to define the derivative of Y in the direction X_θ by

$$\lim_{t\to 0} \frac{Y_{\gamma(t)} - Y_{\gamma(0)}}{t} \tag{3.49}$$

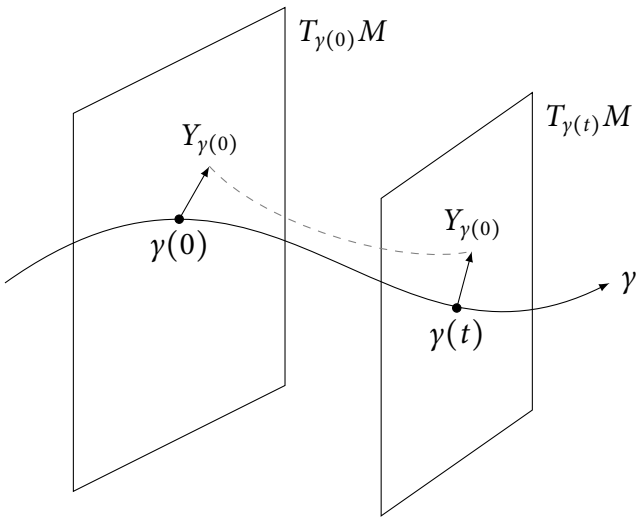


Figure 3.2: Curve γ runs in M. Given a vector field Y in M, the tangent vectors $Y_{\gamma(t)} \in T_{\gamma(t)}M$ and $Y_{\gamma(0)} \in T_{\gamma(0)}M$ are in different vector spaces, and one cannot be subtracted from the other unless we define identifications between the tangent fibers.

since the two tangent vectors $Y_{\gamma(t)} \in T_{\gamma(t)}M$ and $Y_{\gamma(0)} \in T_{\gamma(0)}M$ are in different fibers or vector spaces. See Figure 3.2 for an illustration. Therefore tangent vectors in different fibers should be identified to take derivatives of vector fields.

A classic way in information geometry to identify tangent vectors in different fibers is by α-connections. We review the approach of α-connection first. This is usually done by specifying Christoffel symbols by third-order derivatives of information divergence. See Amari and Nagaoka (2000, Chapter 3) and Zhang (2004). Here we illustrate a more procedural approach. Assume, temporarily, sufficient differentiability and boundedness conditions assumed in, e.g., Amari and Nagaoka (2000, Chapter 3). Let $\theta, \theta' \in M$, and let X, Y be vector fields in M. Then $Y_{\theta'}P_{\theta'}$ is mapped into $L^{1/\alpha}(P_\theta)$ by $T_{\alpha *}$ in (3.37),

$$T_{\alpha *}(Y_{\theta'}P_{\theta'}) = \left|\frac{dP_{\theta'}}{dP_\theta}\right|^\alpha \frac{dY_{\theta'}P_{\theta'}}{dP_{\theta'}}. \tag{3.50}$$

Since $L^{1/\alpha}(P_\theta)$ is a linear space, the differential action

$$X_\theta \cdot T_{\alpha *}(YP)\colon U \to L^{1/\alpha}(P_\theta) \tag{3.51}$$

of the vector field X is well defined. By the condition PAR4, there exists a unique tangent vector $W_\theta \in T_\theta M$ such that

$$T_{\alpha *}W_\theta = \left|\frac{dP_\theta}{dP_\theta}\right|^\alpha \frac{dW_\theta P_\theta}{dP_\theta} = \frac{dW_\theta P_\theta}{dP_\theta} \tag{3.52}$$

is equal to (3.51). We denote such vector $W_\theta = \nabla^{(\alpha)}_{X_\theta} Y$, and the operator $\nabla^{(\alpha)}$ is called the *α-connection*.

Remark 3.6. There may not exist a vector $W_\theta \in T_\theta M$ such that (3.52) is equal to the derivative (3.51), and one has to find the closest solution W_θ instead. One cannot simply use metric projections over Banach space here since linearity must be maintained for $\nabla^{(\alpha)}$ to be an affine connection while metric projections are *not* linear except in simple cases (Deutsch 1982). A solution is to pull back the derivative (3.51) by T_α^{-1} to a unique signed measure in $T_{P_\theta}\mathcal{M}$ where Fisher information inner product (3.40) is defined. With Fisher information (3.40), find the least square solution W_θ of

$$\langle W_\theta, Z_\theta \rangle = \langle X_\theta \cdot T_{\alpha *}(YP), Z_\theta \rangle \tag{3.53}$$

for every $Z_\theta \in T_\theta M$. If everything is locally bounded so that differential and integral may change their order,

$$X_\theta \langle Y, Z \rangle = \int X_\theta \big(T_{\alpha *}(YP) T_{(1-\alpha)*}(ZP) \big)\, dP = \langle \nabla^{(\alpha)}_{X_\theta} Y, Z \rangle + \langle Y, \nabla^{(1-\alpha)}_{X_\theta} Z \rangle. \tag{3.54}$$

Affine connections $\nabla^{(\alpha)}$ and $\nabla^{(1-\alpha)}$ are said to form a *dualistic structure* in M (Amari and Nagaoka 2000, Chapter 3). The claim also shows that 2^{-1}-connection $\nabla^{(1/2)}$ is the metric connection or Levi-Civita connection for the Fisher information.

We do not use α-connections directly in this chapter for some reasons. First, $\nabla^{(\alpha)}$ may not be well-defined since $T_{\alpha *}(YP)$ in (3.51) may not be differentiable. Many differentiability assumptions are hidden in the definition of $\nabla^{(\alpha)}$. Second, it is difficult to find a direct application of α-connection. A more useful concept is *α-transport*, the parallel transportation rule compatible with the α-connection.

The main idea of α-transport is to use the parallel transports in the vector space $L^{1/\alpha}(P_\alpha)$. If $Y_{\theta'} \in T_{\theta'} M$ then the transport of $Y_{\theta'}$ is a tangent vector W_θ such that

$$T_{\alpha *}(W_\theta P_\theta) = \frac{dW_\theta P_\theta}{dP_\theta} = \left| \frac{dP_{\theta'}}{dP_\theta} \right|^\alpha \frac{dY_{\theta'} P_{\theta'}}{dP_{\theta'}} = T_{\alpha *}(Y_{\theta'} P_{\theta'}) \tag{3.55}$$

in least square sense.

Let us discuss the role of α. When $\alpha = 1$,

$$T_{1*}(Y_{\theta'} P_{\theta'}) = \frac{dP_{\theta'}}{dP_\theta} \frac{dY_{\theta'} P_{\theta'}}{dP_{\theta'}} = \frac{dY_{\theta'} P_{\theta'}}{dP_\theta} \tag{3.56}$$

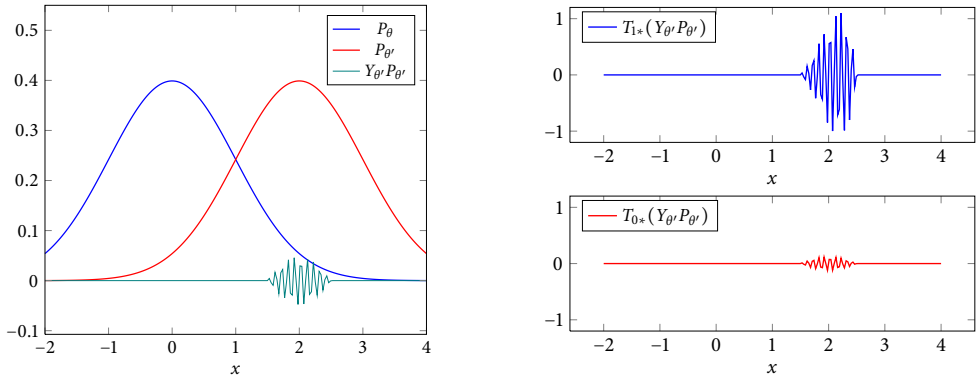


Figure 3.3: In this example, P_θ and $P_{\theta'}$ are Gaussian distributions in $\Omega = \mathbb{R}$ with means at $x = 0$ and $x = 2$. Suppose we have a measure derivative $Y_{\theta'}P_{\theta'}$ around $x = 2$. See the figure on the left. The derivative amplitude is relatively large in the view of P_θ while it is relatively small in the view of $P_{\theta'}$. On the right we compare α-parallel transports for $\alpha = 0, 1$.

by Proposition 3.5. Therefore the perturbation measure $Y_{\theta'}P_{\theta'}$ in $T_{\theta'}M$ is measured in the view of P_θ only, and the place $T_{\theta'}M$ where the perturbation takes place is ignored. On the other hand, when $\alpha = 0$,

$$T_{0*}(Y_{\theta'}P_{\theta'}) = \frac{dY_{\theta'}P_{\theta'}}{dP_{\theta'}} \tag{3.57}$$

in the support of P_θ. This time the perturbation is measured in the view of $P_{\theta'}$ only, and the transport ignores to where the perturbation is sent, P_θ. See Figure 3.3 for an example. When $0 < \alpha < 1$, the transport mixes the measurements. With abuse of notation,

$$T_{\alpha*}(Y_{\theta'}P_{\theta'}) = \frac{dY_{\theta'}P_{\theta'}}{(dP_\theta)^\alpha (dP_{\theta'})^{1-\alpha}}. \tag{3.58}$$

α-divergence, Bregman divergence, and normal charts

The α-divergence (3.7) and the α-variation (3.12) provide another motivation to localize the parameter space by $L^{1/\alpha}(P_\theta)$. Bregman divergence (Bregman 1967; Banerjee *et al.* 2005) is a general class of divergence. Let ψ be a convex function in convex open region U of some vector space. For $y, z \in U$ the Bregman divergence B_ψ is

$$B_\psi(z \parallel y) = \psi(y) - \psi(z) - D\psi(z) \cdot (y - z) \tag{3.59}$$

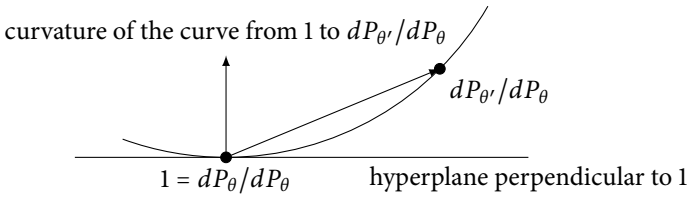


Figure 3.4: α-divergence measures the curvature of the curve from $1 = dP_\theta/dP_\theta$ to $dP_{\theta'}/dP_\theta$ in $L^{1/\alpha}(P_\theta)$.

where D is Gâteaux differential. If ψ is twice differentiable

$$B_\psi(z \parallel y) = \frac{1}{2} D^2\psi((1-r)y + rz) \cdot \big(y - z, y - z\big) \tag{3.60}$$

by the mean value theorem for some $0 \le r \le 1$. Since ψ is convex, the Hessian $D^2\psi$ is positive semidefinite everywhere and for y near z, B_ψ is approximately the Riemannian distance specified by $D^2\psi$, squared. Bregman divergence induces a (pseudo-)Riemannian structure.

It has been shown that α-divergence is a Bregman divergence for finite measurable space Ω when all probability measures are absolutely continuous to each other (Amari 2009). We show that α-variation $\mathcal{D}_\alpha(P_\theta \parallel P_{\theta'})$ is a Bregman divergence for fixed P_θ in general measurable space Ω and relate it to L^p spaces.

If $0 < \alpha < 1$, the norm of $L^{1/\alpha}(P)$ is Gâteaux differentiable except at 0. If $\psi(f) = \|f\|$ for $f \in L^{1/\alpha}(P)$, $f \neq 0$, then (for example, see Leonard and Sundaresan 1974, Theorem 3.1)

$$D\psi(f) \cdot u = \|f\|^{(\alpha-1)/\alpha} \int |f(x)|^{(1-2\alpha)/\alpha} f(x) u(x)\, dP_\theta(x) \tag{3.61}$$

for $u \in L^{1/\alpha}(P_\theta)$, and the integral is over the region where $f(x) \neq 0$.

Choose the convex function ψ in convex open region $U = L^{1/\alpha}(P_\theta) - \{0\}$. Map $\theta, \theta' \in M$ to $(dP_\theta/dP_\theta)^\alpha = 1$ and $(dP_{\theta'}/dP_\theta)^\alpha$, respectively, then

$$B_\psi(P_\theta \parallel P_{\theta'}) = \left(\int \frac{dP_{\theta'}}{dP_\theta}\, dP_\theta\right)^\alpha - \left(\int \frac{dP_\theta}{dP_\theta}\, dP_\theta\right)^\alpha - \int \left(\left(\frac{dP_{\theta'}}{dP_\theta}\right)^\alpha - \left(\frac{dP_\theta}{dP_\theta}\right)^\alpha\right) dP_\theta \tag{3.62}$$

$$= \left(\int \frac{dP_{\theta'}}{dP_\theta}\, dP_\theta\right)^\alpha - \int \left(\frac{dP_{\theta'}}{dP_\theta}\right)^\alpha dP_\theta \tag{3.63}$$

$$= \mathcal{D}_\alpha(P_\theta \parallel P_{\theta'}) \tag{3.64}$$

from (3.12). Therefore $\mathcal{D}_\alpha(P_\theta \parallel P_{\theta'})$ is a Bregman divergence for fixed P_θ. KL variation (3.11) may be understood as the norm curvature limit of $\{L^{1/\alpha}(P_\theta)\}$ as $\alpha \to 1$. See Figure 3.4.

The observation above implies that the Radon-Nikodym derivative plays the role of normal charts in Riemannian geometry. Recall that if M is a Riemannian manifold and $\theta \in M$, then there exists a local coordinate system $(\theta_1, \ldots, \theta_d)$ in a neighborhood U of θ such that for all $\theta' \in U$,

$$\operatorname{dist}(\theta, \theta')^2 = \sum_{i=1}^{d} |\theta_i - \theta_i'|^2 \tag{3.65}$$

where dist denotes the geodesic distance. Such coordinate system is called the normal coordinates. In general, the explicit expression for normal coordinates is intractable. In the case of information divergences, the right side of (3.65) is replaced with B_ψ. Then Radon-Nikodym derivative map $P_{\theta'} \mapsto dP_{\theta'}/dP_\theta$ induces a normal coordinate system in $L^1(P_\theta)$,

$$\mathcal{D}_\alpha(P_\theta \parallel P_{\theta'}) = B_\psi(P_\theta \parallel P_{\theta'}). \tag{3.66}$$

Therefore we obtain a method—Radon-Nikodym derivative—to compute the normal coordinate system. The trade-off is that norm expression on the right side of (3.65) is replaced by more complicated B_ψ.

3.4 Parameter space interpolation

Let P_0, Q_0 be two probability measures. We consider the problem how to construct a "minimal" curve that connects P_0 and Q_0 along probability measures. We have Fisher-Riemann metric and α-parallel transports. Therefore the minimal curves should be auto-parallel curves with respect to the α-transport rules, i.e., they should be re-parameterized to geodesics in $L^{1/\alpha}(P)$.

Let $\gamma\colon [0,1] \to L^{1/\alpha}(P_0)$ be a curve such that $\gamma(0) = (dP_0/dP_0)^\alpha$ and $\gamma(1) = (dQ_0/dP_0)^\alpha$. Let $\mathcal{M}_{P_0} \subset \mathcal{M}$ denote the linear subspace of $\mathcal{M}$ consists of absolutely continuous signed measures with respect to P_0. Let $\tilde{\gamma}\colon [0,1] \to \mathcal{M}$ is a lift of γ such that the sequence

$$[0,1] \xrightarrow{\tilde{\gamma}} \mathcal{M} \longrightarrow \mathcal{M}_P \longrightarrow L^1(P_0) \longrightarrow L^{1/\alpha}(P_0) \tag{3.67}$$

is equal to γ. $\tilde{\gamma}$ is an interpolation between P_0 and Q_0.

Suppose that γ is a straight line. In general, the lift $\tilde{\gamma}$ is not uniquely determined. Since

spheres in $L^{1/\alpha}(P)$ are strictly convex, $\|\gamma(s)\| < 1$ for $0 < s < 1$ and there are infinitely many probability measures that differ in singular parts to P_0 and map into $\gamma(s)$ by (3.67). If $Q_0 = Q_\parallel + Q_\perp$ where $Q_\parallel$ and $Q_\perp$ denote the absolutely continuous and singular components of Q_0 respectively, and $Q_\perp \neq 0$, then the singular part $Q_\perp$ may remove the ambiguity,

$$\tilde{\gamma}(s) = Q_s + rQ_\perp \tag{3.68}$$

where Q_s is the absolutely continuous measure determined by $\gamma(s)$ and $r > 0$ ensures the condition $\|\tilde{\gamma}(s)\| = 1$. This is a reasonable interpolation since $(dQ_0/dP_0)^\alpha$ is inside the unit sphere of $L^{1/\alpha}(P_0)$ and a continuous curve γ must enter inside the sphere anyway.

On the other hand, suppose that $Q_0 = Q_\parallel$ and there is no singular component. In this case, a straight line γ cannot determine the interpolating probability measures at all. Therefore if $Q_0 = Q_\parallel$ then we consider a geodesic curve γ in the unit sphere of $L^{1/\alpha}(P_0)$. From (3.61), this condition may be stated as

$$\int \gamma(s)^{(1-\alpha)/\alpha}\gamma'(s)\, dP_0 = 0, \tag{3.69}$$

and we may always construct the interpolation γ such that $\gamma(s)$ is nonnegative function for all $s \in [0,1]$.

From now on, we restrict our attention to the case $\alpha = 2^{-1}$. There are several reasons to do this. Recall that the interpolation is in $\mathcal{M}$ by $\tilde{\gamma}$. The purpose of γ is only to compute parallel transportations. Since $\mathcal{M}$ is equipped Fisher information metric, for each curve $\tilde{\gamma}$, parallel vector fields along $\tilde{\gamma}$ is determined. These parallel vector fields are transformed to vector fields along γ by α-representation. The question is if the transformed vector fields are also parallel along γ. From (3.54) the answer is negative unless $\alpha = 2^{-1}$.

Another reasons comes from information divergence. As mentioned before, KL divergences D_{KL} does not accurately measure the difference when two probability measures are mutually singular. Therefore one needs to consider Fisher-Riemann distance by

$$\inf \sum_{i=0}^{n-1} \sqrt{2D_{KL}(P_i \parallel P_{i+1})} \tag{3.70}$$

over all finite sequences with $P_n = Q_0$. This Fisher-Riemann metric is isometric to $L^{1/\alpha}(P_0)$ only when $\alpha = 2^{-1}$. See Carter *et al.* (2009a) and Carter, Raich, Finn, and Hero (2011) for an application in dimensionality reduction.

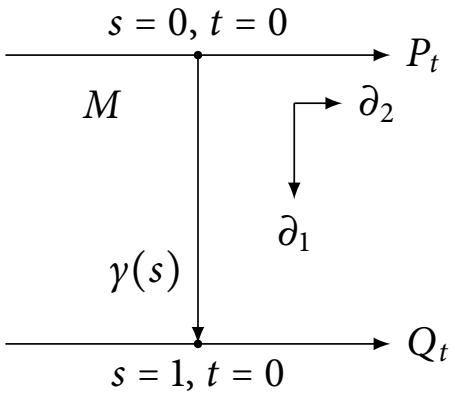


Figure 3.5: An example of $M = [0,1] \times I$. Top horizontal line at $s = 0$ is equal to P_t, and bottom horizontal line at $s = 1$ is equal to Q_t. Each vertical line maps to a geodesic curve in the unit sphere of $L^2(P_t)$.

3.4.1 Interpolation parameter space

Let us go back to one-dimensional parameterizations. Let $I \subset \mathbb{R}$ be an open interval which contains 0, and I will be the domain of one-dimensional parameterization. Suppose that we are given two parameterizations $P_t, Q_t \in \mathcal{M}$, $t \in I$. As planned, P_t and Q_t are connected through geodesic curves, and create a new two-dimensional parameterization $S\colon M = [0,1] \times I \to \mathcal{M}$ such that $S(0,t) = P_t$, $S(1,t) = Q_t$, and the curve $s \mapsto S(s,t)$ is geodesic for all $t \in I$. See Figure 3.5. We assume Q_t is absolutely continuous with respect to P_t for each t, hence dQ_t/dP_t integrates to one. We look for the area of the immersed surface M, and

$$\text{Area} = \int_I \int_0^1 \sqrt{\det(F_{ij})}\, ds\, dt, \tag{3.71}$$

where (F_{ij}) is the Fisher-Riemann metric. In the next subsection, we fix t and focus on the inner integral, the growth rate of area.

3.4.2 Transversal vector field

Let ∂_1 and ∂_2 be vector fields of $M = [0,1] \times I$:

$$\partial_1 = \frac{\partial}{\partial s} \quad \text{and} \quad \partial_2 = \frac{\partial}{\partial t}. \tag{3.72}$$

Let us fix $t = 0$ and let $P = P_0$. Let γ be the vertical curve into $L^2(P_0)$,

$$\gamma\colon M \supset [0,1] \times \{0\} \longrightarrow \mathcal{M} \longrightarrow L^2(P_0). \tag{3.73}$$

We restrict ∂_1 and ∂_2 on γ, and $\partial_1(s), \partial_2(s)$ are vector fields along the curve $\gamma(s)$.

Let us introduce notations for Fisher information along γ:

$$F_{11}(s) = \langle \partial_1(s), \partial_1(s) \rangle, \tag{3.74}$$

$$F_{12}(s) = \langle \partial_1(s), \partial_2(s) \rangle, \tag{3.75}$$

$$F_{22}(s) = \langle \partial_2(s), \partial_2(s) \rangle. \tag{3.76}$$

Since γ is a geodesic curve, $F_{11}(s)$ is constant over s. We simply write it as F_{11}.

Note that $\partial_1(s) = \gamma'(s)$, and $\partial_1(s)$ is completely determined by the end points $\gamma(0)$ and $\gamma(1)$. For the other vector field ∂_2, note that the two-dimensional parameterization ζ is a variation of a geodesic curve, γ, through geodesics. Such variation is well studied in Riemannian geometry and it is solution to the Jacobi equation (Lang 1999, Chapter IX)

$$\nabla_{\partial_1} \nabla_{\partial_1} \partial_2 = R(\partial_1, \partial_2)\partial_1, \tag{3.77}$$

where R is the Riemannian curvature. Since the unit sphere in $L^2(P_0)$ has constant sectional curvature 1, the Jacobi equation translates to

$$\nabla_{\partial_1} \nabla_{\partial_1} \partial_2 = \langle \partial_1, \partial_2 \rangle \partial_1 - \langle \partial_1, \partial_1 \rangle \partial_2. \tag{3.78}$$

Jacobi equation is a linear second-order differential equation with initial values: $\partial_2(0)$ and $\nabla_{\partial_1}\partial_2(0)$. It may be verified that the solution to the equation takes the form

$$\partial_2(s) = (c_1 + c_2 s)\partial_1(s) + \frac{X(s)}{\omega}\cos(\omega s) + \frac{Y(s)}{\omega}\sin(\omega s), \tag{3.79}$$

where c_1, c_2 are constants, $\omega = \sqrt{F_{11}}$, and X, Y are parallel vector fields along γ that are perpendicular to ∂_1. The term $c_1\partial_1(s) + \omega^{-1}X(s)\cos(\omega s)$ is determined by the initial condition $\partial_2(0)$, and the term $c_2 s\partial_1(s) + \omega^{-1}Y(s)\sin(\omega s)$ is determined by the initial condition $\nabla_{\partial_1}\partial_2(0)$. Since X and Y are perpendicular to ∂_1,

$$\frac{X(s)}{\omega}\cos(\omega s) + \frac{Y(s)}{\omega}\sin(\omega s) = \partial_2(s) - \frac{\langle \partial_1(s), \partial_2(s) \rangle}{\langle \partial_1(s), \partial_1(s) \rangle}\partial_1 \tag{3.80}$$

$$= \partial_2(s) - \frac{F_{12}(s)}{F_{11}}\partial_1(s), \tag{3.81}$$

and hence

$$\left\| X(s)\cos(\omega s) + Y(s)\sin(\omega s) \right\|^2 = F_{11}F_{22}(s) - F_{12}(s)^2. \tag{3.82}$$

Therefore the inner integral in (3.71) becomes

$$J = \int_0^1 \|X(s)\cos(\omega s) + Y(s)\sin(\omega s)\|\, ds. \tag{3.83}$$

and let J denote the inner integral, or the growth rate of area. Again, $\omega = \sqrt{F_{11}}$.

3.4.3 Area growth rate calculation from Fisher information

The surface M and the parameterization S are produced from given one-dimensional parameterizations P_t and Q_t. Therefore it is a reasonable assumption that we have some knowledge of P_t and Q_t. In this subsection, we seek to find expressions for the integrand (3.82) in (3.83) in terms of Fisher information values evaluated on two parameterizations: $F_{ij}(0)$ and $F_{ij}(1)$ for $i, j = 1, 2$.

Let us expand the above expression,

$$\begin{aligned} &\|X(s)\cos(\omega s) + Y(s)\sin(\omega s)\|^2 \\ &\qquad = \langle X, X\rangle \cos(\omega s)^2 + \langle X, Y\rangle \sin(2\omega s) + \langle Y, Y\rangle \sin(\omega s)^2. \end{aligned} \tag{3.84}$$

Therefore the key part is to calculate the inner products $\langle X, X\rangle$, $\langle X, Y\rangle$, and $\langle Y, Y\rangle$.

From (3.82), when $s = 0$ and $s = 1$, we have

$$\langle X, X\rangle = F_{11}F_{22}(0) - F_{12}(0)^2, \tag{3.85}$$

and

$$\langle X, X\rangle(\cos\omega)^2 + \langle Y, Y\rangle(\sin\omega)^2 + \langle X, Y\rangle \sin(2\omega) = F_{11}F_{22}(1) - F_{12}(1)^2. \tag{3.86}$$

The equations above are linear in terms of the inner product values, and we have two linear equations for three unknowns. Therefore we need one more linear equation to solve the problem.

To derive the third equation, we calculate the parallel transports of ∂_2 along the geodesic curves in the unit sphere. The following proposition can be verified with direct computation.

Proposition 3.7. *Let A be a point on the unit sphere of $L^2(P_0)$, and γ be a geodesic curve in the sphere passing through A: $\gamma(0) = A$. If X is a parallel vector field along γ and $X(s)$ is perpendicular to $\gamma'(s)$, then $X(s)$ is equal to the parallel transport of $X(0)$ along the straight line between $\gamma(0)$ and $\gamma(s)$ in $L^2(P_0)$. That is, $L^2(P_0)$ expression of an orthogonal parallel vector field $X(s)$ is constant for all s.*

From the proposition above, we compare the values of ∂_2 in (3.79) at $s = 0, 1$ in $L^2(P_0)$,

$$\langle \partial_2(0), \partial_2(1) \rangle \tag{3.87}$$

$$= \left\langle c_1 \partial_1(0) + \frac{X(0)}{\omega}, (c_1 + c_2)\partial_1(1) + X(1)\frac{\cos\omega}{\omega} + Y(1)\frac{\sin\omega}{\omega} \right\rangle \tag{3.88}$$

$$= c_1(c_1 + c_2)\langle \partial_1(0), \partial_1(1) \rangle + \frac{\langle X, X \rangle}{F_{11}} \cos\omega + \frac{\langle X, Y \rangle}{F_{11}} \sin\omega. \tag{3.89}$$

By taking inner products of (3.79) with ∂_1,

$$c_1 = \frac{F_{12}(0)}{F_{11}} \quad \text{and} \quad c_1 + c_2 = \frac{F_{12}(1)}{F_{11}}. \tag{3.90}$$

Therefore the equation above becomes

$$F_{11}\langle \partial_2(0), \partial_2(1) \rangle - \cos\omega F_{12}(0) F_{12}(1) = \langle X, X \rangle \cos\omega + \langle X, Y \rangle \sin\omega. \tag{3.91}$$

Combine (3.85), (3.86), and (3.91) to have a linear system of equations

$$\begin{pmatrix} 1 & 0 & 0 \\ \cos\omega & \sin\omega & 0 \\ (\cos\omega)^2 & \sin(2\omega) & (\sin\omega)^2 \end{pmatrix} \begin{pmatrix} \langle X, X \rangle \\ \langle X, Y \rangle \\ \langle Y, Y \rangle \end{pmatrix} = \begin{pmatrix} F_{11}F_{22}(0) - F_{12}(0)^2 \\ F_{11}\langle \partial_2(0), \partial_2(1) \rangle - \cos\omega F_{12}(0)F_{12}(1) \\ F_{11}F_{22}(1) - F_{12}(1)^2 \end{pmatrix}. \tag{3.92}$$

Invert the matrix on the left-hand side, and the solution $(\langle X, X \rangle, \langle X, Y \rangle, \langle Y, Y \rangle)$ is

$$\begin{pmatrix} 1 & 0 & 0 \\ -\cot\omega & \csc\omega & 0 \\ (\cot\omega)^2 & -2\csc\omega\cot\omega & (\csc\omega)^2 \end{pmatrix} \begin{pmatrix} F_{11}F_{22}(0) - F_{12}(0)^2 \\ F_{11}\langle \partial_2(0), \partial_2(1) \rangle - \cos\omega F_{12}(0)F_{12}(1) \\ F_{11}F_{22}(1) - F_{12}(1)^2 \end{pmatrix}. \tag{3.93}$$

3.4.4 Fisher information calculation

Calculation for sampled densities

Fisher information definition involves differentiations. Since parameterization values are at prescribed only at $s = 0, 1$, the interpolation curve γ should be computed and numerical differentiation should be done in general. However, numerical differentiation may be avoided with sphere geometry.

Since γ is a geodesic curve in the unit sphere of $L^2(P_0)$,

$$F_{11} = F_{11}(0) = \left(\arccos \int \sqrt{\frac{dQ_0}{dP_0}}\, dP_0\right)^2. \tag{3.94}$$

This formula works for F_{12} and F_{22} as well. In practice, no continuum of P_t and Q_t are known, but rather the parameterizations are specified only at some discrete values of t. That is we know P_t and Q_t for some sequence $\cdots < T_{-1} < T_0 = 0 < T_1 < \cdots$. Therefore it works the same as before except that Q_0 is replaced with P_T. Then

$$F_{22}(0) = \frac{1}{T_1^2}\left(\arccos \int \sqrt{\frac{dP_{T_1}}{dP_0}}\, dP_0\right)^2 \tag{3.95}$$

and

$$F_{12}(0) = \langle \partial_1(0), \partial_2(0)\rangle = \frac{\arccos(A)\arccos(B)}{T_1\sqrt{(1-A^2)(1-B^2)}}(C - AB) \tag{3.96}$$

where

$$A = \int \sqrt{\frac{dP_{T_1}}{dP_0}}\, dP_0, \quad B = \int \sqrt{\frac{dQ_0}{dP_0}}\, dP_0, \quad C = \int \sqrt{\frac{dP_{T_1}}{dP_0}\frac{dQ_0}{dP_0}}\, dP_0, \tag{3.97}$$

3.5 Approximations

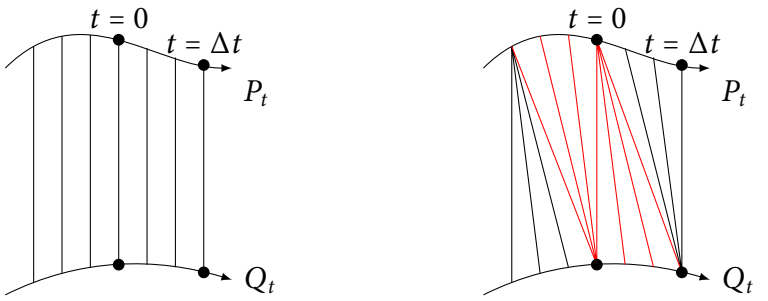
3.5.1 Triangulation approximation

The exact area growth J in (3.83) requires numerical integration. However, there are cases where the solution becomes simpler. Let us consider the case where one of the one-dimensional parameterizations is constant. Say $P_t = P_0 = P$ for all valid t. In other words, the image of M is no longer rectangular-like but it now has sector-like shape. In this case, $\partial_2(0) = 0$ and $X(s) = 0$ for all $s \in [0,1]$. Therefore $\langle X, X\rangle = 0$ and $\langle X, Y\rangle = 0$. Then by (3.82)

$$F_{11}F_{22}(1) - F_{12}(1)^2 = \langle Y, Y\rangle(\sin\omega)^2 \tag{3.98}$$

and

$$\|X(s)\cos(\omega s) + Y(s)\sin(\omega s)\|^2 = \langle Y, Y\rangle \sin(\omega s)^2 \tag{3.99}$$



(a) Visualization of M. (b) Visualization of N.

Figure 3.6: Vertical lines in (a) represent the geodesics used in M. In triangulation approximation, vertical lines are replaced by diagonal lines. See red lines in (b). The red lines on the right side of $t = 0$ represents J_Q where P_t is fixed at P_0 and Q_t proceeds. The red lines on the left side of $t = 0$ represents J_P where Q_t is fixed at Q_0 and P_t proceeds.

for all $s \in [0,1]$. Note that the sine value here is always nonnegative. Therefore J in (3.83) becomes

$$\|Y\| \int_0^1 \sin(\omega s)\, ds = \frac{\|Y\|}{\omega}(1-\cos\omega) = \sqrt{\frac{1-\cos\omega}{1+\cos\omega}}\sqrt{F_{22}(1) - \frac{F_{12}(1)^2}{F_{11}}}, \tag{3.100}$$

or

$$\tan\left(\frac{\omega}{2}\right)\sqrt{F_{22}(1) - \frac{F_{12}(1)^2}{F_{11}}}. \tag{3.101}$$

Previously the area growth J is calculated while P_t and Q_t evolve simultaneously. Now let us alter the problem slightly and break the surface increment into two pieces. Q_t is incremented while P_t is fixed at P_0 from $t = 0$ till $t = \Delta t > 0$. Then the roles are switched and P_t is incremented from P_0 to $P_{\Delta t}$ while Q_t is fixed at $Q_{\Delta t}$. Let J_P and J_Q denote the increments while Q_t and P_t are fixed, respectively. Define $J_T = J_P + J_Q$,

$$J_T = J_P + J_Q = \tan\left(\frac{\omega}{2}\right)\left(\sqrt{F_{22}(0) - \frac{F_{12}(0)^2}{F_{11}}} + \sqrt{F_{22}(1) - \frac{F_{12}(1)^2}{F_{11}}}\right). \tag{3.102}$$

J_T is a heuristic approximate solution to J, and we will refer to it as *triangular approximation*.

To break a problem with multiple dependent factors into mulptiple stages and to solve in one factor at a time with the others fixed is a common approach. For example, Gauss-Seidel method separates a matrix into super-diagonal component and sub-diagonal component, and then updates the solution with each factor fixed at a time.

Let us visualize and see what J_T represents. Let $\Delta > 0$, and consider the surface spanned by P_t and Q_t. Let M be the immersed surface discussed so far, therefore the vertical line segments map into geodesic curves in $L^2(P)$. Let N be another immersed surface such that

(1) the immersion images of curves $\{r \mapsto (r, (rs+n)\Delta) : 0 \le r \le 1, 0 \le s \le 1, n \in \mathbb{Z}\}$ are geodesic curves,

(2) the immersion images of curves $\{r \mapsto (r, (r(1-s)+s+n)\Delta) : 0 \le r \le 1, 0 \le s \le 1, n \in \mathbb{Z}\}$ are geodesic curves.

See Figure 3.6 for an illustration of N. Therefore N is another surface construction slightly different from M. While N has computational advantage, its drawback is that it depends on increment unit Δ.

3.5.2 Surface energy

Another approximate solution or another concept of dissimilarity between the curves is the energy of the surface spanned by the parameterizations. That is, rather than to integrate the square root of the Fisher information determinant, integrate the determinant first and take the square root after. Define surface energy by

$$\text{Energy} = \int_I E_t \, dt \tag{3.103}$$

where

$$E_t = \int_0^1 (F_{11}(s,t)F_{22}(s,t) - F_{12}(s,t)^2) \, ds. \tag{3.104}$$

From (3.82),

$$E_0 = \int_0^1 \|X(s)\cos(\omega s) + Y(s)\sin(\omega s)\|^2 \, ds \tag{3.105}$$

$$= \|X\|^2 \int_0^1 \cos(\omega s)^2 \, ds + \|Y\|^2 \sin(\omega s)^2 \, ds + \langle X, Y\rangle \int_0^1 \sin(2\omega s) \, ds \tag{3.106}$$

$$= \Big(\frac{1}{2} + \frac{\sin(2\omega)}{4\omega}\Big)\|X\|^2 + \Big(\frac{1}{2} - \frac{\sin(2\omega)}{4\omega}\Big)\|Y\|^2 + \frac{(\sin\omega)^2}{\omega}\langle X, Y\rangle \tag{3.107}$$

and the inner product values may be obtained from (3.92). Compared to J, the computation of E_0 avoids the numerical integral part.

The natural concern will be the relation between J and E. By the Jensen's inequality,

$$\sqrt{E_0} \ge \int_0^1 \sqrt{F_{11}(s,0)F_{22}(s,0) - F_{12}(s,0)^2} \, ds = J, \tag{3.108}$$

and the equality holds if and only if $F_{11}(s,0)F_{22}(s,0) - F_{12}(s,0)^2$ is constant over $s \in [0,1]$ since square root is strictly concave. Unlike curves, a non-flat surface cannot have a local isometry to any flat surface. Since the curvature is intrinsic and does not depend on the choice of the charts, the integrand in general cannot be manipulated into constant. Therefore E_0 and J will have certain difference depending on the intrinsic geometry imposed by the parameterization. We will see the empirical difference in Section 3.7.1.

3.6 Comparison to other curve distances

The surface area, denoted as d_A, as a curve dissimilarity measure is compared with traditional curve distances. We consider Hausdorff distance d_H and Chamfer distance d_C. The curves are functions of some fixed interval I into some fixed metrizable vector space with metric d_X. Given two curves f and g, define

$$d_H(f \to g) = \sup_{t\in I} \inf_{x\in I} d_X(f(t), g(x)) \tag{3.109}$$

$$d_H(f,g) = \max\{d_H(f \to g), d_H(g \to f)\} \tag{3.110}$$

and

$$d_C(f \to g) = \int_I \inf_{s\in I} d_X(f(t), g(s))\, dt \tag{3.111}$$

$$d_C(f,g) = \frac{1}{2}\big(d_C(f \to g) + d_C(g \to f)\big). \tag{3.112}$$

Clearly the surface area d_A is a geometric generalization of Chamfer distance d_C, and we may expect some similarities between them.

Perhaps the most important difference between the surface area and traditional curve distances like Hausdorff or Chamfer distance is that both traditional curve distances require global knowledge of the curves for distance computation whereas surface area requires only local knowledge and is additive. To be more specific, Hausdorff and Chamfer distances require the computation of closest point. For each $f(t)$, the infimum over the whole curve domain

$$\inf_{x\in I} d_X(f(t), g(x)) \tag{3.113}$$

is required. Therefore the distance cannot be computed until the whole curve is retrieved. On the other hand, surface area computation may be carried out locally. That is, if $I_1, \ldots, I_n$

is a partition of the curve domain I,

$$d_A(f,g) = \sum_{k=1}^{n} d_A\big(f|I_k, g|I_k\big), \tag{3.114}$$

and each summand computation is independent from the others.

This property is largely beneficial when we have long sequences. For example, suppose that we have video sequences to compare. As the video sequences become longer, the global knowledge requirement becomes more demanding. As mentioned above, we need to find out the closest point for each frame. For the closest point computation in (3.113), we need to search over the whole video sequences and the time complexity grows with the video length. And such large search set must be contained in the workspace hence the space complexity grows as well. In the case of surface area, however, there is no need to sweep over the whole set, and the time complexity of the area growth J is fixed with respect to the video length. Therefore the time complexity of surface area grows linearly with respect to the video length while it grows quadratically for Hausdorff and Chamfer distances.

This advantage comes with a cost when compared to Hausdorff distance. Note that Hausdorff distance is not restricted to curves but is a general metric between any kinds of subsets. In particular, it does not depend on the parameterizations of the curves. Surface area measure is also invariant under re-parameterizations of the curves. It does, however, depend on the synchronizations of the curves. To illustrate, let $\alpha\colon I \to I$ be a re-parameterization, and compare f and $g \circ \alpha$ instead. Then the surface spanned by the new curves is different from the old one since the vertical geodesic curves are connecting $f(t)$ and $g(\alpha(t))$ instead of $f(t)$ and $g(t)$, and the surface area must have changed. However, this cost is inevitable for local computation. Suppose that one would like to modify Hausdorff or Chamfer distance so that its computation becomes local. A simple option would be to partition the curve domain I as (3.114) and to compute the distances within each segment. The choice of partition correspond to synchronization.

3.6.1 Sensitivity analysis

Consider the Gâteaux derivative

$$\frac{d}{dr} d_A(f + rh, g)\bigg|_{r=0} \tag{3.115}$$

and similar derivatives for d_H and d_C. Here h is another curve.

Let us fall back to a simple case where the curves map into $\mathbb{R}^2$. Let $g(t)$ is a straight line segment. Without loss of generality, we may assume that $g(t) = (t, 0)$. Assume that $f(t)$ is

the graph of a function $F: I \to \mathbb{R}$, i.e., $f(t) = (t, F(t))$. Then

$$d_A(f, g) = \int_I |F(t)|\, dt \tag{3.116}$$

and the sensitivity s_A is

$$s_A = \frac{d}{dr} d_A(f + rh, g) = \int_I \operatorname{sign}(F(t)) H(t)\, dt \tag{3.117}$$

where H is defined from h in the same way F was from f.

For Hausdorff distance, note that for fixed $t \in I$,

$$\inf_{x \in I} \|f(t) - g(x)\| = \inf_{x \in I} \|(t, F(t)) - (x, 0)\| = \|(t, F(t)) - (t, 0)\| = \|f(t) - g(t)\|, \tag{3.118}$$

from the convexity of the axis $\{(x, 0)\}$. From the obvious inequality $\inf_{x \in I} \|f(x) - g(t)\| \leq \|f(t) - g(t)\|$,

$$d_H(f, g) = \sup_{t \in I} |F(t)|. \tag{3.119}$$

Restrict our attention to compact intervals and let $T \in I$ be the place where the supremum occurs. Then the sensitivity s_H is

$$s_H = \frac{d}{dr} d_H(f + rh, g)\bigg|_{r=0} = \operatorname{sign}(F(T)) H(T). \tag{3.120}$$

When compared with s_H, s_A takes average of the function $\operatorname{sign}(F(t))H(t)$ over t while s_H evaluates the same function at fixed point T. Therefore the sensitivity of Hausdorff distance is determined by the perturbation at some determined place. Note that $f(T)$ is where f is farthest from g. Suppose that the curves are noisy so that there are some outliers. Then it is likely that the point $f(T)$ is an outlier and the sensitivity s_H becomes dependent on the perturbation at the outlier point. In this kind of scenario, both d_H and s_H are controlled by an exceptional point, and are less robust.

For Chamfer distance, the analysis is more complicated. From Hausdorff distance calculation, we know $d_C(f \to g)$ is equal to $\int_I |F(t)|\, dt$. The coincidence of $d_C(f \to g)$ and d_A here is due to the parameterization assumed for f. Chamfer distance is variant under re-parameterizations, and $d_C(f \to g)$ may be expressed as a Stieltjes integral $\int_I |F(t)|\, d\beta(t)$ where β is a re-parameterization function. We ignore re-parameterizations for simplicity.

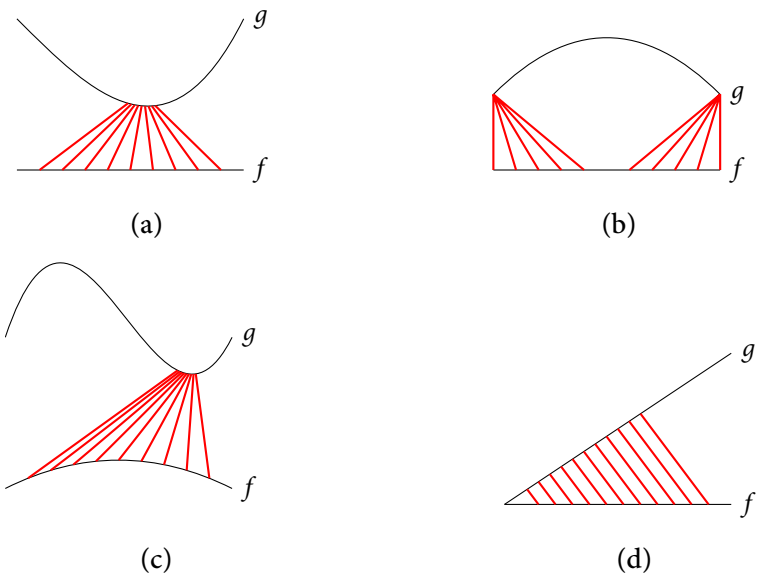


Figure 3.7: Illustrations of $\inf_{t\in I}\|f(x) - g(t)\|$. Red lines denote the nearest neighbor pairs for $f(x)$. Except for case Figure 3.7d, $d_C(g \to f)$ depends on some small portions of f.

Back to the sensitivity analysis, the sensitivity s_C of Chamfer distance is

$$s_C = \frac{1}{2}\int_I \operatorname{sign}(F(t))H(t)\,dt + \frac{d}{2dr}d_C(g \to f + rh)\Big|_{r=0}. \tag{3.121}$$

The second term on the right side often depends only on small parts of f. See Figure 3.7.

Lastly, we note the difference that the sensitivity of Hausdorff distance is determined by the points that are far away from the other curve while the sensitivity of Chamfer distance is determined by the points that are close to the other curve.

3.7 Simulations

3.7.1 Jeffreys' prior

For general parameter space M, its volume may be defined through Fisher information. This volume is called (unnormalized) Jeffreys' prior in Bayesian statistics. Assume that this volume is finite. Then after normalization, the volume form transforms into a probability measure form, and this is defined purely by the parameterization. One of the difficulties in Bayesian approach is to determine the prior. When no prior knowledge exists, Jeffreys' prior serves as a candidate.

The main purpose of this experiment is to examine the effect of approximation schemes from Section 3.5, surface energy and triangulation. To establish ground truth, we use a simple model. Let

$$p_t(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-t)^2}{2}\right) \tag{3.122}$$

and

$$q_t(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-t)^2}{2\sigma^2}\right), \tag{3.123}$$

where $x, t \in \mathbb{R}$, so p_t, q_t are parameterizations of Gaussian distributions in $\mathbb{R}$ running through mean shifts. p_t has fixed variance 1, and q_t has variance $\sigma^2 \neq 1$.

To calculate the theoretical expectation, note that a translation on real line is a measure-preserving isomorphism. Therefore $F_{11}(s,t)$ is constant over t as well as s. If m denotes Lebesgue measure,

$$\cos\omega = \int_{\mathbb{R}} \sqrt{p_t q_t}\, dm = \sqrt{\frac{2\sigma}{1+\sigma^2}}. \tag{3.124}$$

For F_{22},

$$F_{22}(0) = \frac{1}{4} \quad \text{and} \quad F_{22}(1) = \frac{1}{4\sigma^2}. \tag{3.125}$$

For F_{12},

$$F_{12}(0) = \frac{1}{\operatorname{sinc}\omega} \int_{\mathbb{R}} \sqrt{q_t} \frac{\partial}{\partial t} \sqrt{p_t}\, dm = 0, \tag{3.126}$$

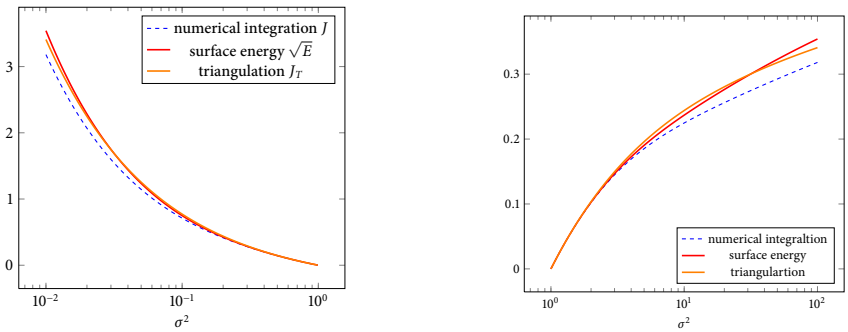
$$F_{12}(1) = \frac{1}{\operatorname{sinc}\omega} \int_{\mathbb{R}} \sqrt{p(t)} \frac{\partial}{\partial t} \sqrt{q_t}\, dm = 0. \tag{3.127}$$

For F_{22x},

$$F_{22x} = \int_{\mathbb{R}} \frac{\partial\sqrt{p_t}}{\partial t} \frac{\partial\sqrt{q_t}}{\partial t}\, dm = \sqrt{\frac{\sigma}{2(1+\sigma^2)^3}} \tag{3.128}$$

Then we have a system of linear equations from (3.92)

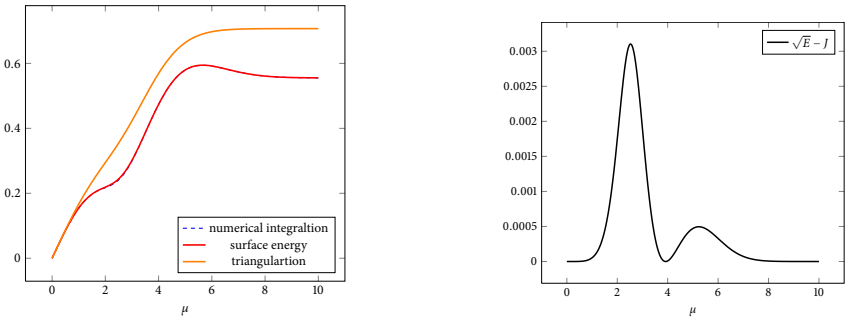
$$\begin{pmatrix} 1 & 0 & 0 \\ \cos\omega & \sin\omega & 0 \\ (\cos\omega)^2 & \sin(2\omega) & (\sin\omega)^2 \end{pmatrix} \begin{pmatrix} \langle X, X\rangle \\ \langle X, Y\rangle \\ \langle Y, Y\rangle \end{pmatrix} = F_{11} \begin{pmatrix} 4^{-1} \\ \sqrt{\sigma(2(1+\sigma^2))^{-3}} \\ (4\sigma^2)^{-1} \end{pmatrix}. \tag{3.129}$$



(a) When $0 < \sigma^2 < 1$.

(b) When $\sigma^2 > 1$.

Figure 3.8: Plot of numerical integral J (- - -), triangulation approximation J_T (——), and square root surface energy $\sqrt{E}$ (——) against σ^2.



(a) When $\mu > 0$.

(b) $\sqrt{E} - J$.

Figure 3.9: Plot of numerical integral J (- - -), triangulation approximation J_T (——), and square root surface energy $\sqrt{E}$ (——) against μ. In (a), J is difficult to notice since $\sqrt{E}$ overlaps. The difference of J and $\sqrt{E}$ is shown in (b).

From the linear equation, we have inner product values. We compute the area increment

$$\int_0^1 \| X\cos(\omega s) + Y\sin(\omega s)\| \, ds \tag{3.130}$$

through numerical integration, and this quantity must be close to the true value. Then we compare the numerical integral with approximation values: triangular approximation and square root of surface energy:

$$\sqrt{\int_0^1 \| X\cos(\omega s) + Y\sin(\omega s)\|^2 \, ds}. \tag{3.131}$$

Figure 3.8 shows plots the comparisons with varying σ^2 in q_t.

On the other hand, we can change the parameterization, so that curve evolves in the

direction of variance σ^2 rather than mean, so the new parameterization becomes

$$u_t(x) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{x^2}{2t}\right) \quad \text{and} \quad v_t(x) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(x-\mu)^2}{2t}\right), \tag{3.132}$$

where $\mu \neq 0$. Figure 3.9 compares the quantities in interest.

3.7.2 Action recognition

Although this chapter is motivated by the applications of time-parameterized probability measures, the surface area computation methods proposed in this chapter apply to broader class of time-parameterized objects. Note that the only mathematical property used in Section 3.4.3 is the constant sectional curvature property of spheres in finite or infinite-dimensional Hilbert spaces. Therefore the area computation approach in this chapter is valid for any curves in unit spheres of Hilbert spaces, whether they are L^2 spaces or not. Examples of curves not in an L^2 space include quantum state trajectories in quantum state space. See Chapter 4 for more discussion.

In this subsection we illustrate the surface are approach for action recognition. We will compare surface area, Hausdorff distance, and Chamfer distance in terms of classification error performance.

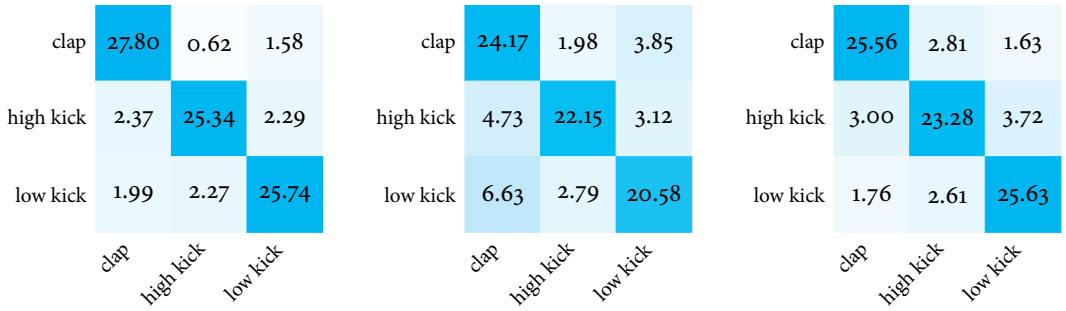
Surface area is computed by triangulation approximation J_T instead of numerical integration as it has much faster run time. From now on, by surface area we mean triangulation approximated area. As defined, surface area needs to be slightly modified for application to streaming data like video since the surface area monotonically increases as more frames are added to the action sequences. For this reason, the surface area is normalized by the product of two curve lengths L, where

$$L = \int_I \sqrt{F_{22}}(t)\, dt. \tag{3.133}$$

We apply the normalized surface area to the problem of classification of PGM Kinnect action data.[3] The data contains three actions: *clap, high kick, low kick*. There are 30 sequences for each action type, thus 90 sequences in total. Action sequences have at minimum 11 frames and at maximum 24 frames. Each frame has xy-coordinates and orientations of ten human body parts: *torso, head, left arm, left forearm, right arm, right forearm, left thigh, left leg, right thigh, right leg*.

We extracted relative orientations of body parts from each frame. For example, if the coordinates of the endpoints of the left forearm are $x, y \in \mathbb{R}^2$, then the unit orientation vector

[3] https://www.coursera.org/course/pgm



(a) Surface area. Classification error .1184

(b) Hausdorff. Classification error .2629

(c) Chamfer. Classification error .1699

Figure 3.10: Confusion matrix of PGM data. 5-fold cross-validation. Average over 100 runs.

$(x - y)/\|x - y\|$ is one feature. Therefore each feature represents the relative orientation of a body part, and such orientation features are collected for all body parts listed above. Since there are ten body parts in the dataset, the features are in $(\mathbb{S}^1)^{10}$, ten-fold Cartesian product of circles. After normalization in $(\mathbb{S}^1)^{10} \subset (\mathbb{R}^2)^{10} \cong \mathbb{R}^{20}$, each action sequence is now a discretely sampled curve in $\mathbb{S}^{19}$. Therefore the action sequences are in a sphere, and we are able to apply the surface area computation method from Section 3.4.3.

A Nearest-neighbor classifier was employed on the relative orientation data. Figure 3.10 has confusion matrices created by 5-fold cross-validation. The cross-validation error rates are 11.84%, 26.29%, and 16.99% for surface area, Hausdorff, and Chamfer distance, respectively. Therefore, for this experiment the surface area significantly improves upon the Hausdorff and Chamfer distances in terms of classification performance.

We repeated the experiment on the MSR-Action3D data described in Li, Zhang, and Liu (2010) and Wang, Liu, Wu, and Yuan (2012). The data contains twenty human actions: *high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up and throw*. These actions were collected by depth cameras. Each frame in MSR data has 20 joints of human body in $\mathbb{R}^3$. Figure 3.11 has sample frames of the data set.

As in PGM data, we use body part orientations as features. Since there are 20 joints in $\mathbb{R}^3$, each frame is represented by $\mathbb{S}^{56}$. Figure 3.12 shows confusion matrices created by 5-fold cross-validation. The median cross-validation error rates across the 20 classes are 8.50%, 10.82%, and 4.96% for surface area, Hausdorff, and Chamfer distance, respectively.

While it outperforms Hausdorff in this MSR experiment, the surface area does not outperform the Chamfer distance in terms of average classification performance. Figure 3.12a

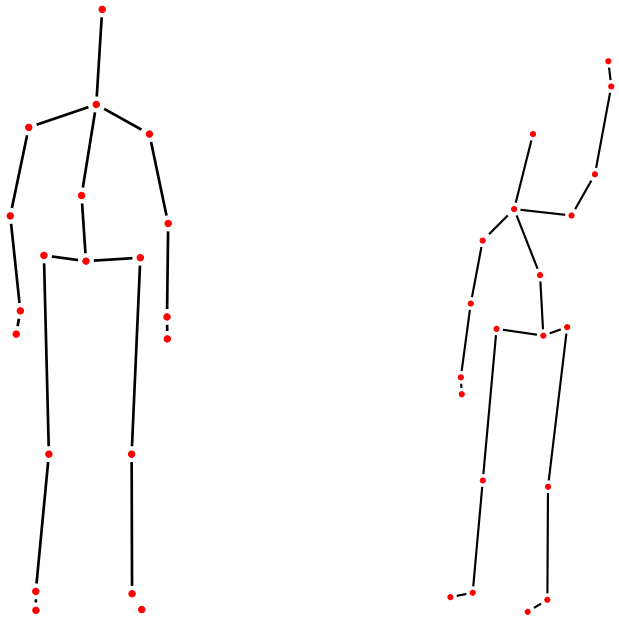
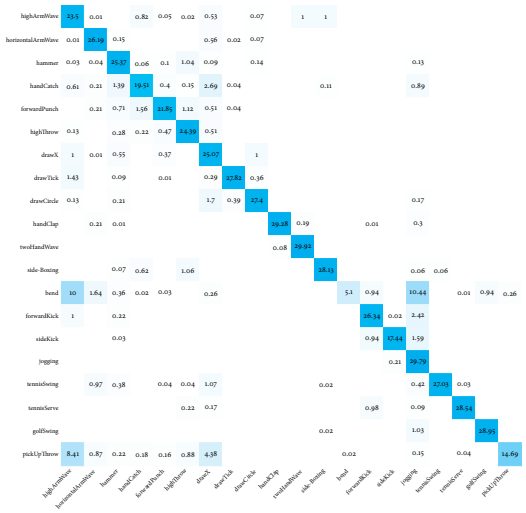


Figure 3.11: Sample skeleton images from MSR-Action3D data.

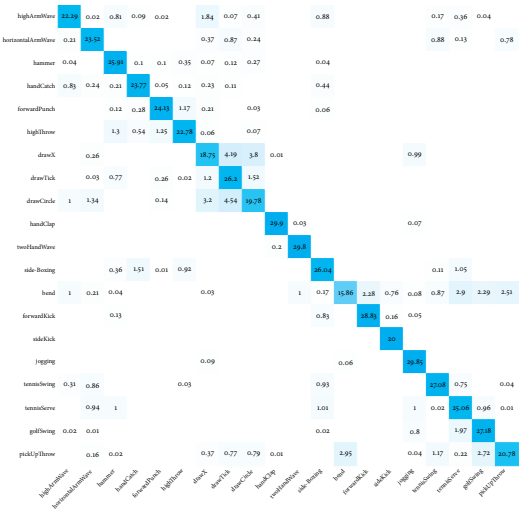
shows that the most significant factor is *bend* action class. Compared to Hausdorff and Chamfer distances, surface area misclassifies *bend* action sequences to *jogging* actions.

It turns out that the MSR-Action3D data has defects in several action classes including *bend* and *pickup and throw*. The depth camera and the human skeleton tracking algorithm fail when the camera cannot see the whole subject body. For instance, when a human subject bends in the scene, the head, the torso, and many other body parts overlap in the depth camera view. Therefore the camera cannot track the body parts effectively, and many frames of the corresponding actions have degenerate data.

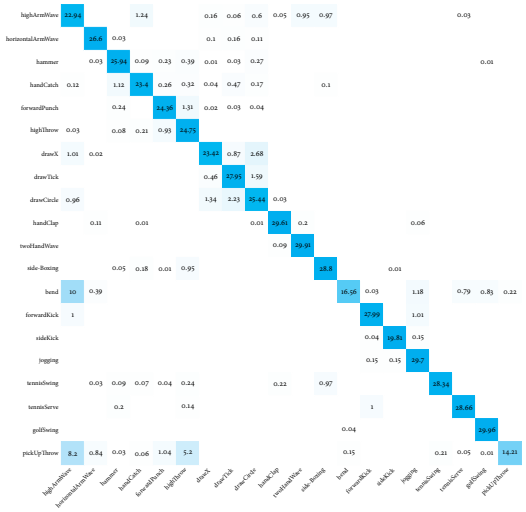
If *bend* and *pickup and throw* classes are discarded from the data, the 5-fold cross-validation error rates have the medians 6.93%, 9.28%, and 4.71% for surface area, Hausdorff, and Chamfer distances. Further investigation is needed to better understand these differences in classification performance.



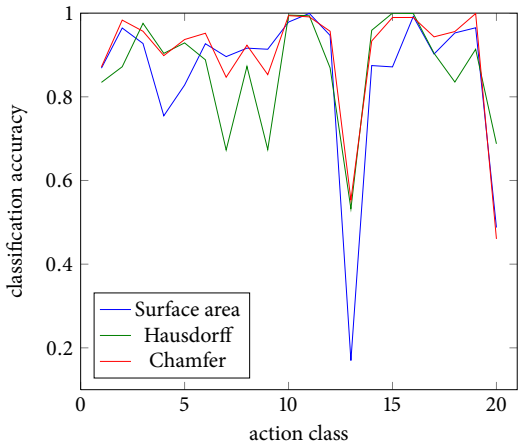
(a) Surface area. Error: 8.50%



(b) Hausdorff. Error: 10.82%



(c) Chamfer. Error: 4.96%



(d) Cross-validation accuracy per class

Figure 3.12: Confusion matrix of MSR-Action3D data. 5-fold cross-validation. Average over 100 runs.

Chapter 4
Quantum Cluster Analysis

4.1 Introduction

In Chapter 2 and Chapter 3, we proposed novel dissimilarity measures that enjoyed good properties for random samples. Dissimilarity measures are frequently used in spectral clustering, in which one performs dimensionality reduction by eigen-decomposition on the Gram matrix of a transformed version of these dissimilarities. For example, Gaussian kernel and Markov diffusion transition probabilities are common in manifold learning and spectral clustering (Belkin and Niyogi 2003; Coifman and Lafon 2006; Ng *et al.* 2002). In this chapter, we develop a new framework for dimensionality reduction and clustering based on a quantum mechanical interpretation of the latent variables, called states, that identify cluster membership. Similar to other "soft clustering" methods, e.g., fuzzy k-means (Duda, Hart, and Stork 2000, Chapter 10.4.4) and probabilistic mixtures of experts (Tipping and Bishop 1999), the quantum mechanical framework results in replacing a binary valued membership function with a probability valued membership function. However, in the quantum mechanical framework it is a cluster equivalence class indicator function that is relaxed to a probability.

The quantum mechanical framework for cluster analysis based on quantum states.[1] In quantum mechanics, dynamics of systems are described by the quantum states associated to a system, and each state is a unit vector (or a unit operator) in some Hilbert space H. The main idea behind quantum cluster analysis is to establish a mapping from the sample space to the state space H, and to study the interactions between the subsystems by the interactions between its state vectors.

[1] This should not be confused with Horn and Gottlieb (2002) which is based on the Schrödinger's equation.

4.2 Mixture models and k-means

Mixture model is a probabilistic approach of cluster analysis. In (finite) mixture models, sample points follow a probability density function of the form

$$\sum_{k=1}^{m} a_k f_k \tag{4.1}$$

where $\sum_k a_k = 1$, $a_k \geq 0$, and f_k is a probability density function in $\mathbb{R}^d$ for $k = 1, 2, \ldots, m$. The density functions f_k are called the mixture density functions, and they are often chosen to be Gaussian density functions. Each sample point X_i, drawn from (4.1), has a *latent variable* Y_i, which takes value k with probability a_k, $k = 1, 2, \ldots, m$. The latent variable Y_i controls the distribution of X_i such that the conditional density function of X_i conditioned by $Y_i = k$ is f_k.

The outcome of the latent variable Y_i determines what mixture density function f_k the sample point X_i belongs to. Therefore in mixture model, the cluster analysis attempts to estimate the latent variable Y_i given the point X_i. A typical approach is maximum-likelihood estimation. If $\hat{Y}_i$ is the estimation of Y_i, then

$$\hat{Y}_i = \arg\max_k f_k(X_i). \tag{4.2}$$

This choice of cluster assignment is similar to the k-means algorithm, which is probably the most prevalent geometric approach of cluster analysis. In its greedy implementation, the k-means algorithm assumes that there are finite m clusters with centroids at $\mu_k \in \mathbb{R}^d$, $k = 1, 2, \ldots, m$ (Lloyd 1982). Then each sample point X_i belongs to the cluster with the nearest centroid. If the mixture model had Gaussian mixture functions f_k with means at μ_k and the identity covariance I_d, then the maximum-likelihood estimation $\hat{Y}_i$ in (4.2) is equivalent to the k-means,

$$\arg\max_k f_k(X_i) = \arg\min_k |X_i - \mu_k|. \tag{4.3}$$

The mixture model and k-means have interpretations in terms of geometric and statistical models. The mixture models provide a statistical interpretation but do not actively consider the underlying geometric structure of the data space. On the other hand, the k-means and its extensions such as fuzzy k-means do not assume a probabilistic model, and their designs are motivated by geometric interpretations.

In the next subsection, we show how quantum mechanical model can provide a frame-

work that can account for geometry and statistical models simultaneously. In soft-clustering, fuzzy clustering, and probabilistic latent variable mixture models, the indicator function of the event "X_i belongs to cluster Y_j" is replaced with a continuous valued membership function in the interval $[0,1]$. In contrast, in the proposed quantum clustering approach a different indicator function is relaxed to $[0,1]$: the indicator of the equivalence relation "X_i is in the same cluster as X_j". In Section 4.3, we show a connection between the proposed quantum mechanical clustering method and mixture models using the spectral theory of Hilbert-Schmidt operators.

4.2.1 Quantum mechanical generalization

Suppose that we have n sample points $X_1, X_2, \ldots, X_n$. Like latent variables in the mixture model, define cluster labels $Y_1, Y_2, \ldots, Y_n$ associated to the sample points.

We design the algorithm so that the geometry learning process and the data partition process are separated. In this chapter, we focus on the latter. Therefore the sample points $X_1, \ldots, X_n$ are not used directly. In fact, we do not even specify in what space these sample points are. Instead it is assumed that some black-box geometry learning process provides a dissimilarity matrix $D_{ij}, 1 \le i < j \le n$.

The dissimilarity D may be any real function. However, for the cluster analysis to make sense, we require D to have the following properties.

1. D is topologically compatible. For example, $D(x,x) \le D(x,y)$ for all x, y.
2. D is symmetric, i.e., $D(x,y) = D(y,x)$.

The state optimization problem is motivated by the k-means algorithm (MacQueen 1967; Hastie, Tibshirani, and Friedman 2009, Chapter 14). The k-means algorithm minimizes the within-point scatter defined as

$$W(Y) = \frac{1}{2} \sum_k D_{ij} 1_{\{Y_i = Y_j\}} \tag{4.4}$$

where the index k runs through all values of Y_i, $i = 1, 2, \ldots, n$. Our approach is to express (4.4) with pairwise probability p_{ij},

$$p_{ij} = \Pr\{Y_i = Y_j \mid X_i = x_i, X_j = x_j\}. \tag{4.5}$$

The cluster analysis no longer provides deterministic output of the cluster prediction Y_i. Rather the cluster predictions are randomized and cluster analysis provides the probability

p_{ij} instead. If the predictions are random, then the expected within-point scatter (4.4) is

$$EW(Y) = \frac{1}{2}\sum_{i\neq j} D_{ij} E 1_{\{Y_i = Y_j\}} = \frac{1}{2}\sum_{i\neq j} D_{ij} p_{ij}. \tag{4.6}$$

The mean scatter (4.6) may be interpreted as half the sum of the mean cluster dispersions $\sum_j D_{ij} p_{ij}$ for all $1 \leq i \leq n$.

We call the minimization of (4.6) *quantum state optimization*. While the k-means—the base method—is purely geometric and make hard cluster memberships of the data points, the quantum state optimization is a soft cluster analysis, and the cluster memberships are expressed in probability. Some examples of previous soft cluster analysis methods are the fuzzy k-means, the expectation-maximization (EM), or latent Dirichlet allocation (Blei, Ng, and Jordan 2003). The key difference of the quantum state optimization from the examples is the bivariate membership, i.e., the previous examples estimate the distribution of cluster labels Y_i, $i = 1, 2, \ldots, n$, whereas the quantum state optimization answers the probability that $Y_i = Y_j$, for all $i \neq j$.

Note that the minimization of (4.6) is currently not well defined yet since a trivial solution $p_{ij} = 0$ for all $i \neq j$ would attain a minimum. The next section provides a statistical cluster analysis framework for the quantum state optimization, and an optimization constraint for the minimization of (4.6).

4.3 Quantum cluster analysis

We introduce the proposed quantum cluster analysis framework, then provide arguments on how to apply the framework to cluster analysis.

4.3.1 Quantum mechanical background

This subsection is a brief introduction to the mathematical background and basic postulates in quantum mechanics (Shankar 1994, Chapter 4; Sakurai 1993, Chapter 1). Bear in mind that the purpose of this short introduction is to borrow some concepts and mathematical utilities from quantum mechanics, not to discuss and develop theories of physics in depth. For that reason, the introduction below is modified and tuned for the purpose of the thesis, and one may observe some differences from the formal definitions in quantum mechanics. For example, the Hilbert space H is usually chosen to be a complex linear space in quantum physics but in this chapter, H is always chosen to be a real vector space. Lastly, the *bra-ket* notation is very common in quantum mechanics literature. That notation is not used in this

thesis.

A *(pure) quantum state* Ψ of a system is a one-dimensional subspace in some Hilbert space H. Therefore the set of pures states is the projection space of H,

$$\{\mathsf{span}(x) : x \in H\}. \tag{4.7}$$

It is convenient to represent a state Ψ by a normalized vector in the subspace. We will use ψ to denote a unit vector in the subspace Ψ. A (normalized) quantum state is also called a wave function, a probability amplitude, or an ensemble.

A measurement or an observable Ω in quantum mechanics is a self-adjoint operator of H. Ω is self-adjoint, hence it admits a spectral decomposition[2]

$$\Omega = \sum_j \lambda_j \omega_j \omega_j^* \tag{4.8}$$

where $\lambda_j \in \mathbb{R}$ is an eigenvalue, $\omega_j \in H$ is a unit eigenvector corresponding to λ_j, and $\omega_j^* \in H^*$ is the linear functional

$$\omega_j^*(x) = \langle \omega_j, x \rangle \tag{4.9}$$

for all $x \in H$.

Suppose that a system is in quantum state ψ. When an observable Ω is measured, then the possible outcomes of the measurement are λ_j's,

$$\Pr\{\text{observation is } \lambda_j\} = |\langle \omega_j, \psi \rangle|^2. \tag{4.10}$$

Note that it is possible $\lambda_j = \lambda_k$ for $j \neq k$. When the observation outcome is λ_j, the state of the system transits from ψ to ω_j. We say the state ψ collapsed into ω_j. By the Parseval's theorem, the probability is invariant over different choices of the orthonormal bases for the eigenspaces.

4.3.2 Quantum states for cluster analysis

Assume n sample points $X_1, \ldots, X_n$ in some measurable space. Each point X_i is associated a normalized state variable ψ_i in some Hilbert space H. In the language of physics, the system consists of n subsystems and each subsystem X_i has state ψ_i, $i = 1, 2, \ldots, n$. We estimate the *cluster label* Y_i for X_i. If $Y_i = Y_j$ for some indicies i, j, then X_i and X_j are *observed* to be in

[2] An operator may have continuous spectra in general H. We assume the spectra are always discrete.

the same cluster. In quantum mechanics the notion of observation is somewhat different from the classical statistical notion.

Define a *partition operator* as an observable, i.e., a self-adjoint operator Ω of H. If $\Omega = \sum_j \lambda_j \omega_j \omega_j^*$, then

$$Y_i = \lambda_j \tag{4.11}$$

with probability $|\langle \omega_j, \psi_i \rangle|^2$. We define the cluster labels to be the outcome eigenvalues, and *clusters* to be the corresponding eigenspaces of the partition. Note that the cluster label values have no specific meanings and their sole purpose is identification of cluster membership of each sample point. When comparing clusters from different partitions, clusters are represented by the eigenspaces rather than the eigenvalues.

The strength of this framework is that a partition operator Ω may be any operator we wish to be.

Example 4.1. Suppose that there are two input variables X_1 and X_2, and associate the states $\psi_1 = (1, 0)$ and $\psi_2 = (\sqrt{1/3}, \sqrt{2/3}) \in \mathbb{R}^2 = H$, respectively. Choose a partition operator $\Omega = \lambda_1 \omega_1 \omega_1^* + \lambda_2 \omega_2 \omega_2^*$,

$$\lambda_1 = 1, \quad \omega_1 = (1, 0) \in \mathbb{R}^2, \tag{4.12}$$

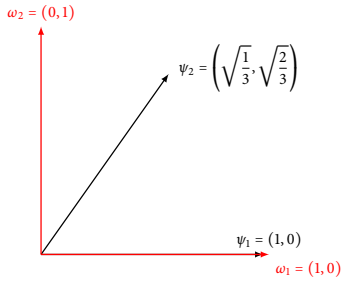
$$\lambda_2 = 2, \quad \omega_2 = (0, 1) \in \mathbb{R}^2. \tag{4.13}$$

Then X_1 has probability one to collapse into cluster $\lambda_1 = 1$ since $|\langle \omega_1, \psi_1 \rangle|^2 = 1$, and has probability zero to collapse into cluster $\lambda_2 = 2$ as $|\langle \omega_2, \psi_1 \rangle|^2 = 0$. For X_2, it has probabilities $1/3$ and $2/3$ to collapse into cluster 1 and 2, respectively. See Figure 4.1 for a visualization.

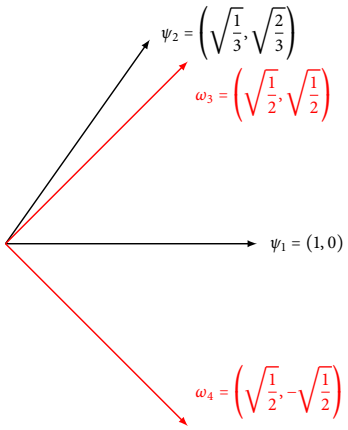
Now pick another partition operator $\Omega' = \lambda_3 \omega_3 \omega_3^* + \lambda_4 \omega_4 \omega_4^*$ and choose

$$\lambda_3 = 3, \quad \omega_3 = \left(\sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}} \right) \in \mathbb{R}^2, \tag{4.14}$$

$$\lambda_4 = 4, \quad \omega_4 = \left(\sqrt{\frac{1}{2}}, -\sqrt{\frac{1}{2}} \right) \in \mathbb{R}^2. \tag{4.15}$$



(a) Partition operator with ω_1, ω_2.



(b) Partition operator with ω_3, ω_4.

Figure 4.1: Quantum states are ψ_1 and ψ_2. Any partition operator may applied to the states.

Then

$$\Pr\{Y_1 = k\} = \begin{cases} \frac{1}{2}, & k = 3, \\ \frac{1}{2}, & k = 4, \end{cases} \tag{4.16}$$

$$\Pr\{Y_2 = k\} = \begin{cases} \frac{1}{2} + \frac{\sqrt{2}}{3}, & k = 3, \\ \frac{1}{2} - \frac{\sqrt{2}}{3}, & k = 4. \end{cases} \tag{4.17}$$

4.3.3 Mixture models in quantum cluster analysis

In this subsection, we compare quantum cluster analysis with the finite mixture model described in Section 4.2. In mixture models, each sample point X_i has a latent variable Y_i, and the probability distribution of the latent variable is determined by the maximum-likelihood estimation (4.2).

Quantum cluster analysis provides a dual to the mixture models. Let H be the L^2 space in $\mathbb{R}^d$. Associate a normalized quantum state or wave function $\psi \in H$ to each sample point, which is like a particle in quantum physics. Then the squared magnitude $|\psi|^2$ is a probability density function governing where the particle may be observed. Of course a sensible wave function should have its mean at the sample point it is associated to. We design the partition operator based on position operators. Let $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$ be mean vectors of the mixtures. Choose a small neighborhood U of $0 \in \mathbb{R}^d$ such that the collection $\{\mu_j + U\}_{j=1}^k$ are disjoint.

Then define the partition operator Ω,

$$\Omega\psi(x) = \sum_{j=1}^{k} 1_{\{x\in\mu_j+U\}} \int_{\mu_j+U} \psi(u)\, du \tag{4.18}$$

where $1_{\{x\in\mu_j+U\}}$ is one if $x \in \mu_j + U$ and is zero otherwise. Since

$$\left|\int_{\mu_j+U} \psi(u)\, du\right|^2 \le \int_{\mu_j+U} |\psi(u)|^2\, du \le \int_{\mathbb{R}^d} |\psi(u)|^2\, du, \tag{4.19}$$

Ω is bounded, hence is a continuous operator. The eigenfunctions of Ω are $1_{\{x\in\mu_j+U\}}$, $j = 1, \ldots, k$. Then the probability that a sample point X_i collapses into the cluster μ_j is proportional to

$$\frac{1}{|U|}\left|\int_{\mu_j+U} \psi_i(u)\, du\right|^2 \approx |\psi_i(\mu_j)|^2 |U| \tag{4.20}$$

where $|U|$ denotes the volume of U. The approximation holds when U is small enough. In that case, (4.20) matches the likelihood function used by the maximum-likelihood estimation (4.2) in the mixture models.[3]

4.4 State space dimensionality

Quantum states model the dependency between sample points. Let H be a Hilbert state space, and let $\psi_i \in H$ be the normalized quantum state for the sample point X_i, $i = 1, 2, \ldots, n$ where $\langle \psi_i, \psi_j \rangle = 1$. If $\psi_1, \psi_2, \ldots, \psi_n$ are orthonormal then there exists a non-trivial partition operator

$$\Omega = \sum_{i=1}^{n} i\psi_i\psi_i^* \tag{4.21}$$

which assigns every X_i to a singleton cluster with probability one. On the other hand, if $\psi_1 = \psi_2 = \cdots = \psi_n$ are identical then for any partition operator, the cluster observation process is independent and identically distributed, i.e., $Y_1, Y_2, \ldots, Y_n$ are i.i.d.

[3]One may have noticed that (4.20) does not sum to one over $j = 1, 2, \ldots, k$. This is because there is the cluster which corresponds to the zero eigenvalue, i.e., the null space of Ω. There may be multiple ways to handle or interpret the null space cluster. One way is to consider the null space cluster as the $(k+1)$-th cluster. Another way to handle it is to restrict the wave functions ψ_i to the domain $\bigcup_j (\mu_j + U)$. Other possible approach is the update the partition operator and add another cluster with a mean μ_{k+1}, as in the Dirichlet process (Rasmussen 2000; Blei *et al.* 2003).

Apply the partition operator

$$\Psi_i = \psi_i \psi_i^*, \tag{4.22}$$

then the i-th class label $Y_i = 1$ with probability one. In addition, if the state of X_i collapses into some cluster η with probability one, then ψ_i must be in the subspace of η. In that sense, we say the one-dimensional subspace spanned by ψ_i is the *smallest* cluster that contains X_i almost surely. Define p_{ij} as the probability that X_j collapses into the smallest cluster prototypical to X_i, i.e.,

$$p_{ij} = |\langle \psi_i, \psi_j \rangle|^2. \tag{4.23}$$

Note the symmetry $p_{ij} = p_{ji}$.

The minimization may be rewritten as

$$\arg\min_{\{\psi_i\}} \frac{1}{2} \sum_{i \neq j} D_{ij} p_{ij} = \frac{1}{2} \sum_{i \neq j} D_{ij} |\langle \psi_i, \psi_j \rangle|^2 \tag{4.24}$$

for unit vectors $\psi_1, \ldots, \psi_n \in H$.

The dimensionality of finite-dimensional Hilbert space $H = \mathbb{R}^m$ plays the role of the optimization constraint. If $m \geq n$, then the states $\psi_1, \ldots, \psi_n$ may be chosen to be an orthonormal sequence. In that case, $p_{ij} = 0$ for all $i \neq j$,

$$\frac{1}{2} \sum_{i \neq j} D_{ij} p_{ij} = 0 \tag{4.25}$$

regardless of the given dissimilarity matrix D_{ij}. On the other hand, if $m = 1$, then $p_{ij} = 1$ for all i, j and the optimization problem is trivial since there is no degree of freedom for the quantum states.

Therefore it is important to choose an intermediate value for the dimensionality m of H between 1 and n. If the dimensionality m is strictly less than n, then it is impossible to reach the trivial solution $p_{ij} = 0$ for all $i \neq j$. The trivial solution is achieved when $\psi_1, \psi_2, \ldots, \psi_n$ are orthonormal vectors, and it is impossible to have such configuration when $m < n$. A basic rule is that m should indicate the maximal number of clusters allowed for the point configuration. In diffusion terms, the dimensionality determines the approximate number of neighborhoods in the space.

4.5 Quantum state optimization: implementation

In this section, we propose a general strategy to minimize (4.24). As discussed above, the dimensionality of the state space H must be less than the number of sample points n to avoid trivial solutions. Since finite-dimensional Hilbert spaces are isomorphic to Euclidean spaces, without loss of generality, we can assume that the normalized state variables in (4.24) are on the unit sphere $\mathbb{S}^m$ in $\mathbb{R}^{m+1}$. Standard optimization techniques operate over Euclidean domains. Therefore to apply the already developed optimization methods, we use chart maps of the smooth manifold $\mathbb{S}^m$. This section describes the details and the related computations.

Let J denote the objective function in (4.24), where the normalized state variables are denoted by $\psi_1, \psi_2, \ldots, \psi_n \in \mathbb{S}^m$. We use stereographic projections to settle the variable points in Euclidean domains. We review stereographic projections in the next subsections.

4.5.1 Stereographic projection

Previous dimensionality reduction methods such as spherical multidimensional scaling (Cox and Cox 2001) or spherical Laplacian information maps (Carter, Raich, and Hero 2009b) use spherical coordinates to represent vectors on the unit spheres for its nonlinear optimization. Spherical coordinates, however, are not local homeomorphisms at the poles, and they often lead to optimization instabilities.

A more systematic approach is to use chart maps. In quantum state optimization, stereographic projection chart maps are used. Stereographic projections have some advantages for gradient descent methods. For instance, single stereographic chart covers all but one point on the sphere, and two stereographic charts are sufficient to cover the entire $\mathbb{S}^m$. Compare it with gnomonic projections where $2(m+1)$ charts are required to cover the entire $\mathbb{S}^m$.

Let $U = \mathbb{S}^m - (1, 0, 0, \ldots, 0)$, and $\varphi^{-1}\colon U \to \mathbb{R}^m$ be the stereographic projection with its projection focal point at $(1, 0, \ldots, 0)$,

$$\varphi^{-1} : \psi \in U \longmapsto \xi = \frac{\pi_2 \psi}{1 - \pi_1 \psi}, \tag{4.26}$$

where $\pi_1\colon \mathbb{R}^{m+1} \to \mathbb{R}$ is the projection onto the first coordinate, and $\pi_2\colon \mathbb{R}^{m+1} \to \mathbb{R}^m$ is the projection onto the last m coordinates. And its projection inverse $\varphi\colon \mathbb{R}^m \to U$,

$$\varphi\colon \xi \longmapsto \psi = \left(1 - \frac{2}{1 + |\xi|^2}, \frac{2\xi}{1 + |\xi|^2}\right). \tag{4.27}$$

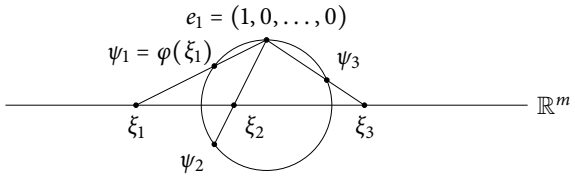


Figure 4.2: Examples of a stereographic projection. The normalized state variables ψ_1, ψ_2, ψ_3 on $\mathbb{S}^m$ are mapped onto $\xi_1, \xi_2, \xi_3 \in \mathbb{R}^m$, respectively. The focal point is $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^{m+1}$. When ψ_i approaches and is close to the focal point e_1 for some i, the state variable ψ_i is replaced by $-\psi_i$.

See Figure 4.2 for a visualization.

The stereographic projection map φ^{-1} cannot map its focal point $(1, 0, \ldots, 0) \in \mathbb{S}^m$ into $\mathbb{R}^m$. Note that two normalized vectors ψ and $-\psi$ represent the same quantum state. Indeed, the objective function (4.24) is invariant when ψ_i is replaced with $-\psi_i$, for $i = 1, 2, \ldots, n$. Therefore when a normalized state variable ψ is near the focal point, we replace it with $-\psi$ to avoid the singularity of the stereographic projection.

Let $\xi_+ = \varphi^{-1}(\psi)$ and $\xi_- = \varphi^{-1}(-\psi)$. One may check that

$$\xi_+ = -\frac{\xi_-}{|\xi_-|^2} \quad \text{and} \quad \xi_- = -\frac{\xi_+}{|\xi_+|^2}. \tag{4.28}$$

We describe an optimization strategy for quantum state optimization. Choose a threshold $R > 1$.

Step 1 Initialize $\psi_1, \psi_2, \ldots, \psi_n \in \mathbb{S}^m$, and compute $\xi_1, \ldots, \xi_n \in \mathbb{R}^m$ using (4.26).

Step 2 Update $\xi_1, \xi_2, \ldots, \xi_n$ using a preferred iterative descent method on $J(\varphi(\xi_1), \ldots, \varphi(\xi_n))$. A broad class of iterative methods may be adopted here. For example, Newton's method, gradient descent, interior point method, etc.

Step 3 After an update, check if any variable needs to be flipped to avoid the singularity of the stereographic projection. For $i = 1, 2, \ldots, n$, if $|\xi_i| > R$, then replace ξ_i by $-\xi_i/|\xi_i|^2$, as in (4.28).

Step 4 Repeat Step 2 and Step 3 until $\xi_1, \ldots, \xi_n$ converge.

4.5.2 Derivatives under stereographic projections

Many iterative optimization methods that may be adopted in Step 2 require the knowledge of the optimization objective gradient, and sometime the Hessian of the objective function

J. This subsection provides computation results for the iterative updates.

For the first-order partial derivatives of J,

$$\frac{\partial J}{\partial \xi_i} = \frac{1}{2}\sum_{j\neq i} D_{ij}\frac{\partial}{\partial \xi_i}|\langle \psi_j, \psi_i\rangle|^2 = \sum_{j\neq i} D_{ij}\langle \psi_j, \psi_i\rangle\left\langle \psi_j, \frac{d\psi_i}{d\xi_i}\right\rangle, \tag{4.29}$$

and from (4.27),

$$\frac{d\psi_i}{d\xi_i} = \left(\frac{2(2\xi_i)^*}{(1+|\xi_i|^2)^2}, \frac{2\cdot I_m}{1+|\xi_i|^2} - \frac{(2\xi_i)(2\xi_i)^*}{(1+|\xi_i|^2)^2}\right) \tag{4.30}$$

where $I_m\colon \mathbb{R}^m \to \mathbb{R}^m$ is the identity operator, and ξ^* denotes the dual functional $\xi^*(\cdot) = \langle \xi, \cdot\rangle$. One can check that its representation in terms of ψ_i is

$$\frac{d\psi_i}{d\xi_i} = \Big((1-\pi_1\psi_i)\pi_2\psi_i^*, (1-\pi_1\psi_i)I_m - \pi_2\psi_i\pi_2\psi_i^*\Big). \tag{4.31}$$

One can check that if we map $u \in \mathbb{R}^m$ into $(0,u) \in \mathbb{R}\times\mathbb{R}^m \cong \mathbb{R}^{m+1}$, then at each ξ_i, the derivative (4.31) is the Householder transformation incorporating a scaling by $1-\pi_1(\psi_i)$. The Householder transformation maps ψ_i to $(1,0,\ldots,0)$. Note that a Householder transformation is orthonormal. Therefore a stereographic projection chart maps each tangent space orthogonally. Use of Householder transformations to results in a numerically stable implementation of the gradient computation.

For the second derivatives of J, by the chain rule,

$$\frac{\partial^2 J}{\partial \xi_i \partial \xi_j} = \frac{\partial^2 J}{\partial \psi_i \partial \psi_j}\Big(\frac{\partial \psi_i}{\partial \xi_i}, \frac{\partial \psi_j}{\partial \xi_j}\Big) + \sum_{i=1}^{n}\frac{\partial J}{\partial \psi_i}\cdot\frac{\partial^2 \psi_i}{\partial \xi_i^2}, \tag{4.32}$$

therefore we need to know $\partial^2\psi_i/\partial\xi_i^2$ and $\partial^2 J/\partial\psi_i\partial\psi_j$.

For $\partial^2\psi_i/\partial\xi_i^2$, a direct computation shows that

$$\begin{aligned}\frac{d^2\psi_i}{d\xi_i^2} = \Big(&-(1-\pi_1\psi_i)\big(2\pi_2\psi_i^*\otimes\pi_2\psi_i^* - (1-\pi_1\psi_i)\langle\cdot,\cdot\rangle\big),\\ &\big(2\pi_2\psi_i^*\otimes\pi_2\psi_i^* - (1-\pi_1\psi_i)\langle\cdot,\cdot\rangle\big)\pi_2\psi_i\\ &-(1-\pi_1\psi_i)\big(I_m\otimes\pi_2\psi_i^* + \pi_2\psi_i^*\otimes I_m\big)\Big),\end{aligned} \tag{4.33}$$

where $\langle\cdot,\cdot\rangle$ is dot product in bilinear form.

Now we compute $\partial^2 J/\partial\psi_i\partial\psi_j$. For the block-diagonal components, i.e., when $i = j$,

$$\frac{\partial^2 J}{\partial\psi_i^2} = \sum_{j\neq i} D_{ij}\langle\cdot,\psi_j\rangle\langle\cdot,\psi_j\rangle. \tag{4.34}$$

For the off-diagonal components, i.e., when $i \neq j$,

$$\frac{\partial^2 J}{\partial\psi_i\partial\psi_j} = \frac{\partial}{\partial\psi_j}\sum_{k\neq i} D_{ik}\langle\psi_i,\psi_k\rangle\langle\cdot,\psi_k\rangle \tag{4.35}$$

$$= D_{ij}\Big(\langle\psi_i,\psi_j\rangle\langle\cdot,\cdot\rangle + \langle\cdot,\psi_j\rangle\langle\psi_i,\cdot\rangle\Big). \tag{4.36}$$

4.5.3 Switching stereographic projections

A stereographic projection has a singularity at its focal point, i.e., the projection is not well-defined at the focal point. We circumvented the problem above by flipping the normalized quantum state variables ψ that are close to the focal point into $-\psi$. See (4.28) for the formulae for ξ.

An alternative way to handle the singularity is to use another stereographic projection map, and switch between the projections as needed. Let φ_s^{-1} denote the stereographic projection with the focal point at $(-1, 0, \ldots, 0)$. If $U_s = \mathbb{S}^m - \{(-1, 0, \ldots, 0)\}$, then

$$\varphi_s^{-1} : \psi \in U_s \longmapsto \xi_s = \frac{\pi_2\psi}{1+\pi_1\psi}, \tag{4.37}$$

and its projection inverse $\varphi_s\colon \mathbb{R}^m \to U_s$,

$$\varphi_s\colon \xi_s \longmapsto \psi = \left(-1 + \frac{2}{1+|\xi_s|^2}, \frac{2\xi_s}{1+|\xi_s|^2}\right). \tag{4.38}$$

Note that the domain of φ and φ_s are U and U_s, respectively, and

$$U \cup U_s = (\mathbb{S}^m - \{(1, 0, \ldots, 0)\}) \cup (\mathbb{S}^m - \{(-1, 0, \ldots, 0)\}) = \mathbb{S}^m, \tag{4.39}$$

hence these two projections φ and φ_s can map every point in $\mathbb{S}^m$ into $\mathbb{R}^m$. One may check that

$$\xi = \frac{\xi_s}{|\xi_s|^2} \quad \text{and} \quad \xi_s = \frac{\xi}{|\xi|^2}. \tag{4.40}$$

As mentioned, we may choose to switch between the projections to avoid the singularities, instead of to flip the state vectors. Then Step 3 may be modified as follows.

Step 3' After an update, check if any state variable needs to change its projection. For $i = 1, 2, \ldots, n$, if $|\xi_i| > R$, then replace ξ_i by $\xi_i/|\xi_i|^2$. Use φ_s for ξ_i instead of φ, or vice versa if ξ_i were using φ_s.

4.6 Experiments

The figures in this section use color to denote different clusters and classes in the visulization. Please see the electronic version for the color information in the figures.

4.6.1 Three circles dataset

We apply quantum state optimization to synthetic three circles data. A similar dataset used in Grikschat, Costa, Hero, and Michel (2006).

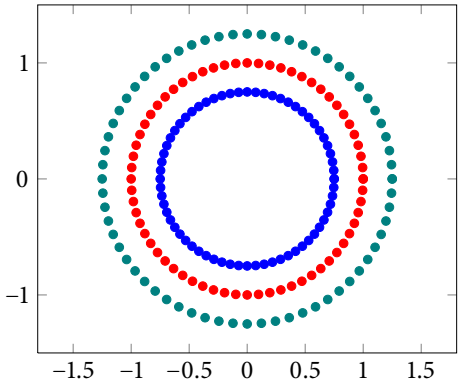
The data in its original feature space is shown in Figure 4.3a. It contains three thin layers of concentric circles. Each circle has 65 data points and the total number of points is 195. This data is difficult to split with spectral clustering. For instance, see Laplacian Eigenmaps result in Figure 4.3b. The inner circle points (•) are separated from the others but the middle circle points (•) and the outer circle points (•) are not separated.

We use power-weighted shortest path lengths from Chapter 2 to measure the dissimilarities in this data. The graph weights were given by $w(x, y) = \cosh(20|x - y|) - 1$. See Section 2.4.1 for super-additive power-weighted shortest paths.

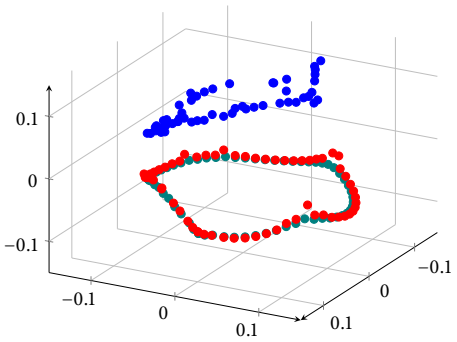
In Figure 4.4, the state vectors ψ_i and the associate probabilities p_{ij} are plotted when $H = \mathbb{R}^3$. The color code agrees with that used in Figure 4.3a. The probability matrix is nearly block-diagonal and is close to a perfect split. In Figure 4.4a the state vectors form an orthonormal basis of $H = \mathbb{R}^3$.

In Figure 4.5, the state vectors ψ_i and the associate probabilities p_{ij} are plotted when $H = \mathbb{R}^2$. The color code still agrees with that used in Figure 4.3a. Since the dimensionality is less than the number of circles, some correlation is introduced across the points in different circles. In Figure 4.5b the most inter-class correlations occur between the first 65 points (inner circle •) and the second 65 points (middle circle •), while the least inter-class correlations occur between the first 65 points (inner circle •) and the last 65 points (outer circle •). This is as expected since the inner circle has the smallest scatter and the outer circle has the largest scatter, in terms of summands in Equation 4.4.

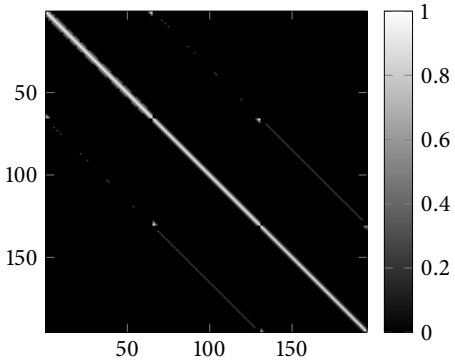
If the dimension becomes greater than three, the state vectors begin to spread. The increase in the state space dimensionality allows state vectors in single cluster to span more than one-dimensional subspaces while maintaining orthogonality against the state vectors



(a) Original feature space in $\mathbb{R}^2$.

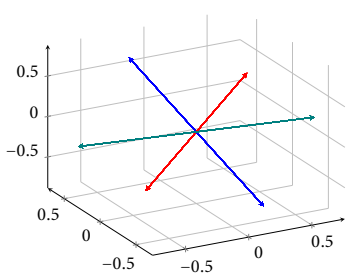


(b) Laplacian Eigenmaps in $\mathbb{R}^3$.

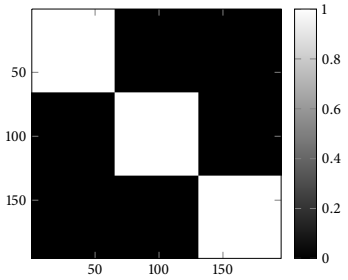


(c) Gaussian heat kernel matrix used for Laplacian Eigenmaps.

Figure 4.3: Three circles data. (a) is the raw form of the data. (b) is Laplacian Eigenmaps with three eigenvectors. (c) is the Gaussian heat kernel matrix used as the weight matrix for the Laplacian Eigenmaps result in (b). Unlike Figure 4.4b, the cluster structure is not obvious from (c).

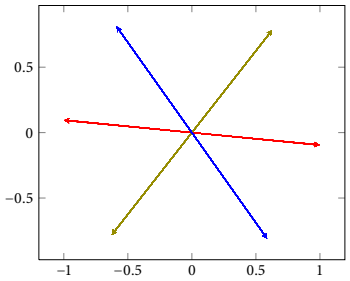


(a) Quantum states ψ_i.

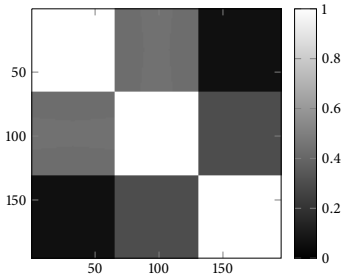


(b) Probability matrix p_{ij}.

Figure 4.4: Three-dimensional states, $H = \mathbb{R}^3$. Quantum states and association probability matrix of the three circles data. Blue arrows for the first 65 points (inner circle in Figure 4.3a), red arrows for the second 65 points (middle circle in Figure 4.3a), and teal arrows for the last 65 points (outer circle in Figure 4.3a). Unlike Figure 4.3b, the data points from different circles are well separated in terms of their state vectors. (a) shows that the state vectors from different circles are nearly orthogonal. (b) confirms the near orthogonality of the state vectors by the block-diagonal structure of the probability matrix. The diagonal blocks have the association probability $p_{ij} \approx 1$ while the off-diagonal blocks have $p_{ij} \approx 0$.



(a) Quantum states ψ_i.



(b) Probability matrix p_{ij}.

Figure 4.5: Two-dimensional states, $H = \mathbb{R}^2$. Quantum states and association probability matrix of the three circles data. Blue arrows for the first 65 points, red arrows for the second 65 points, and teal arrows for the last 65 points.

in other clusters. That is, one of the clusters among the three become a two-dimensional cluster in the Hilbert space.

4.6.2 Other datasets

The quantum state optimization was introduced as a relaxed version of the k-means in Section 4.2.1. We examine its coherence and differences from the conventional centroid approach to the k-means. There are many ways to quantitatively evaluate cluster analysis

results. Here we use Rand index (Rand 1971),

$$\text{Rand}(Y, \hat{Y}) = \frac{A + D}{A + B + C + D} \tag{4.41}$$

where Y and $\hat{Y}$ are two cluster labels,

- A is the number of pairs i, j such that $Y_i = Y_j$ and $\hat{Y}_i = \hat{Y}_j$,
- B is the number of pairs i, j such that $Y_i = Y_j$ and $\hat{Y}_i \neq \hat{Y}_j$,
- C is the number of pairs i, j such that $Y_i \neq Y_j$ and $\hat{Y}_i = \hat{Y}_j$,
- D is the number of pairs i, j such that $Y_i \neq Y_j$ and $\hat{Y}_i \neq \hat{Y}_j$.

First we compare the difference between the quantum state optimization and the k-means with conventional centroid-based implementation. For the implementation of the conventional k-means, we use the `kmeans` function from MATLAB statistics toolbox. For the partition operator in the proposed quantum cluster analysis, we use the mean projection operator,

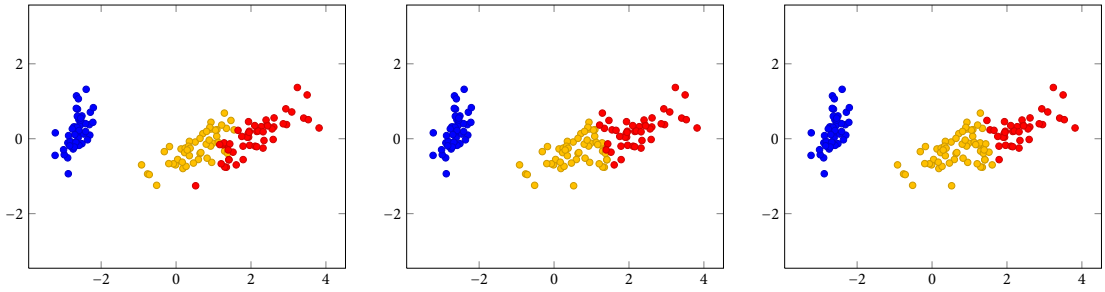
$$\Omega = \frac{1}{n} \sum_{i=1}^{n} \psi_i \psi_i^*. \tag{4.42}$$

We pick iris datasets from UCI machine learning repository.[4] The data has 150 instances of four continuous variables, and has three classes where each class refers to a type of iris plant. One class is linearly separable from the others, while the other two classes are not easily separable. Figure 4.6 shows a comparison and visualization. Euclidean distance is used for the both methods. Note that the quantum state optimization improved slightly over the centroid approach in terms of the Rand index, from 0.8797 to 0.8923.

We also carry out the same comparison for Wisconsin diagnostic breast cancer dataset from UCI machine learning repository. This dataset has 569 instances and 32 variables. Each instance is either malignant or benign. Euclidean distance is used again for the both methods. See Figure 4.7 for visualization and a comparison. Like iris dataset comparison, the quantum state optimization improves over the centroid approach. The Rand index rises to 0.8308 from 0.7504. The improvement is much more significant for this dataset.

For the two datasets from UCI repository, Euclidean distances were used. However, the quantum state optimization may work with a broad general class of dissimilarities of its input data, hence a quantum cluster analysis may benefit from a careful choice of dissimilarity.

[4] http://archive.ics.uci.edu/ml/ (Frank and Asuncion 2010)

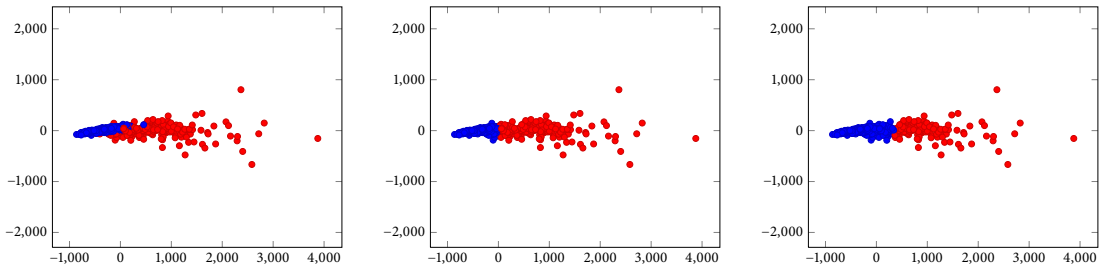


(a) True labels

(b) Labels from the quantum state optimization. The Rand index is 0.8923.

(c) Labels from the k-means. The Rand index is 0.8797.

Figure 4.6: Cluster analysis comparison for iris dataset from UCI repository. The employed dissimilarity measure is the Euclidean metric. For IRIS dataset, the quantum state optimization performs slightly better than the conventional centroid-based k-means. The data is projected onto the plane by MDS for visualization purpose.

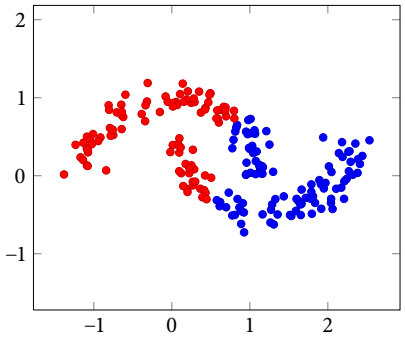


(a) True labels

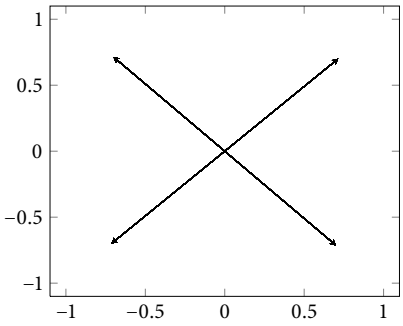
(b) Labels from the quantum state optimization. The Rand index is 0.8308.

(c) Labels from the k-means. The Rand index is 0.7504.

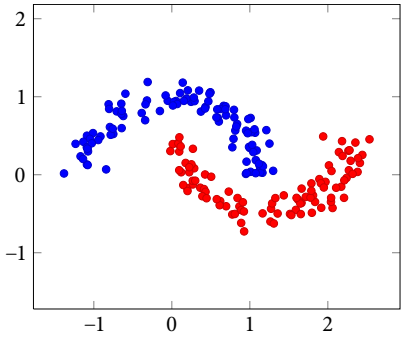
Figure 4.7: Cluster analysis comparison for Wisconsin diagnostic breast cancer dataset from UCI repository. The employed dissimilarity measure is the Euclidean metric. The quantum state optimization significantly improved the cluster analysis result. The data is projected onto the plane by MDS for visualization purpose.



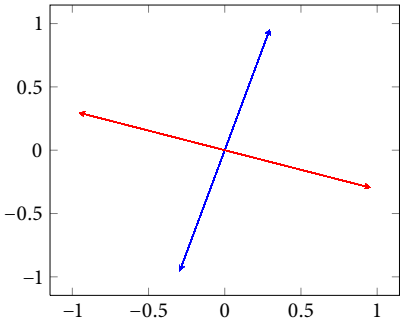
(a) Partition with Euclidean distance.



(b) Quantum states optimized with Euclidean distance.



(c) Partition with shortest path distance.



(d) Quantum states optimized with shortest path distance.

Figure 4.8: Quantum cluster analysis results for two-moons data. (a) and (b) are the optimized quantum states and partition result based on Euclidean distance. The Rand index score is 0.5047. (c) and (d) are the optimized quantum states and partition result based on super-additive shortest path distance. The Rand index score is 1, which indicates the perfect partition.

Figure 4.8 illustrates an example. The data shown in the figure is often called two-moons or half-circles data. This dataset has been used to demonstrate non-convex point distribution in machine learning methods. The figures on the top row visualize the optimized state vectors $\{\psi_i\}_i$ based on Euclidean distance, and a partition result based on the state vectors. The cluster analysis result is not different from the k-means partition with Euclidean distance. This is as anticipated since the quantum state optimization is a relaxed version of the k-means.

When the Euclidean distance is replaced by super-additive shortest path lengths, Figure 4.8c and Figure 4.8d shows an improved partition result, and in fact, the perfect partition is achieved.

4.7 Further discussions

4.7.1 Connections to spectral clustering

We compare our clustering scheme with spectral clustering (Shi and Malik 2000; Ng *et al.* 2002; von Luxburg 2007). Spectral clustering finds the eigenvalues and the eigenvectors of the graph Laplacian matrix. Similarity weights $a_{ij} = a_{ji} \geq 0$ for $i \neq j$ are provided. The (unnormalized) graph Laplacian L—having ij-th element L_{ij}—is

$$L_{ij} = \begin{cases} d_i, & i = j, \\ -a_{ij}, & i \neq j, \end{cases} \tag{4.43}$$

where $d_i = \sum_{j \neq i} a_{ij}$.

Let $e_1, e_2, \ldots, e_n \in \mathbb{R}^n$ be the standard orthonormal sequence. Choose

$$\psi_i = \frac{1}{\sqrt{d_i}} e_i \wedge \sum_{j \neq i} e_j \sqrt{a_{ij}} \tag{4.44}$$

in $\mathbb{R}^d \wedge \mathbb{R}^d$ for $i = 1, 2, \ldots, n$. The factor $\sqrt{d_i}$ is introduced to normalize the state vector ψ_i. Then

$$L'_{ij} = \langle \psi_i, \psi_j \rangle = \begin{cases} 1, & i = j, \\ -\dfrac{a_{ij}}{\sqrt{d_i d_j}}, & i \neq j. \end{cases} \tag{4.45}$$

This Gram matrix L'—having ij-th element L'_{ij}—of the normalized quantum state vectors $\{\psi_i\}_i$ is known as the normalized graph Laplacian.

Choose L' as the partition operator. Let $\lambda_\alpha > 0$ be an eigenvalue of L' and let $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ denote the corresponding eigenvector. Note that $\alpha^* L' \alpha = \lambda_\alpha \alpha^* \alpha$, where α^* denotes the transpose of α. Then the probability that a state vector ψ_i collapses into the cluster α is

$$\frac{1}{\lambda_\alpha} \left| \left\langle \sum_j \alpha_j \psi_j, \psi_i \right\rangle \right|^2 = \frac{1}{\lambda_\alpha} \left| \left\langle \sum_j \alpha_j e_j, L' e_i \right\rangle \right|^2 = \lambda_\alpha |\alpha_i|^2 \tag{4.46}$$

where the inner product in $\mathbb{R}^d$ is the Euclidean dot product.

Spectral clustering manifests in many different forms (Shi and Malik 2000; Ng *et al.* 2002; Zelnik-Manor and Perona 2005; von Luxburg 2007). One form of spectral clustering is to look at the sign of eigenvector elements (von Luxburg, Belkin, and Bousquet 2008). That is,

an eigenvector α induces two clusters, and the cluster into which X_i falls is determined by the sign of α_i for each $i = 1, 2, \ldots, n$. Since $\lambda_\alpha > 0$, the sign of α_i is the same as the sign of $\lambda_\alpha \alpha_i$. Therefore the cluster assignment is determined by the sign of

$$\lambda_\alpha \alpha_i = \left\langle \sum_j \alpha_j e_j, L' e_i \right\rangle = \left\langle \sum_j \alpha_j \psi_j, \psi_i \right\rangle. \tag{4.47}$$

(4.46) and (4.47) illustrate the main difference between quantum cluster analysis and spectral clustering. The former puts emphasis on the squared magnitude of the inner product values (projective geometry), while the latter discriminates by the simple inner product without magnitude squaring.

4.7.2 Quantum state optimization and couplings

In this subsection, we connect quantum cluster analysis to a coupling problem. For a general introduction to the coupling problems, we refer the reader to Villani (2009).

Let $\mathcal{X}_0, \mathcal{X}_1$ be measurable spaces, μ_0, μ_1 be positive measures in $\mathcal{X}_0, \mathcal{X}_1$, respectively. A *coupling* of μ_0 and μ_1 is a positive measure π in $\mathcal{X}_0 \times \mathcal{X}_1$ such that

$$\pi(A \times \mathcal{X}_1) = \mu_0(A) \quad \text{and} \quad \pi(\mathcal{X}_0 \times B) = \mu_1(B) \tag{4.48}$$

for all measurable $A \subset \mathcal{X}_0$, $B \subset \mathcal{X}_1$. If X_0, X_1 are random variables with probability law μ_0, μ_1, respectively, then a coupling is a joint probability law of X_0 and X_1. If there exists a map $T: \mathcal{X}_0 \to \mathcal{X}_1$ such that $T(X_0) = X_1$, then the coupling is said to be *deterministic*.

Set $\mathcal{X}_0 = \mathcal{X}_1 = \mathcal{X}$ and $\mu_0 = \mu_1 = \mu$. We will call this special case as an *auto-coupling* problem. If an auto-coupling is deterministic, it is a measure-preserving transformation in the ergodic theory. The problem may be phrased as follows. Suppose X_0 is a random variable in $\mathcal{X}$. An auto-coupling is to generate another random variable X_1 dependent on X_0 so that the marginal probability law of X_1 is identical to that of X_0.

This auto-coupling problem has both practical and theoretical deficiencies.

1. In practice, the probability law μ of X_0 is unknown.

2. There are at least two trivial solutions. One is the independent sampling, i.e., X_0 and X_1 are independent and identically distributed. Another is the identity solution $X_1 = X_0$.

We need more constraints or structure in the coupling model to remove these deficiencies.

The first deficiency of unknown probability law is not difficult to remove. Suppose that the coupling π is symmetric, i.e.,

$$\pi(A \times B) = \pi(B \times A) \tag{4.49}$$

for all measurable subsets $A, B \subset \mathcal{X}$. Substitute $\mathcal{X}$ for B. Then the marginal probability law of X_0 and X_1 are identical.

For the second deficiency of trivial solutions, the approach of quantum cluster analysis is to use the state space H. Assume H is a separable Hilbert space. Suppose there exists a measurable function from $\mathcal{X}$ to H, and let $\psi_i \in H$ be the unit state vector associated to $x_i \in \mathcal{X}$. If Ω is a quantum observation, i.e., a self-adjoint operator of H and $\omega_1, \omega_2, \cdots \in H$ is a countable orthonormal basis of eigenvectors of Ω, then the probability that X_0 and X_1 are observed to be in the same cluster conditioned on that $X_i = x_i$ in quantum cluster analysis is

$$|\langle \psi_0, \psi_1 \rangle|^2 = \left| \sum_j \langle \psi_0, \omega_j \rangle \langle \psi_1, \omega_j \rangle \right|^2. \tag{4.50}$$

Note that if X_0, X_1 were independent, then the probability that X_0 and X_1 collapse to the same observation would have been

$$\sum_j |\langle \psi_0, \omega_j \rangle|^2 |\langle \psi_1, \omega_j \rangle|^2 \tag{4.51}$$

and the quantum state model imposes dependency between points.

Quantum state optimization determines the mapping into the state space H. It is a variation of the optimal transport problem, which is defined as follows. Suppose there exists a cost function $D: \mathcal{X} \times \mathcal{X} \to (-\infty, +\infty]$. Then the optimal transport problem is to find a joint measure π which minimizes

$$E_\pi D(X_0, X_1) \tag{4.52}$$

where E_π denote the expectation under the probability law π. The minimal expectation $E_\pi D(X_0, X_1)$ is also called the earth mover's distance. If D satisfies a regularity condition $D(x_0, x_0) \leq D(x_0, x_1)$ for all $x_0, x_1 \in \mathcal{X}$, then the optimal transport problem is trivial by setting $X_1 = X_0$. In terms of the state space map, the trivial minimization occurs when $\langle \psi_0, \psi_1 \rangle = 0$ for all $x_0 \neq x_1$. If $\mathcal{X}$ were an open subset of some real or complex vector space, then $\mathcal{X}$ is uncountable and any map of $\mathcal{X}$ into a separable Hilbert space H must have pairs of points $x_0 \neq x_1$ in $\mathcal{X}$ such that $\langle \psi_0, \psi_1 \rangle \neq 0$ since H has at most countable orthonormal sequences.

Chapter 5
Conclusion

Differential geometry may serve as a theoretical ground for a large class of statistical inference problems which emphasize the local neighborhoods. This thesis provides several examples in manifold learning, information theory, cluster analysis, and diffusion dynamics.

In Chapter 2, we proved the convergence properties of power-weighted and super-additive shortest path lengths. The path lengths converged to Riemannian distances under a class of conformal deformations of the manifold where the data points were sampled. The conformal deformation provided a hybrid measure of geometry and statistics. The convergence proofs were provided in several domains e.g., compact manifolds, complete manifolds, and embedded manifolds. The results from the chapter holds theoretical importance in random graph theory as an extension of the Beardwood-Halton-Hammersley theorem.

In Chapter 3, a formal theory of information geometry was developed. The main object of study in information theory is the space of probability measures. Information geometry presents a differential geometric view for the measure space. We focused on L^2 interpretation of the measures through Radon-Nikodym derivatives and exponentiations. The theory also provided a link between parameterizations and non-parametrics.

One of major theories where probability measures have L^2 representations is quantum mechanics. This motivates and leads to the topic of the quantum cluster analysis.

In Chapter 4, we proposed quantum cluster analysis framework. The framework used projective spaces as the state space to model the dependencies between the data points. The presented framework provides physical interpretations of data clustering procedures, and the quantum state optimization algorithm bridges the notion of dissimilarities or metrics to the notion of similarities or random walk transition probabilities. The optimization algorithm also provides a general strategy for the gradient descent methods over spherical domains.

There are many related and open problems for future work. A natural conjecture which follows from the convergence result developed in Chapter 2, is that some other minimal

graph problems such as minimal cuts in power-weighted graphs may satisfy similar asymptotic convergence to certain geometric quantities in Riemannian geometry. A successful development of the theoretical connection between graphs and deformations will lead to a better understanding of graphs on manifolds.

Chapter 4 posed the unsupervised clustering problem in quantum mechanics. However, we have not yet introduced the Schrödinger's equation into the context of the quantum cluster analysis, which is the central equation is quantum physics. Nadler *et al.* (2006) explains the relationship of the diffusion map with the Schrödinger operator. And indeed the quantum state association probabilities p_{ij} presented are similar to the Markov chain transition probabilities in the diffusion maps. By a connection between the quantum state optimization and the Schrödinger's equation, the quantum cluster framework may be reinforced as a unified framework which integrates cluster analysis and spectral analysis.

Another area for future work is a connection of quantum cluster framework with the optimal transports. The optimal transport problem is one of mathematical theories in active research with many practical applications. In Section 4.7.2, the quantum state optimization was presented as an alternative problem to the optimal transport. A successful future work in this perspective may bring the theories already developed in the context of coupling problems to machine learning, in a different perspective from the earth-mover distance.

Bibliography

S. M. Ali and S. D. Silvey (1966). "A general class of coefficients of divergence of one distribution from another". *Journal of the Royal Statistical Society, Series B (Methodological)* 28.1, pp. 131–142. JSTOR: 2984279.

Shun-Ichi Amari (2009). "α-divergence is unique, belonging to both f-divergence and Bregman divergence classes". *IEEE Transactions on Information Theory* 55.11, pp. 4925–4931. DOI: 10.1109/TIT.2009.2030485.

Shun'ichi Amari and Hiroshi Nagaoka (2000). *Methods of information geometry*. Trans. by D. Harada. Translations of Mathematical Monographs 191. American Mathematical Society.

Adrian Baddeley (2007). "Spatial point processes and their applications". In: *Stochastic geometry*. Ed. by W. Weil. Lecture Notes in Mathematics 1892. Lectures given at the C.I.M.E. Summer School held in Martina Franca, Italy, September 13--18, 2004. Springer. DOI: 10.1007/3-540-38174-0.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh (2005). "Clustering with Bregman divergences". *Journal of Machine Learning Research* 6, pp. 1705–1749. URL: http://jmlr.csail.mit.edu/papers/v6/banerjee05b.html.

Jillian Beardwood, J. H. Halton, and J. M. Hammersley (1959). "The shortest path through many points". In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 55, pp. 299–327. DOI: 10.1017/S0305004100034095.

Mikhail Belkin and Partha Niyogi (2003). "Laplacian eigenmaps for dimensionality reduction and data representation". *Neural Computation* 15.6, pp. 1373–1396. DOI: 10.1162/089976603321780317.

— (2008). "Towards a theoretical foundation for Laplacian-based manifold methods". *Journal of Computer and System Sciences* 74.8, pp. 1289–1308. DOI: 10.1016/j.jcss.2007.08.006.

Mira Bernstein, Vin de Silva, John C. Langford, and Joshua B. Tenenbaum (20, 2000). "Graph approximations to geodesics on embedded manifolds". URL: http://web.mit.edu/cocosci/isomap/isomap.html (visited on 08/03/2012).

Arthur L. Besse (1987). *Einstein manifolds*. Ergebnisse der Mathematik und ihrer Grenzgebiete 10. Springer.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet allocation". *Journal of Machine Learning Research* 3, pp. 993–1022. URL: http://jmlr.csail.mit.edu/papers/v3/blei03a.html.

L. M. Bregman (1967). "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming". Trans. by

G. Kiss. *USSR Computational Mathematics and Mathematical Physics* 7.3, pp. 200–217. DOI: 10.1016/0041-5553(67)90040-7.

Kevin M. Carter, Raviv Raich, William G. Finn, and Alfred O. Hero III (2011). "Information geometric dimensionality reduction". *IEEE Signal Processing Magazine* 28.2: *Dimensionality Reduction. Via Subspace and Submanifold Learning*, pp. 89–99. DOI: 10.1109/MSP.2010.939536.

Kevin M. Carter, Raviv Raich, and Alfred O. Hero III (2009a). "FINE: Fisher information nonparameteric embedding". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.11, pp. 2093–2098. DOI: 10.1109/TPAMI.2009.67. arXiv: 0802.2050v1 [stat.ML].

— (2009b). "Spherical Laplacian information maps (SLIM) for dimensionality reduction". In: *IEEE/SP 15th Workshop on Statistical Signal Processing, 2009. SSP '09.* (Cardiff, UK, Aug. 31–Sept. 3, 2009), pp. 405–408. DOI: 10.1109/SSP.2009.5278554.

Olivier Chapelle and Alexander Zien (2005). "Semi-supervised classification by low density separation". In: *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*. (Bridgetown, Barbados, Jan. 6–8, 2005). Ed. by R. G. Cowell and Z. Ghahramani, pp. 57–64. URL: http://www.gatsby.ucl.ac.uk/aistats/.

Yilun Chen, Ami Wiesel, and Alfred O. Hero III (2011). "Robust shrinkage estimation of high-dimensional covariance matrices". *IEEE Transactions on Signal Processing* 59.9, pp. 4097–4107. DOI: 10.1109/TSP.2011.2138698. arXiv: 1009.5331v1 [stat.ME].

Ronald R. Coifman and Stéphane Lafon (2006). "Diffusion maps". *Applied and Computational Harmonic Analysis* 21.1, pp. 5–30. DOI: 10.1016/j.acha.2006.04.006.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein (2009). *Introduction to Algorithms*. 3rd ed. The MIT Press.

Jose A. Costa and Alfred O. Hero (2004). "Geodesic entropic graphs for dimension and entropy estimation in manifold learning". *IEEE Transactions on Signal Processing* 52.8, pp. 2210–2221. DOI: 10.1109/TSP.2004.831130.

Thomas M. Cover and Joy A. Thomas (2006). *Elements of information theory*. 2nd ed. Wiley. DOI: 10.1002/0471200611.

Trevor F. Cox and Michael A. A. Cox (2001). *Multidimensional scaling*. 2nd ed. Monographs on Statistics and Applied Probability 88. CRC Press.

I. Csiszár (1967). "Information type measures of difference of probability distributions and indirect observations". *Studia Scientiarum Mathematicarum Hungarica* 2, pp. 299–318.

I. Csiszár and P. C. Shields (2004). "Information theory and statistics: a tutorial". *Foundations and Trends in Communications and Information Theory* 1.4, pp. 417–528. DOI: 10.1561/0100000004.

Frank Deutsch (1982). "Linear selections for the metric projection". *Journal of Functional Analysis* 49.3, pp. 269–292. DOI: 10.1016/0022-1236(82)90070-2.

Inderjit S. Dhillon and Dharmendra S. Modha (2001). "Concept decompositions for large sparse text data using clustering". *Machine Learning* 42.1--2, pp. 143–175. DOI: 10.1023/A:1007612920971.

Manfredo Perdigão do Carmo (1992). *Riemannian geometry*. Trans. by F. Flaherty. Birkhäuser.

David L. Donoho and Carrie Grimes (2003). "Hessian eigenmaps: locally linear embedding techniques for high-dimensional data". *Proceedings of the National Academy*

of Sciences of the United States of America 100.10, pp. 5591–5596. DOI: 10.1073/pnas.1031596100.

Richard O. Duda, Peter E. Hart, and David G. Stork (2000). *Pattern classification*. 2nd ed. Wiley.

A. Frank and A. Asuncion (2010). *UCI machine learning repository*. URL: http://archive.ics.uci.edu/ml.

Steve Grikschat, Jose A. Costa, Alfred O. Hero III, and Olivier Michel (2006). "Dual rooted-diffusions for clustering and classification on manifolds". In: *2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings*. (Toulouse, France, May 14–19, 2006). DOI: 10.1109/ICASSP.2006.1661431.

Richard S. Hamilton (1982). "The inverse function theorem of Nash and Moser". *Bulletin (New Series) of the American Mathematical Society* 7.1, pp. 65–222. MR: 656198. URL: http://projecteuclid.org/euclid.bams/1183549049.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman (2009). *The elements of statistical learning. Data mining, inference, and prediction*. 2nd ed. Springer Series in Statistics. Springer. DOI: 10.1007/978-0-387-84858-7.

Alfred O. Hero III (2007). "Geometric entropy minimization (GEM) for anomaly detection and localization". In: *Advances in Neural Information Processing Systems 19*. (Vancouver, BC, Canada, Dec. 4–7, 2006). Ed. by B. Schölkopf, J. C. Platt, and T. Hoffman, pp. 585–592. URL: http://books.nips.cc/nips19.html.

Alfred Hero, Bing Ma, Olivier Michel, and John Gorman (2002). "Applications of entropic spanning graphs". *IEEE Signal Processing Magazine* 19.5: *Mathematics in Imaging*, pp. 85–95. DOI: 10.1109/MSP.2002.1028355.

David Horn and Assaf Gottlieb (2002). "The method of quantum clustering". In: *Advances in Neural Information Processing Systems 14*. (Vancouver, BC, Canada, Dec. 3–8, 2001). Ed. by T. G. Dietterich, S. Becker, and Z. Ghahramani, pp. 769–776. URL: http://books.nips.cc/nips14.html.

Robert E. Kass and Paul W. Vos (1997). *Geometrical foundations of asymptotic inference*. Wiley Series in Probability and Statistics 125. Wiley.

S. Kullback and R. A. Leibler (1951). "On information and sufficiency". *The Annals of Mathematical Statistics* 22.1, pp. 79–86. DOI: 10.1214/aoms/1177729694.

Serge Lang (1993). *Real and functional analysis*. Graduate Texts in Mathematics 142. Springer.

— (1999). *Fundamentals of differential geometry*. Graduate Texts in Mathematics 191. Springer.

Olivier Ledoit and Michael Wolf (2004). "A well-conditioned estimator for large-dimensional covariance matrices". *Journal of Multivariate Analysis* 88.2, pp. 365–411. DOI: 10.1016/S0047-259X(03)00096-4.

John M. Lee (2003). *Introduction to smooth manifolds*. Graduate Texts in Mathematics 218. Springer.

I. E. Leonard and K. Sundaresan (1974). "Smoothness and Duality in $L_p(E, \mu)$". *Journal of Mathematical Analysis and Applications* 46, pp. 513–522.

Wanqing Li, Zhengyou Zhang, and Zicheng Liu (2010). "Action recognition based on a bag of 3d points". In: *2010 IEEE Computer Society Conference on Computer Vision and*

Pattern Recognition Workshops (CVPRW). (San Francisco, CA, USA, June 13–18, 2010), pp. 9–14. DOI: 10.1109/CVPRW.2010.5543273.

Stuart P. Lloyd (1982). "Least squares quantization in PCM". *IEEE Transactions on Information Theory* 28.2, pp. 129–137. DOI: 10.1109/TIT.1982.1056489.

J. MacQueen (1967). "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. (Berkeley, CA, USA). Ed. by L. M. Le Cam and J. Neyman, pp. 281–297. MR: 0214227.
URL: http://projecteuclid.org/euclid.bsmsp/1200512992.

Ronald Meester and Rahul Roy (1996). *Continuum percolation*. Cambridge University Press.

J. Milnor (1963). *Morse theory*. Princeton University Press.

Shigeyuki Morita (2001). *Geometry of differential forms*. Trans. by T. Nagase and K. Nomizu. Translations of Mathematical Monographs 201. American Mathematical Society.

Boaz Nadler, Stéphane Lafon, Ronald R. Coifman, and Ioannis G. Kevrekidis (2006). "Diffusion maps, spectral clustering and reaction coordinates of dynamical systems". *Applied and Computational Harmonic Analysis* 21.1, pp. 113–127.
DOI: 10.1016/j.acha.2005.07.004.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss (2002). "On spectral clustering: analysis and an algorithm". In: *Advances in Neural Information Processing Systems 14*. (Vancouver, BC, Canada, Dec. 3–8, 2001). Ed. by T. G. Dietterich, S. Becker, and Z. Ghahramani, pp. 849–856. URL: http://books.nips.cc/nips14.html.

Barrett O'Neill (1983). *Semi-Riemannian geometry*. Academic Press.

Mathew D. Penrose and J. E. Yukich (2003). "Weak laws of large numbers in geometric probability". *The Annals of Applied Probability* 13.1, pp. 277–303.
DOI: 10.1214/aoap/1042765669.

— (2011). *Limit theory for point processes in manifolds*. arXiv: 1104.0914v1 [math.PR].

Ziad Rached, Fady Alajaji, and L. Lorne Campbell (2001). "Rényi divergence and entropy rates for finite alphabet Markov sources". *IEEE Transactions on Information Theory* 47.4, pp. 1553–1561. DOI: 10.1109/18.923736.

— (2004). "The Kullback-Leibler divergence rate between Markov sources". *IEEE Transactions on Information Theory* 50.5, pp. 917–921.
DOI: 10.1109/TIT.2004.826687.

William M. Rand (1971). "Objective criteria for the evaluation of clustering methods". *Journal of the American Statistical Association* 66.336, pp. 846–850. JSTOR: 2284239.

Carl Edward Rasmussen (2000). "The infinite Gaussian mixture model". In: *Advances in Neural Information Processing Systems 12*. (Denver, CO, USA, Nov. 29–Dec. 4, 1999). Ed. by S. A. Solla, T. K. Leen, and K.-R. Müller, pp. 554–560.
URL: http://books.nips.cc/nips12.html.

Alfréd Rényi (1961). "On measures of entropy and information". In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. (Berkeley, CA, USA, June 20–July 30, 1960). Ed. by J. Neyman, pp. 547–561.
URL: http://projecteuclid.org/euclid.bsmsp/1200512181.

Wansoo T. Rhee (1993). "A matching problem and subadditive Euclidean functionals". *The Annals of Applied Probability* 3.3, pp. 794–801. DOI: 10.1214/aoap/1177005364.

Sam T. Roweis and Lawrence K. Saul (2000). "Nonlinear dimensionality reduction by locally linear embedding". *Science* 290.5500, pp. 2323–2326. DOI: 10.1126/science.290.5500.2323.

Walter Rudin (1986). *Real and complex analysis*. 3rd ed. McGraw-Hill.

— (1991). *Functional analysis*. 2nd ed. McGraw-Hill.

J. J. Sakurai (1993). *Modern quantum mechanics*. Revised ed. Addison-Wesley.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson (2001). "Estimating the support of a high-dimensional distribution". *Neural Computation* 13.7, pp. 1443–1471. DOI: 10.1162/089976601750264965.

R. Shankar (1994). *Principles of quantum mechanics*. 2nd ed. Springer.

C. E. Shannon (1948). "A mathematical theory of communication". *Bell System Technical Journal* 27, pp. 379–423. URL: http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html.

Jianbo Shi and Jitendra Malik (2000). "Normalized cuts and image segmentation". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.8, pp. 888–905. DOI: 10.1109/34.868688.

J. Michael Steele (1981). "Subadditive Euclidean functionals and nonlinear growth in geometric probability". *The Annals of Probability* 9.3, pp. 365–376. DOI: 10.1214/aop/1176994411.

— (1988). "Growth rates of Euclidean minimal spanning trees with power weighted edges". *The Annals of Probability* 16.4, pp. 1767–1787. DOI: 10.1214/aop/1176991596.

— (1997). *Probability theory and combinatorial optimization*. CBMS-NSF Regional Conference Series in Applied Mathematics 69. Society for Industrial and Applied Mathematics. DOI: 10.1137/1.9781611970029.

Michel Talagrand (1995). "Concentration of measure and isoperimetric inequalities in product spaces". *Publications Mathématiques de l'IHÉS* 81.1, 73:205. DOI: 10.1007/BF02699376. arXiv: math/9406212v1 [math].

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei (2006). "Hierarchical Dirichlet processes". *Journal of the American Statistical Association* 101.476, pp. 1566–1581. DOI: 10.1198/016214506000000302.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford (2000). "A global geometric framework for nonlinear dimensionality reduction". *Science* 290.5500, pp. 2319–2323. DOI: 10.1126/science.290.5500.2319.

Michael E. Tipping and M. Bishop Christopher (1999). "Mixtures of probabilistic principal compoent analyzers". *Neural Computation* 11.2, pp. 443–482. DOI: 10.1162/089976699300016728.

Constantino Tsallis (1988). "Possible generalization of Boltzmann-Gibbs statistics". *Journal of Statistical Physics* 52.1–2, pp. 479–487. DOI: 10.1007/BF01016429.

Cédric Villani (2009). *Optimal transport. Old and new*. Grundlehren der mathematischen Wissenschaften 338. Springer. DOI: 10.1007/978-3-540-71050-9.

Ulrike von Luxburg (2007). "A tutorial on spectral clustering". *Statistics and Computing* 17.4, pp. 395–416. DOI: 10.1007/s11222-007-9033-z. arXiv: 0711.0189v1 [cs.DS].

Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet (2008). "Consistency of spectral clustering". *The Annals of Statistics* 36.2, pp. 555–586. DOI: 10.1214/009053607000000640.

Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan (2012). "Mining actionlet ensemble for action recognition with depth cameras". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (Providence, RI, USA, June 16–21, 2012), pp. 1290–1297. DOI: 10.1109/CVPR.2012.6247813.

Kilian Q. Weinberger, Fei Sha, and Lawrence K. Saul (2004). "Learning a kernel matrix for nonlinear dimensionality reduction". In: *Proceedings of the 21st International Conference on Machine Learning*. (Banff, AB, Canada, July 4–8, 2004). Ed. by R. Greiner and D. Schuumans, pp. 106–113. DOI: 10.1145/1015330.1015345. URL: http://machinelearning.org/icml2004_proc.html.

J. E. Yukich (2000). "Asymptotics for weighted minimal spanning trees on random points". *Stochastic Processes and their Applications* 85.1, pp. 123–138. DOI: 10.1016/S0304-4149(99)00068-X.

Joseph E. Yukich (1998). *Probability theory of classical Euclidean optimization problems*. Lecture Notes in Mathematics 1675. Springer. DOI: 10.1007/BFb0093472.

Lihi Zelnik-Manor and Pietro Perona (2005). "Self-tuning spectral clustering". In: *Advances in Neural Information Processing Systems 17*. (Vancouver, BC, Canada, Dec. 13–18, 2004). Ed. by L. K. Saul, Y. Weiss, and L. Bottou, pp. 1601–1608. URL: http://books.nips.cc/nips17.html.

Jun Zhang (2004). "Divergence function, duality, and convex analysis". *Neural Computation* 16.1, pp. 159–195. DOI: 10.1162/08997660460734047.