Geometrical Aspects of Statistical Learning Theory

Vom Fachbereich Informatik
der Technischen Universität Darmstadt
genehmigte

Dissertation

zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.)
vorgelegt von
Dipl.-Phys.

Matthias Hein

aus Esslingen am Neckar

Prüfungskommission:

Vorsitzender:	Prof. Dr. B. Schiele
Erstreferent:	Prof. Dr. T. Hofmann
Korreferent :	Prof. Dr. B. Schölkopf

Tag der Einreichung: 30.9.2005
Tag der Disputation: 9.11.2005

Darmstadt, 2005
Hochschulkennziffer: D17

Abstract

Geometry plays an important role in modern statistical learning theory, and many different aspects of geometry can be found in this fast developing field. This thesis addresses some of these aspects. A large part of this work will be concerned with so called manifold methods, which have recently attracted a lot of interest. The key point is that for a lot of real-world data sets it is natural to assume that the data lies on a low-dimensional submanifold of a potentially high-dimensional Euclidean space. We develop a rigorous and quite general framework for the estimation and approximation of some geometric structures and other quantities of this submanifold, using certain corresponding structures on neighborhood graphs built from random samples of that submanifold. Another part of this thesis deals with the generalization of the maximal margin principle to arbitrary metric spaces. This generalization follows quite naturally by changing the viewpoint on the well-known support vector machines (SVM). It can be shown that the SVM can be seen as an algorithm which applies the maximum margin principle to a subclass of metric spaces. The motivation to consider the generalization to arbitrary metric spaces arose by the observation that in practice the condition for the applicability of the SVM is rather difficult to check for a given metric. Nevertheless one would like to apply the successful maximum margin principle even in cases where the SVM cannot be applied. The last part deals with the specific construction of so called Hilbertian metrics and positive definite kernels on probability measures. We consider several ways of building such metrics and kernels. The emphasis lies on the incorporation of different desired properties of the metric and kernel. Such metrics and kernels have a wide applicability in so called kernel methods since probability measures occur as inputs in various situations.

Zusammenfassung

Geometrie spielt eine wichtige Rolle in der modernen statistischen Lerntheorie. Viele Aspekte der Geometrie können in diesem sich schnell entwickelnden Feld gefunden werden. Diese Dissertation beschäftigt sich mit einigen dieser Aspekte. Ein großer Teil dieser Arbeit befasst sich mit sogenannten Mannigfaltigkeits-Methoden. Die Hauptmotivation liegt darin, daß es für Datensätze in Anwendungen eine in vielen Fällen zutreffende Annahme ist, daß die Daten auf einer niedrig-dimensionalen Untermannigfaltigkeit eines potentiell hoch-dimensionalen Euklidischen Raumes liegen. In dieser Arbeit wird ein mathematisch strenger und allgemeiner Rahmen für die Schätzung und Approximation von geometrischen Strukturen und anderen Größen der Untermannigfaltigkeit entwickelt. Dazu werden korrespondierende Strukturen auf einem durch eine Stichprobe von Punkten der Untermannigfaltigkeit erzeugten Nachbarschaftsgraphen genutzt. Ein weiterer Teil dieser Dissertation behandelt die Verallgemeinerung des sogenannten „maximum-margin“-Prinzips auf allgemeine metrische Räume. Durch eine neue Sichtweise auf die sogenannte „support vector machine“(SVM) folgt diese Verallgemeinerung auf natürliche Weise. Es wird gezeigt, daß die SVM als ein Algorithmus gesehen werden kann, der das „maximum-margin“-Prinzip auf eine Unterklasse von metrischen Räumen anwendet. Die Motivation für diese Verallgemeinerung entstand durch das in der Praxis häufig auftretende Problem, daß die Bedingungen für die Verwendung einer bestimmten Metrik in der SVM schwer zu überprüfen sind. Trotzdem würde man gerne selbst in Fällen in denen die SVM nicht angewendet werden kann das erfolgreiche „maximum-margin“-Prinizp verwenden. Der abschließende Teil dieser Arbeit beschäftigt sich mit der speziellen Konstruktion von sogenannnten Hilbert'schen Metriken und positiv definiten Kernen auf Wahrscheinlichkeitsmaßen. Mehrere Möglichkeiten solche Metriken und Kerne zu konstruieren werden untersucht. Der Schwerpunkt liegt dabei auf der Integration verschiedener gewünschter Eigenschaften in die Metrik bzw. den Kern. Solche Metriken und Kerne haben vielfältige Anwendungsmöglichkeiten in sogenannten Kern-Methoden, da Wahrscheinlichkeitsmaße als Eingabeformate in verschiedensten Situationen auftreten.

Wissenschaftlicher Werdegang des Verfassers

10/1996–02/2002	Studium der Physik mit Nebenfach Mathematik an der Universität Tübingen.
02/2002	Diplom in Physik Thema der Diplomarbeit: Numerische Simulation axialsymmetrischer, isolierter Systeme in der Allgemeinen Relativitätstheorie. Betreuer: PD. Dr. J. Frauendiener
06/2002–11/2005	Wissenschaftlicher Mitarbeiter am Max-Planck-Institut für biologische Kybernetik in Tübingen in der Abteilung von Prof. Dr. Bernhard Schölkopf.

Erklärung

Hiermit erkläre ich, daß ich die vorliegende Arbeit - mit Ausnahme der in ihr ausdrücklich genannten Hilfen - selbständig verfasst habe.

Acknowledgements

First of all I would like to thank Bernhard Schölkopf for giving me the possibility to do my doctoral thesis in an excellent research environment. He gave me the freedom to look for my own lines of research while always providing ideas how to progress. I also very much appreciated his advice and support in times when it was needed.

I am especially thankful to Olivier Bousquet for guiding me into the world of learning theory. In our long discussions we usually grazed through all sorts of topics ranging from pure mathematics to machine learning to theoretical physics. This was very inspiring and raised my interest in several branches of mathematics. He always had time for questions and was a constant source of ideas for me.

I want to thank Thomas Hofmann for giving me the opportunity to do my thesis at the TU Darmstadt. I am very thankful for his support in these last steps towards the thesis.

A special thanks goes to Olaf Wittich for reading parts of the second chapter and for giving helpful comments which improved the clarity of this part.

During these three years I had the pleasure to work or discuss with several other nice people. They all influenced in the way I think about learning theory. I thank all of them for their time and help: Jean-Yves Audibert, Goekhan Bakır, Stephane Boucheron, Olivier Chapelle, Jan Eichhorn, André Elisseeff, Matthias Franz, Arthur Gretton, Jeremy Hill, Kwang-In Kim, Malte Kuss, Matti Kääriäinen, Navin Lal, Cheng Soon Ong, Petra Philips, Carl Rasmussen, Gunnar Rätsch, Lorenzo Rosasco, Alexander Smola, Koji Tsuda, Ulrike von Luxburg, Felix Wichmann, Olaf Wittich, Dengyong Zhou, Alexander Zien, Laurent Zwald.

I would like to thank all the AGBS team and in particular all the PhD students in our lab for a very nice atmosphere and a lot of fun. In particular I would like to thank our pioneer Ulrike von Luxburg for pleasant and helpful discussions and for the mutual support of our small 'theory' group, Navin Lal for a nice time here in Tübingen, Malte Kuss for providing me his Matlab script to produce the nice manifold figures, my office mate Arthur Gretton for his subtle jokes and the nice atmosphere and all AOE participants for relaxing afterhours in our lab.

Finally I would like to thank my family for their unconditional help and support during my studies and to Kathrin for her understanding and for reminding me sometimes that there is more in life than a thesis.

Inhaltsverzeichnis

Kapitel 1

Introduction

Geometry plays an important role in modern statistical learning theory, and many different aspects of geometry can be found in this fast developing field. This thesis addresses some of these aspects. A large part of this work will be concerned with so called manifold methods, which have recently attracted a lot of interest. The key point is that for a lot of real-world data sets it is natural to assume that the data lies on a low-dimensional submanifold of a potentially high-dimensional Euclidean space. We develop a rigorous and quite general framework for the estimation and approximation of some geometric structures and other quantities of this submanifold, using certain corresponding structures on neighborhood graphs built from random samples of that submanifold. Another part of this thesis deals with the generalization of the maximal margin principle to arbitrary metric spaces. This generalization follows quite naturally by changing the viewpoint on the well-known support vector machines (SVM). It can be shown that the SVM can be seen as an algorithm which applies the maximum margin principle to a subclass of metric spaces. The motivation to consider the generalization to arbitrary metric spaces arose by the observation that in practice the condition for the applicability of the SVM is rather difficult to check for a given metric. Nevertheless one would like to apply the successful maximum margin principle even in cases where the SVM cannot be applied. The last part deals with the specific construction of so called Hilbertian metrics and positive definite kernels on probability measures. We consider several ways of building such metrics and kernels. The emphasis lies on the incorporation of different desired properties of the metric and kernel. Such metrics and kernels have a wide applicability in so called kernel methods since probability measures occur as inputs in various situations.
As a foundation for the following chapters we first introduce basic notions of statistical learning theory. Then we give a more detailed account of geometry in statistical learning theory. We conclude this chapter with a summary of our contributions to different aspects of this topic.

1.1 Introduction to statistical learning theory

Statistical learning theory is the mathematical theory of learning. Its foundations were mainly laid by Vapnik and Chervonenkis. For a historical account see the book of Vapnik [99]. Statistical learning theory puts the learning process into a mathematical framework. The following should not be understood as a complete

introduction to this field. It mainly serves as an introduction to the basic notions and questions answered by the theory. In particular we will only study classification in detail, where we mainly focus on the introduction of the so called Rademacher averages as a capacity measure of a function class. The Rademacher averages will then be used in Chapter 4. We mainly follow the two review articles of Bousquet, Boucheron and Lugosi [16, 17].
The data is assumed to be from $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the 'input' space and $\mathcal{Y}$ is the space of possible labels. For binary classification $\mathcal{Y}$ it is simply $\{-1, 1\}$, whereas $\mathcal{Y} = \mathbb{R}$ for regression. The key assumptions in statistical learning theory are twofold:

- There exists a data-generating probability measure P on $\mathcal{X} \times \mathcal{Y}$,
- The data $(X_i, Y_i)_{i=1,\dots,n}$ is drawn independent[1] and identically distributed[2](i.i.d.) from P.

The main difference with respect to classical statistical inference problems, where one works with a parametric model P_θ, is that in statistical learning theory no assumptions[3] are made on the probability measure P. A learning algorithm is then simply a way to choose a function from a given hypothesis class of functions after one has seen the data.

Definition 1.1 (Learning algorithm) *A learning algorithm $\mathcal{A}$ is a mapping $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}$ which assigns to any sample $(X_i, Y_i)_{i=1,\dots,n}$ a function $f_n : \mathcal{X} \to \mathcal{Y}$ chosen from a given class of functions $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$. In classification the function f_n is called the classifier.*

In order to assess the quality of a learning algorithm one needs a way to measure its performance.

Definition 1.2 (Loss function) *A loss function $l : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ measures how errors in the prediction are penalized. The (expected) loss of a function f is then defined as $L(f) = \mathbb{E}\, l(X, f(X), Y)$ In classification one usually chooses $l(X, f(X), Y) = \mathbb{1}_{f(X)\neq Y}$. Then the loss of a function f is the probability of error $L(f) = \mathbb{E}\, \mathbb{1}_{f(X)\neq Y} = \mathrm{P}(f(X) \neq Y)$, which is also often called the risk of a function.*

It is then straightforward to define the best possible loss. Before we do so, let us introduce the useful concept of the regression function η, defined as $\eta(x) = \mathbb{E}[Y|X = x]$. In the case of binary classification one has $\eta(x) = 2\mathrm{P}(Y = 1|X = x) - 1$. For a deterministic rule the regression function attains only the values $\{-1, 1\}$, whereas if the rule is completely random, that is $\mathrm{P}(Y = 1|X = x) = \mathrm{P}(Y = -1|X = x) = \frac{1}{2}$, then $\eta(x) \equiv 0$. For simplicity we will restrict ourselves in the following to the case of binary classification with loss function $l(x, f(x), y) = \mathbb{1}_{f(x)\neq y}$.

Definition 1.3 (Bayes classifier, Bayes loss) *The minimal achievable loss is called the Bayes loss (Bayes risk) and is defined as*

$$L^* = \inf\{L(f) \,|\, f \text{ measurable}\}$$

A function f^* which attains L^* is called the Bayes classifier and one has

$$f^*(x) = \operatorname{sgn} \eta(x).$$

[1] More general settings allow also dependencies in the data.
[2] In the following big letters X will denote random variables and small letters x a point in $\mathcal{X}$. If not indicated otherwise expectations $\mathbb{E}$ are taken of all random quantities.
[3] Tsybakov has introduced recently so called noise conditions on P, however they are completely different from assumptions made in classical statistics.

For a given function class $\mathcal{F}$ the minimal loss in this class $L^*_{\mathcal{F}}$ is defined[4] as

$$L^*_{\mathcal{F}} = \inf_{f \in \mathcal{F}} L(f)$$

A question that immediately arises is how the loss of the function f_n chosen by the algorithm differs from the loss of the Bayes classifier, i.e. the best possible classifier. Indeed one can write the difference of the loss $L(f_n)$ of the function $f_n \in \mathcal{F}$ to the Bayes loss as:

$$L(f_n) - L^* = (L(f_n) - L^*_{\mathcal{F}}) \; + \; (L^*_{\mathcal{F}} - L^*)$$

The first term on the right hand side is called the estimation error. It is a random quantity since f_n depends on the sample and measures how close the chosen function is to the best possible function in $\mathcal{F}$. The second term is called the approximation error. It measures the difference between the loss of the best possible function in $\mathcal{F}$ and the Bayes classifier. The richer the function class $\mathcal{F}$, the lower the approximation error.
The core problem now is that the probability measure P is in general unknown. Therefore the loss $L(f) = \mathbb{E}\, l(X, f(X), Y)$ of a function f cannot be computed and so the performance of a learning algorithm is not directly accessible. Given the sample one can instead estimate the empirical loss (risk)

$$L_n(f) = \frac{1}{n} \sum_{i=1}^{n} l(X_i, f(X_i), Y_i) = \mathbb{E}_{P_n}\, l(X, f(X), Y),$$

where $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$ is the empirical measure of P.
We have now defined all the basic notions in order to formulate the questions which statistical learning theory tries to answer:

- How close is the empirical loss of the chosen classifier to its true loss ?
- Does the difference between true and empirical loss of the classifier converge to zero as the sample size goes to infinity ?
- How fast does this convergence happen ?
- Does the approximation error approach[5] zero as $n \to \infty$?

We will now study some of these questions for the most simple type of learning algorithm: the empirical risk minimization. Afterwards we briefly discuss regularized empirical risk minimization, which is the basic principle behind many modern learning algorithms.

1.1.1 Empirical risk minimization

Empirical risk minimization (ERM) is a very simple way to choose the classifier. One simply takes the empirical risk minimizer from a predefined function class $\mathcal{F}$. That is, ERM is defined as

$$ERM: \quad f_n = \inf_{f \in \mathcal{F}} L_n(f)$$

[4]For technical simplicity we assume here that the minimal loss $L^*_{\mathcal{F}}$ can be achieved by some f in $\mathcal{F}$.
[5]This question only makes sense if one allows that the function class used in the learning algorithm grows with the sample size.

Whereas it is simple to write down the principle of ERM, algorithmically ERM is usually very complex and leads to NP-hard problems, even for simple function classes $\mathcal{F}$. Moreover it suffers in general from overfitting and instability, that is small changes in the data can lead to large differences in the final classifier.
Nevertheless it is a valid algorithm, and as we will see for appropriately chosen function classes it is consistent in the sense that as $n \to \infty$ ERM finds a classifier with the best risk attainable using classifiers in $\mathcal{F}$. Moreover the rather simple setting allows one to introduce rather straightforwardly the basic methods used in statistical learning theory.
In the following we will be interested in getting upper bounds for the probability of uniform deviation of the empirical loss from the true loss. We have the following trivial inequality:

$$L(f_n) \leq L_n(f_n) + \sup_{f \in \mathcal{F}} \big(L(f) - L_n(f)\big).$$

The quantity $\sup_{f \in \mathcal{F}} \big(L(f) - L_n(f)\big)$ is random. In a first step towards an upper bound we will show that it is close to its expectation using a concentration inequality; more precisely, we employ the so called bounded differences inequality of McDiarmid.

Theorem 1.4 (see [63]) *Let $g : \mathcal{X}^n \to \mathbb{R}$ be a function such that for some nonnegative $c_1, \ldots, c_n$,*

$$\sup_{x_1,\ldots,x_n,x_i' \in \mathcal{X}} |g(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i, \quad 1 \leq i \leq n,$$

and define $C = \sum_{i=1}^n c_i^2$. Furthermore, let $X_1, \ldots, X_n$ be independent random variables. Then the random variable $U = g(X_1, \ldots, X_n)$ satisfies:

$$\mathrm{P}(|U - \mathbb{E}\, U| > t) \leq 2e^{-2t^2/C}.$$

Now denote by $L_n^i(f)$ the loss of f where the i-th variable is modified. Using $l(X_i, f(X_i), Y_i) = \mathbb{1}_{f(X_i) \neq Y_i}$, it follows that

$$\Big| \sup_{f \in \mathcal{F}} \big(L(f) - L_n(f)\big) - \sup_{f \in \mathcal{F}} \big(L(f) - L_n^i(f)\big) \Big| \leq \sup_{f \in \mathcal{F}} |L_n^i(f) - L_n(f)| \leq \frac{1}{n}.$$

This means the random variable $\sup_{f \in \mathcal{F}} \big(L(f) - L_n(f)\big)$ satisfies the bounded differences inequality with $c_i = \frac{1}{n}$ and therefore with probability $1 - \delta$,

$$\sup_{f \in \mathcal{F}} \big(L(f) - L_n(f)\big) \leq \mathbb{E} \sup_{f \in \mathcal{F}} \big(L(f) - L_n(f)\big) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

One can bound now the expectation of the deviation using the symmetrization technique. Let $(X_1', Y_1'), \ldots, (X_n', Y_n')$ denote a ghost sample, independent of (X_i, Y_i) and distributed identically, and denote by $L_n'(f) = \frac{1}{n} \sum_{i=1}^n l(X_i', f(X_i'), Y_i')$ the empirical loss with respect to the ghost sample. Then with $L(f) = \mathbb{E}\, L_n'(f)$ and Jensen's inequality, we get

$$\mathbb{E} \sup_{f \in \mathcal{F}} \big(L(f) - L_n(f)\big) = \mathbb{E} \sup_{f \in \mathcal{F}} \big(\mathbb{E}\, L_n'(f) - L_n(f)\big) \leq \mathbb{E} \sup_{f \in \mathcal{F}} \big(L_n'(f) - L_n(f)\big).$$

Let us now introduce independent Rademacher variables[6], σ_i, $i = 1, \ldots, n$, in order to rewrite the last expression as:

$$\begin{aligned}\mathbb{E} \sup_{f \in \mathcal{F}} \big(L'_n(f) - L_n(f) \big) &= \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \big[l(X_i, f(X_i), Y_i) - l(X'_i, f(X'_i), Y'_i) \big] \\ &\leq 2\, \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i\, l(X_i, f(X_i), Y_i).\end{aligned}$$

Let Z_i be random variables in $\mathcal{Z}$ and $\mathcal{G}$ a function class with domain $\mathcal{Z}$. Then $\mathbb{E}_\sigma \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(Z_i)$ is called the Rademacher average $\widehat{R}_n(\mathcal{G})$ of $\mathcal{G}$. Note that it is conditioned on the data Z_i. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and denote by $\mathcal{G}$ the function class indexed by $f \in \mathcal{F}$, where

$$g(z) = \mathbb{1}_{f(x) \neq y}.$$

Given $l(x, f(x), y) = \mathbb{1}_{f(x) \neq y}$, one has

$$\widehat{R}_n(\mathcal{G}) = \mathbb{E}_\sigma \sup_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(Z_i) = \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i\, l(X_i, f(X_i), Y_i),$$

since there is a one-to-one relationship between the set $\mathcal{G}$ and $\mathcal{F}$.
Then using the derived upper bound, one gets with probability $1 - \delta$,

$$\sup_{f \in \mathcal{F}} \big(L(f) - L_n(f) \big) \leq 2\, \mathbb{E}\, \widehat{R}_n(\mathcal{G}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

One can also check that $\widehat{R}_n(\mathcal{G})$ satisfies the bounded differences inequality with $c_i = \frac{1}{n}$. Then one gets

$$\sup_{f \in \mathcal{F}} \big(L(f) - L_n(f) \big) \leq 2\, \widehat{R}_n(\mathcal{G}) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

Note that this is a data-dependent performance bound, which can be expected to be tighter than the more classical distribution independent bounds. The Rademacher average $\widehat{R}_n(\mathcal{G})$ can be seen as a capacity measure of the function class $\mathcal{G}$ respectively $\mathcal{F}$. An intuitive interpretation of the Rademacher average $\widehat{R}_n(\mathcal{G})$ is that it measures the ability of $\mathcal{G}$ to fit random noise. In fact if $\mathcal{G}$ is very large then there will always exist a function in $\mathcal{G}$ which fits the σ_i, that is $g(Z_i) = 1$ if $\sigma_i = 1$ and $g(Z_i) = 0$ otherwise, and therefore $\widehat{R}_n(\mathcal{G}) = \frac{1}{2}$. Then there is no hope of uniform convergence of the difference between empirical and true risk to zero. Furthermore it turns out that one can upper bound the Rademacher average in terms of more classical capacity terms like the distribution-independent shattering coefficients or the VC-dimension, see [17] for more on this topic.
As a final remark let us note that bounding the probability of the deviation

$$\delta_n(\mathcal{F}) := \sup_{f \in \mathcal{F}} \big| \mathbb{E}_{P_n}\, L(f) - \mathbb{E}_P\, L(f) \big|$$

is known as the Glivenko-Cantelli problem. In particular a function class $\mathcal{F}$ where

$$\lim_{n \to \infty} \delta_n(\mathcal{F}) = 0 \quad \text{almost surely}$$

[6] A Rademacher variable σ is a random variable with $\mathrm{P}(\sigma = 1) = \mathrm{P}(\sigma = -1) = \frac{1}{2}$.

is known as Glivenko-Cantelli class[7]. A uniform Glivenko-Cantelli class is a class of functions $\mathcal{F}$ where this convergence takes even place uniformly over all probability measures. The uniform Glivenko-Cantelli classes of binary-valued functions have been characterized as follows:

Theorem 1.5 (Vapnik and Chervonenkis, see [65]) *A class of binary-valued functions is a uniform Glivenko-Cantelli class if and only if it has finite VC-dimension.*

1.1.2 Regularized empirical risk minimization

There are several problems using empirical risk minimization as a learning algorithm which we summarize now:

- in particular for small sample sizes, ERM leads in general to overfitting: that is the data and the noise is fitted, which leads to poor generalization,
- ERM is unstable: small changes in the data can lead to large differences in the classifier,
- since the function class $\mathcal{F}$ is fixed, the approximation error is also fixed and is in general nonzero. However that means that even as the sample size goes to infinity one cannot expect that one obtains a classifier with the smallest possible loss, the Bayes loss.

All these three points motivate algorithms based on regularized empirical risk minimization, which can be formulated for a given function class $\mathcal{F}$ as follows

$$f_n = \min_{f \in \mathcal{F}} L_n(f) + \lambda\, \Omega(f)$$

where $\Omega : \mathcal{F} \to \mathbb{R}_+$ is the so called regularization functional and $\lambda > 0$ a parameter controlling the trade-off between empirical risk and regularization. Note that even for binary classification one uses in general as the output space $\mathcal{Y}$ either $[-1, 1]$ or even the whole real line. The final classifier is then determined by the sign of the function. Let us now discuss rather informally how the introduction of the regularizer $\Omega(f)$ addresses the three disadvantages of ERM:

- Usually $\Omega(f)$ is a measure of smoothness of the function f, which in turn leads to a trade-off between fit of the data and smoothness in the choice of the classifier f_n. The idea is that smooth functions generalize better, having small true loss. In other words the preference for smooth functions reflects the implicit assumption that closeness in $\mathcal{X}$ implies in general closeness of the labels in $\mathcal{Y}$. Note however that this is an assumption which need not be a priori valid. We only think of it as a rather natural assumption since it is fortunately fulfilled in most real-world data sets.
- The ERM principle is unstable: in other words, the problem of learning based on ERM is ill-posed. Then it is a standard approach in the theory of regularization to add a regularization term to transform the problem into a well-posed one.

[7] Equivalently: for a Glivenko-Cantelli class the strong law of large numbers holds uniformly over the function class $\mathcal{F}$

- The introduction of the regularizer leads to an effective reduction of the function class $\mathcal{F}$, since in general only functions with small penalization term $\Omega(f)$ will be chosen as the classifier. This in turn allows one to choose a rather large $\mathcal{F}$, so that the approximation error is small or even zero.

A large part of this thesis is concerned with algorithms known as kernel methods. Many of them, the support vector machine (SVM) in particular, can be formulated as regularized empirical risk minimization. The function class $\mathcal{F}$ is in this case a reproducing kernel Hilbert space $\mathcal{H}_k$ (RKHS) associated to a positive definite kernel k, and the regularizer $\Omega(f)$ is taken to be the squared norm in that RKHS: $\Omega(f) = \|f\|^2_{\mathcal{H}_k}$. We will introduce in Chapter 3 positive definite kernels and associated structures and their properties. For a detailed account of kernel methods we refer to the book of Schölkopf and Smola, see [81]. The consistency of the regularized empirical risk minimization for kernel methods, and in particular the universal consistency of SVM's, has been studied by Steinwart in [89, 90].

1.2 Geometry in statistical learning theory

Until now geometrical aspects in statistical learning theory have not been stressed. In this section we try to emphasize several explicit but also some more implicit occurrences of geometry in statistical learning.

- 'Input spaces $\mathcal{X}$ as metric spaces': Traditionally learning algorithms are based on a feature representation of the data in $\mathbb{R}^d$. However in the learning algorithm itself a metric[8] is used for the data either implicitly or explicitly . Therefore in our opinion it is in most cases equivalent to start initially with $(\mathcal{X}, d)$, that is a set $\mathcal{X}$ with metric d. The clear advantage of such a setting is that one can deal with much more general structures than $\mathbb{R}^d$. However a disadvantage is that many algorithms use the linear structure of $\mathbb{R}^d$ which is in general not available in a metric space. Therefore some algorithms cannot be directly transferred to this setting. In Chapter 4 we will discuss how one can transfer the maximal margin principle to arbitrary metric spaces.

- 'Smoothness regularizer': Smoothness of a function means that closeness in the inputs in $\mathcal{X}$ implies closeness in the outputs in $\mathcal{Y}$. Now what closeness means in either $\mathcal{X}$ or $\mathcal{Y}$ is usually determined by a metric in these spaces, so that the underlying metric structure defines which functions we consider to be smooth. As a more explicit example take the Lipschitz constant $\mathrm{Lip}(f)$ of a function f, defined as

$$\mathrm{Lip}(f) = \sup_{x,y \in \mathcal{X}} \frac{|f(x) - f(y)|}{d(x,y)}$$

  as smoothness regularizer on the metric space $\mathcal{X}$. Here it is very obvious that the underlying metric structure of $\mathcal{X}$ determines what we consider as smooth functions. In Chapter 4 we will show that the Rademacher averages for the maximal margin classifier for arbitrary metric spaces can be upper bounded in terms of the covering numbers of the metric space $(\mathcal{X}, d)$; that is, the geometry of $\mathcal{X}$ directly determines the capacity of the classifier.

[8] Note here that one can always express an inner product in terms of distances.

- 'Regularizers which are adaptive to submanifold structure in the data': If one has a large number of features, that is the data lives in a high-dimensional space $\mathbb{R}^d$, it is often observed that the features are not independent of each other. This in turn can lead to the effect that the data lies either on or close to a low-dimensional submanifold in $\mathbb{R}^d$. It is then much more natural to penalize functions with respect to the geometry of this submanifold. However since the submanifold structure is a priori unknown, one cannot directly define a regularizer which is adapted to the geometry of the submanifold. Instead one can construct regularizers which are only defined on the data and implicitly approximate a geometric regularizer of the submanifold. In that sense the regularizer adapts to the underlying structure as one gets more samples. In Chapter 2 we will consider and identify the limit of smoothness functionals defined on an approximating neighborhood graph built from random samples as the number of samples goes to infinity.

- 'Metrics and kernels specifically designed for a given structure': The metric resp. the kernel encode the dissimilarity and similarity in a space. In turn this notion of similarity should encode our prior knowledge about the problem. As an example take a metric between color histograms of images. If we only want to compare the mass assigned to different colors and we have no preference among the color spaces employed the choice of the color space should not influence the metric. More formally the metric should be covariant, meaning it should be invariant with respect to coordinate transformations. In Chapter 5 we develop metrics and kernels of this kind and also kernels which capture other properties of histograms resp. probability measures.

The study of such geometrical properties is not at the core of traditional statistical learning theory, since usually only feature representations in $\mathbb{R}^d$ are considered. However in recent years the need to deal with more complex data structures in all sorts of fields like computer vision, bioinformatics or information retrieval has shown that such questions should also attract more interest in theory. We think that in particular the study of adaptive regularization and the specific design of metrics and kernels will become even more important in the future.

1.3 Summary of Contributions of this thesis

In the following we would like to summarize the main achievements in this thesis.

Chapter II This chapter considers the continuum limit of structures of certain neighborhood graphs built from random samples in $\mathbb{R}^d$. Of particular interest is the case when the data generating probability measure has support on a low-dimensional submanifold. One can roughly see the discrete (random) graph structure as an approximation of the corresponding continuous quantities. Our setting will be quite general:

- The submanifold M is allowed to have a boundary and is in general non-compact. The curvature of M is assumed to be bounded and there are some more technical conditions,

- The data generating probability measure P is only required to have a density with respect to the natural volume element of the submanifold M, which has a certain smoothness.

All these requirements should not exclude any interesting scenario occurring in practice.
We consider in particular the limit of the following structures on the graph:

- The degree function: We show that in general the degree function of the graph converges pointwise towards a function of the density. In the simplest case where one has no data-dependent edge weights it turns out that the extended degree function is nothing else than a kernel density estimator on the submanifold (however using the Euclidean (extrinsic) distance instead of the intrinsic distance).
- The normalized as well as the unnormalized graph Laplacian: We show that the normalized graph Laplacian converges pointwise towards the weighted Laplace-Beltrami operator of the submanifold, whereas the unnormalized graph Laplacian converges only up to a function of the density towards the weighted Laplace-Beltrami operator. It is quite interesting that the influence of the probability measure can be controlled by the way one defines the generally data-dependent edge-weights. One can even eliminate the influence of the probability measure completely so that one gets the standard Laplace-Beltrami operator.
- The smoothness functional: The normalized as well as the unnormalized graph Laplacian induce a smoothness functional for functions on the vertices of the graph. Both smoothness functionals coincide. We then establish that their common limit corresponds to a penalization of the integrated squared norm of the gradient of the function along the submanifold weighted by some function of the density. Also for the smoothness functional one can control the influence of the probability measure by choosing different data-dependent edge weights.

Finally as an application of the framework we develop for the estimation of submanifold structure, we propose a new scheme for intrinsic dimensionality estimation of submanifolds in $\mathbb{R}^d$.

Chapter III The goal of this chapter is to survey some of the results relevant for machine learning on positive definite kernels and associated structures scattered in the mathematical literature. In particular we briefly discuss the relationship to Gaussian processes, positive kernel operators and reproducing kernel Hilbert spaces (RKHS). Moreover we answer some interesting questions about RKHS like: When is a RKHS separable ? When does it contain only continuous functions ? What are feature maps ?
We present some new results on the integral operator and covariance operator associated to a class of kernels. We then discuss three ways of generalizing positive definite kernels resp. their associated RKHS. It turns out that two ways, which are positive definite operator-valued kernel functions and the so called framework of Hilbertian subspaces mainly developed by Schwartz, can actually be seen as special cases of the standard framework. The only true generalization is that of positive definite kernels to kernels with negative squares. This type of kernel still induces a reproducing kernel space but with an inner product which is indefinite.

Chapter IV In this chapter we develop a general framework for applying the maximum margin principle to arbitrary metric spaces. The motivation for doing this arises by moving to a new point of view on support vector machines (SVM), in particular on the use of kernels in the SVM. The usual argument is that the kernel is used to map the input space into the RKHS and then maximum margin separation is done there. However one can also use the interpretation that the kernel defines a (semi)-metric on the input space and then this (semi)-metric space is mapped isometrically into the RKHS. However not all metric spaces can be embedded isometrically into a Hilbert space, and thus the SVM can not be applied to all metric spaces. However every metric space can be embedded isometrically into a Banach space. We hence formulate the general framework of maximum margin separation in this Banach space. Unfortunately the problem in a Banach space lacks some of the appealing properties of the maximal margin problem when it is formulated in a Hilbert space. In particular there is no representer theorem. However an approximative solution can be developed which is even exact if one considers the data only as a finite metric space. We show further that for the SVM only the induced metric matters and not the kernel. Indeed the optimization problem as well as the solution of the SVM can be equivalently formulated with a metric.

Chapter V The last chapter deals with the construction of Hilbertian metrics[9] and positive definite kernels on probability measures. At first we consider γ-homogeneous Hilbertian metrics on $\mathbb{R}_+$, which were recently characterized by Fuglede. We extend the parameter range of a two-parameter family of 1/2-homogeneous Hilbertian metrics on $\mathbb{R}_+$ introduced by Fuglede and Topsøe. Then we present using 1/2-homogeneous Hilbertian metrics resp. one-homogeneous positive definite kernels on $\mathbb{R}_+$ the general principle of building Hilbertian metrics and positive definite kernels on probability measures which are covariant. Covariance means that the metric resp. the kernel is invariant under coordinate transformations. Applying this principle to the two-parameter family of Hilbertian metrics on $\mathbb{R}_+$ we get several well-known measures as special cases of an effectively one parameter family of Hilbertian metrics on probability measures (either directly or the square roots of them): The χ^2-measure, the Hellinger distance, the Jensen-Shannon divergence and the total variation. These covariant metrics have the problem that disjoint measures are at maximum distance, so that learning with disjoint measures is not possible with these kind of metrics. Therefore we consider two ways of building so called structural kernels which incorporate a similarity function into the kernel, so that disjoint measures have different distances according to the 'similarity' of their support. We compare all these new kernels in four experiments using SVM's which show that these new kernels perform often better or at least comparable to standard kernels on histogram data.

[9] A Hilbertian metric is a metric which can be isometrically embedded into a Hilbert space.

Kapitel 2

Consistent Continuum Limit for Graph Structure on Point Clouds

In recent years the interest in graph based methods in machine learning has increased rapidly. In particular in semi-supervised resp. transductive learning [110, 107, 18, 9], dimensionality reduction [8, 25] and clustering (see [102] for references) graph based methods have been very successful. There are several reasons for this. First graph based methods can be used with all kind of data. The only requirement is that one has a function of two variables which assigns to every pair of points its similarity respectively dissimilarity, second, there is a large amount of theory in mathematics and theoretical computer science on which one can build on, and third, and that is the topic of this chapter, graphs approximate an underlying continuous structure if they are built in a certain way.
Naturally graphs are inherently discrete objects. However if one has as an underlying continuous structure certain neighborhood graphs and the associated geometric operators can be seen as approximations of the underlying continuous structure respectively of the corresponding continuous operators. The main goal of this chapter is to show how certain structures on neighborhood graphs built from random samples can be defined such that their continuum limit models a desired continuous quantity. In particular we are interested in the limit of the graph Laplacian, its associated smoothness functional, and the graph structure from which it is derived, the Hilbert spaces $\mathcal{H}_V$ and $\mathcal{H}_E$ of functions on the vertices V and the edges E and the derivative operator d.
These limits are well-understood for lattices. For example the second-order approximation of the Laplacian on an equidistant lattice can be seen as the graph Laplacian of a certain nearest neighbor graph on the lattice, see [85]. An already much more complicated setting is when one considers graphs generated by a discretization of a Riemannian manifold[1]. For this type of regular graph several interesting connections have been shown between properties of the graph and the corresponding properties of the approximated Riemannian manifold, see [100] and in particular the chapter 4.4 in the book of Chavel [22]. Although this regular setting has been studied quite intensively in the last twenty years, the bridge to the random setting usually encountered in machine learning applications has only partially been addressed yet. This is somehow surprising since often algorithms in machine learning are motivated

[1] Chavel defines a discretization of a complete Riemannian manifold as an ϵ-separated subset S of M which is an R-covering of M. The neighborhood $N(x)$ of $x \in S$ is then defined as $N(x) := \{S \cap B(x, 2R)\} \backslash \{x\}$.

by properties of the continuous Laplacian.
The Figure 2.1 illustrates the setting we are working in. In general we assume that

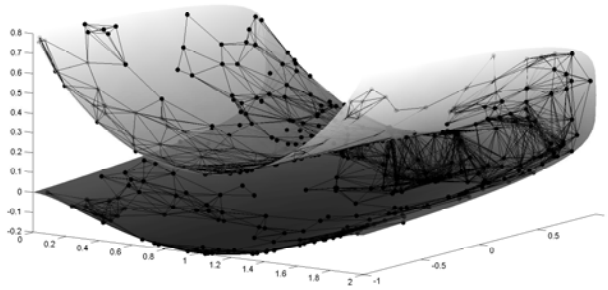


Abbildung 2.1: Random samples X_i of a probability measure of a two-dimensional submanifold in $\mathbb{R}^3$ and the associated neighborhood graph.

we are given random samples X_i from a probability measure P which has support on a submanifold[2] M in $\mathbb{R}^d$. This assumption is motivated by the observation that data in $\mathbb{R}^d$ where d is rather large has often only a small number m of intrinsic parameters, e.g. image sequences of a smoothly varying object, so that it is reasonable to assume that effectively the data lies on a low-dimensional submanifold. We then consider these random samples X_i as vertices of a graph. An edge between two vertices X_i and X_j is defined if they are close with respect to the Euclidean distance in $\mathbb{R}^d$, resulting in a neighborhood graph. This neighborhood graph can be seen as an approximation of the submanifold M. The random graph setting is considerably more difficult than the regular setting emerging from the discretization of a Riemannian manifold. In the latter one has much more control on the distances of points of the discretization and the number of points of the discretization in a certain neighborhood which one has not in the random setting. Additionally we have only access to the Euclidean distance in $\mathbb{R}^d$ between two sample points but not to the intrinsic distance in M. However neighborhoods in the Euclidean distance do in general not correspond to the corresponding neighborhoods of the manifold M measured in the intrinsic distance $d_M(x, y)$ since one has always $d_M(x, y) \geq \|x - y\|_{\mathbb{R}^d}$.
The goal of this chapter is to compute the limits of the following structures on a neighborhood graph as the sample size n goes to infinity and the neighborhood size h shrinks to zero:

- The degree function $d_{h,n}$ corresponding to fixed weights,
- The general degree function $d_{\lambda,h,n}$ corresponding to data-dependent weights,
- The normalized graph Laplacian $\Delta_{\lambda,h,n}$,
- The unnormalized graph Laplacian $\Delta'_{\lambda,h,n}$,
- The inner product $\langle f, g\rangle_{\mathcal{H}(V,d_{\lambda,h,n})}$ of functions f, g on the vertices V,
- The smoothness functional $S(f) = \langle f, \Delta f\rangle$ induced by the normalized and the unnormalized graph Laplacian.

[2]However it includes as a special case the setting where P has full support on $\mathbb{R}^d$.

Particular emphasis is laid on the control of the influence of the density p of P which can be done with the parameter λ. One could ask why this is interesting for machine learning. The answer is that, as we will point out later, there is a lot of freedom to build structure on the graph and it is not clear a priori which kind of structure on the graph is suited for phrasing the problem at hand in an optimal way. From a regularization point of view it might be interesting to look for functions which are smooth in high density regions on the submanifold or for a diffusion process (label propagation in semi-supervised learning) which is mainly directed towards high-density regions on the submanifold M. However it is not known which kind of smoothness functional respectively graph Laplacian models this. The main goal of this chapter is to provide a toolbox for building algorithms on graphs by reverse engineering. By reverse engineering we mean that one first defines the objectives of the algorithm in the continuous case where one has usually more intuition than in the discrete case. Then in a second 'reverse' step one chooses on the graph the structures which will approximate in the large sample limit (as the neighborhood shrinks to zero) the chosen continuous structure by the dictionary of limits we provide in this chapter.
Another goal of this chapter is to fix a certain ambiguity on the graph. Namely in order to define the graph Laplacian one has to introduce three structures on the graph. The Hilbert spaces $\mathcal{H}_V, \mathcal{H}_E$ of functions on the vertices V resp. edges E and the difference operator d. Then the graph Laplacian Δ is defined as $\Delta = d^*d$. As we will show, the choice of a specific graph Laplacian does not fix these structures and even more surprising the induced smoothness functional $S(f) = \langle f, \Delta f \rangle_{\mathcal{H}_V}$ is independent of the choice of $\mathcal{H}_V$. Nevertheless by requiring mutual consistency in the continuum limit one can at least fix the Hilbert space $\mathcal{H}_V$.
The proofs usually work in the following way. First one establishes the convergence of the discrete graph structure to a continuous counterpart ("variance term") as the sample size $n \to \infty$, and in a second step the convergence of this continuous operator to the desired continuous operator ("bias term") is shown as the neighborhood size $h \to 0$. Merging both limit processes, one considers both limits simultaneously.
The second step has already been studied by Belkin [11] for Gaussian weights for the unnormalized graph Laplacian in the case of compact submanifolds without boundary in $\mathbb{R}^d$ and the uniform measure (as a byproduct the same result has also been derived earlier in the work of Smolyanov, Weizsäcker and Wittich in [86]), and was then generalized by Lafon [58] to compact submanifolds with boundary, general isotropic weights, and general densities. We will also use the data-dependent weights introduced by Coifman and Lafon in [25], but further generalize their setting to non-compact manifolds with boundary of bounded geometry. Belkin and Lafon show that the bias term converges pointwise for $h \to 0$, where h controls the neighborhood of the graph. However, the convergence of the variance term was left open in [11] and [58].
The first work where, in a slightly different setting, both limit processes have been studied together is the work of Bousquet, Chapelle and Hein [18]. Using the law of large numbers for U-statistics, the authors studied the convergence of the regularizer $\Omega_n(f) = \langle f, L_n f \rangle$ for sample size $n \to \infty$ (where $f \in \mathbb{R}^V$ and L_n is the unnormalized graph Laplacian on n sample points). Then taking the limit for the bandwidth $h \to 0$, they arrived at a weighted Laplace operator in $\mathbb{R}^d$. The drawback of this approach is that the limits in n and h are not taken simultaneously.

In contrast to this work von Luxburg, Belkin and Bousquet considered in [102] the setting where the bandwidth h was kept fixed while the large sample limit $n \to \infty$ of the graph Laplacian (normalized and unnormalized) was considered. In this setting, the authors show strong convergence results of graph Laplacians to certain limit integral operators which then even imply the convergence of the eigenvalues and eigenfunctions of the graph Laplacian.
The goal of this chapter is to surpass the limitations of previous approaches. We study the convergence of both bias and variance term, where the limits $n \to \infty$ and $h \to 0$ are taken simultaneously. Moreover, we study in general non-compact submanifolds with boundary of bounded geometry. We think that this setting probably includes all interesting cases occurring in machine learning. It is actually not easy to imagine a submanifold which is not in this class. Also the assumptions on the kernel function used to define the weights for the graph as well as the assumptions on the density $p(x)$ of the data-generating measure P are rather weak and should not exclude interesting cases which occur in practice.
Recently Belkin and Niyogi have independently shown in [10] the pointwise convergence of the unnormalized graph Laplacian for a compact submanifold M without boundary and the uniform measure on M using Gaussian weights. This corresponds to a special case of our Theorem 2.37 for non-data-dependent weights.
This chapter is organized as follows. In Section 2.1 we introduce first directed graphs and the structures defined on them: $\mathcal{H}_V, \mathcal{H}_E$ the Hilbert spaces of functions on the vertices V resp. edges E and the difference operator d. Then we specialize the construction to undirected graphs which we will use throughout the chapter. Finally we introduce the smoothness functional $S(f)$ on the graph and its higher order variations and discuss how to build general regularizing functionals on the graph. Then in Section 2.2 we introduce some basic notions from differential geometry which we use repeatedly in the rest of the chapter: submanifolds, normal coordinates, and, a rather non-standard topic in Riemannian geometry, the so called manifolds with boundary of bounded geometry. Then we discuss how the intrinsic geometry of M is connected to the extrinsic geometry of $\mathbb{R}^d$. This is especially important in the following since we can only compute the Euclidean distances between two points but not the intrinsic geodesic distance on M. Finally we introduce the weighted Laplacian on a Riemannian manifold with measure P and the associated smoothness functional on M. In Section 2.3 we introduce the neighborhood graphs with generally data-dependent weights and state our assumptions on the submanifold M, the kernel function k, and the probability measure P. Then after some preliminary work on the asymptotics of convolutions with the Euclidean distance on M we prove our first result: the limit of the normal degree function. It turns out that one can see the degree function as a kernel density estimator on M. This result can be seen as a generalization of the recent work of Pelletier [72] on kernel density estimation for compact Riemannian manifolds without boundary. However, apart from that, we work in the more general setting of manifolds with bounded geometry the main difference is that in [72] it is assumed that one knows the intrinsic distance function $d_M(x, y)$ on M. We cannot make such an assumption since we do not know the submanifold M beforehand, instead we use the Euclidean distance in the ambient space $\mathbb{R}^d$. As a next result we then derive the limit of the general data-dependent degree function $d_{\lambda,h,n}$. The second main result is the pointwise consistency of both the normalized and the unnormalized graph Laplacian. It turns out that in general only the

normalized graph Laplacian has as its limit a weighted Laplace-Beltrami operator on M, whereas the unnormalized graph Laplacian only converges up to a function of the density to the weighted Laplace-Beltrami operator. Our third main result is the limit of the inner product on the vertices and in particular the limit of the smoothness functional $S(f)$ induced by the graph Laplacian (both the normalized and the unnormalized graph Laplacian induce the same smoothness functional $S(f)$). Finally we fix the freedom in the choice of $\mathcal{H}_V$ for both Laplacians by requiring mutual consistency of the limits of $\mathcal{H}_V$ and the limits of the graph Laplacians.

2.1 Abstract Definition of the Graph Structure

In this section we define the abstract structure on a graph which is required in order to define the graph Laplacian. To this end one has to introduce Hilbert spaces H_V and H_E of functions on the vertices V and the edges E, define a difference operator d, and then set the graph Laplacian as $\Delta = d^*d$. We first do this in full generality for directed graphs and then specialize it to undirected graphs. This approach is well-known for undirected graphs in discrete potential theory, see e.g. [106, 64], and was generalized to directed graphs by Zhou, Schölkopf, and Hofmann in [108] for a special choice of H_V, H_E and d. To our knowledge the very general setting introduced here has not been discussed elsewhere.
In many articles graph Laplacians are used without explicitly mentioning d, H_V and H_E. This can be misleading since, as we will show, there always exists a whole family of choices for d, H_V and H_E which all yield the same graph Laplacian. Since we are interested in finding a consistent continuum limit of the randomly sampled graph, one has to be careful how to define this structure.

2.1.1 Hilbert Spaces of Functions on the vertices V and the edges E

Let (V, W) be a graph where V denotes the set of vertices with $|V| = n$ and W a positive $n \times n$ similarity matrix, that is $w_{ij} \geq 0,\ i, j = 1, \ldots, n$. W need not to be symmetric that means we consider here the case of a directed graph. The special case of an undirected graph will be discussed in a following section. We put a directed edge e_{ij} from the vertex i to the vertex j if $w_{ij} > 0$. Moreover we define the outgoing and ingoing sum of weights of a vertex i as

$$d_i^{out} = \frac{1}{n}\sum_{j=1}^{n} w_{ij}, \quad d_i^{in} = \frac{1}{n}\sum_{j=1}^{n} w_{ji}. \tag{2.1}$$

We assume that $d_i^{out} + d_i^{in} > 0,\ i = 1, \ldots, n$, meaning that each vertex has at least one in- or outgoing edge.
The inner product on the function space $\mathbb{R}^V$ is defined as

$$\langle f, g\rangle_V = \frac{1}{n}\sum_{i=1}^{n} f_i\, g_i\, \chi_i,$$

where $\chi_i = (\chi_{out}(d_i^{out}) + \chi_{in}(d_i^{in}))$ with $\chi_{out} : \mathbb{R}_+^* \to \mathbb{R}_+^*$ and $\chi_{in} : \mathbb{R}_+^* \to \mathbb{R}_+^*$ and $\mathbb{R}_+^* = \{x \in \mathbb{R} | x > 0\}$.

We also define an inner product on the space of functions $\mathbb{R}^E$ on the edges:

$$\langle F, G\rangle_E = \frac{1}{2n^2}\sum_{i,j=1}^{n} F_{ij}\, G_{ij}\, \phi(w_{ij}),$$

where $\phi : \mathbb{R}_+^* \to \mathbb{R}_+^*$. Together with the above assumptions this guarantees us that both inner product are well-defined. We denote by $\mathcal{H}(V,\chi) = (\mathbb{R}_V, \langle\cdot,\cdot\rangle_V)$ and $\mathcal{H}(E,\phi) = (\mathbb{R}^E, \langle\cdot,\cdot\rangle_E)$ the corresponding Hilbert spaces. As a last remark let us clarify the roles of $\mathbb{R}^V$ and $\mathbb{R}^E$. The first one is the space of functions on the vertices and can therefore be seen as a normal function space. However elements of $\mathbb{R}^E$ can be interpreted as a flow on the edges so that the function value on a edge e_{ij} corresponds to the massflowing from one vertex i to the vertex j (per unit time).

2.1.2 The difference operator d and its adjoint d^*

Definition 2.1 (Difference operator) *The difference operator $d : \mathcal{H}(V,\chi) \to \mathcal{H}(E,\phi)$ is defined as follows:*

$$\forall\, e_{ij} \in E, \qquad (df)(e_{ij}) = \gamma(w_{ij})\,(f(j) - f(i)), \tag{2.2}$$

where $\gamma : \mathbb{R}_+^* \to \mathbb{R}_+^*$.

Note that d is zero on the constant functions as one would expect it from a derivative. In [107] a different operator d is used:

$$(df)(e_{ij}) = \left(\frac{f(j)}{\sqrt{d(j)}} - \frac{f(i)}{\sqrt{d(i)}}\right), \tag{2.3}$$

which is in general not zero on the constant functions. This in turn leads also to the effect that the associated Laplacian is not zero on the constant functions. For general graphs without any geometric interpretation this is just a matter of choice. However the choice of d matters if one wants a consistent continuum limit of the graph. Since one cannot expect convergence of the graph Laplacian associated to the difference operator d of Equation (2.3) towards a Laplacian. Since each of the graph Laplacians in the sequence is not zero on the constant functions, also the limit operator will share this property unless $\lim_{n\to\infty} d(X_i) = c, \forall i = 1,\dots,n$, where c is a constant. Since this is in general not the case, as we will show in Section 2.3.3, the limit operator cannot be a Laplacian.
Obviously in the finite case d is always a bounded operator. The adjoint operator $d^* : \mathcal{H}(E,\phi) \to \mathcal{H}(V,\chi)$ is defined by

$$\langle df, u\rangle_E = \langle f, d^*u\rangle_V, \quad \forall\, f \in H(V,\chi), \quad u \in \mathcal{H}(E,\phi).$$

Using the indicator function $f(j) = \mathbb{1}_{j=l}$ it is straightforward to derive:

$$\begin{aligned}\frac{1}{n}\chi_l\,(d^*u)(l) = \langle d\mathbb{1}_{m=l}, u\rangle_E &= \frac{1}{2n^2}\sum_{i,j=1}^{n}(d\mathbb{1}_{j=l})_{ij}u_{ij}\phi(w_{ij}) \\ &= \frac{1}{2n^2}\sum_{i=1}^{n}\gamma(w_{il})u_{il}\phi(w_{il}) - \frac{1}{2n^2}\sum_{i=1}^{n}\gamma(w_{li})u_{li}\phi(w_{li})\end{aligned}$$

In total

$$(d^*u)(l) = \frac{1}{2\chi_l}\left(\frac{1}{n}\sum_{i=1}^{n}\gamma(w_{il})\,u_{il}\,\phi(w_{il}) - \frac{1}{n}\sum_{i=1}^{n}\gamma(w_{li})\,u_{li}\,\phi(w_{li})\right). \tag{2.4}$$

The left term of the last equation can be interpreted as the outgoing flow, whereas the right term can be seen as the ingoing flow. The corresponding continuous counterpart of d is the gradient of a function and for d^* it is the divergence of a vector-field, measuring the infinitesimal difference between in- and outgoing flow.

2.1.3 The general graph Laplacian

Definition 2.2 (graph Laplacian for a directed graph) *Given Hilbert spaces $\mathcal{H}(V,\chi)$ and $\mathcal{H}(E,\phi)$ and the difference operator $d : \mathcal{H}(V,\chi) \to \mathcal{H}(E,\phi)$ the graph Laplacian $\Delta : \mathcal{H}(V,\chi) \to \mathcal{H}(V,\chi)$ is defined as*

$$\Delta = d^*d.$$

Explicitly

$$\begin{aligned}(\Delta f)(l) =& \frac{1}{2\chi_l}\Big[f(l)\frac{1}{n}\sum_{i=1}^{n}\big(\gamma(w_{il})^2\phi(w_{il}) + \gamma(w_{li})^2\phi(w_{li})\big)\\ &- \frac{1}{n}\sum_{i=1}^{n} f(i)\,\big(\gamma(w_{il})^2\phi(w_{il}) + \gamma(w_{li})^2\phi(w_{li})\big)\Big].\end{aligned} \tag{2.5}$$

The explicit expression Δ can be easily derived by plugging the expression of d^* and d together:

$$\begin{aligned}(d^*df)(l) =& \frac{1}{2\chi_l}\left(\frac{1}{n}\sum_{i=1}^{n}\gamma(w_{il})^2[f(l)-f(i)]\phi(w_{il}) - \frac{1}{n}\sum_{i=1}^{n}\gamma(w_{li})^2[f(i)-f(l)]\phi(w_{li})\right)\\ =& \frac{1}{2\chi_l}\Big[f(l)\frac{1}{n}\sum_{i=1}^{n}\big(\gamma(w_{il})^2\phi(w_{il}) + \gamma(w_{li})^2\phi(w_{li})\big)\\ &- \frac{1}{n}\sum_{i=1}^{n} f(i)\,\big(\gamma(w_{il})^2\phi(w_{il}) + \gamma(w_{li})^2\phi(w_{li})\big)\Big].\end{aligned}$$

Proposition 2.3 *Δ is self-adjoint and positive semi-definite.*

Proof: This follows directly from the definition since

$$\langle f, \Delta g\rangle_V = \langle df, dg\rangle_E = \langle \Delta f, g\rangle_V\,, \quad \langle f, \Delta f\rangle_V = \langle df, df\rangle_E \geq 0.$$

□

2.1.4 The special case of an undirected graph

In the case of an undirected graph we have $w_{ij} = w_{ji}$ that is whenever we have an edge from i to j there is an edge with the same value from j to i. In this case also, $d_i^{out} \equiv d_i^{in}$, so that we will denote the degree function by d with $d_i = \frac{1}{n}\sum_{j=1}^{n} w_{ij}$.

The same for the weights in H_V, $\chi_{out} \equiv \chi_{in}$, so that we have only one function χ. If one likes to interpret functions on E as flows, it is reasonable to restrict the space $\mathcal{H}_E$ to antisymmetric functions since symmetric functions are associated to flows which transport the same mass from vertex i to vertex j and back. Therefore, as a net effect, no mass is transported at all so that from a physical point of view these functions cannot be observed at all. However we will not do this restriction explicitly since in our case we consider anyway only functions on the edges of the form df (where f is in $\mathcal{H}_V$) which are by construction antisymmetric.
The adjoint d^* simplifies to

$$(d^*u)(l) = \frac{1}{2\chi(d_l)}\frac{1}{n}\sum_{i=1}^{n}\gamma(w_{il})\phi(w_{il})(u_{il} - u_{li}),$$

and the general graph Laplacian on an undirected graph has the following form:

Definition 2.4 (graph Laplacian for an undirected graph) *Given Hilbert spaces $\mathcal{H}(V,\chi)$ and $\mathcal{H}(E,\phi)$ and the difference operator $d : \mathcal{H}(V,\chi) \to \mathcal{H}(E,\phi)$ the graph Laplacian $\Delta : \mathcal{H}(V,\chi) \to \mathcal{H}(V,\chi)$ is defined as*

$$\Delta = d^*d$$

Explicitly

$$(\Delta f)(l) = (d^*df)(l) = \frac{1}{\chi(d_l)}\left[f(l)\frac{1}{n}\sum_{i=1}^{n}\gamma^2(w_{il})\phi(w_{il}) - \frac{1}{n}\sum_{i=1}^{n}f(i)\gamma^2(w_{il})\phi(w_{il})\right] \tag{2.6}$$

In the literature one finds the following special cases of the general graph Laplacian. The first one is called the 'normalized' graph Laplacian:

$$(\Delta_{\text{norm}}f)(i) = f(i) - \frac{1}{d(i)}\frac{1}{n}\sum_{j=1}^{n}w_{ij}f(j), \quad \Delta_{norm}f = (\mathbb{1} - P)f, \tag{2.7}$$

where the matrix P is defined as $P = D^{-1}W$ with $D_{ij} = d_i\,\delta_{ij}$. Note that P is a stochastic matrix and therefore can be used to define a Markov random walk on V, see e.g. [106] for more on this connection. The 'unnormalized' graph Laplacian is defined as

$$(\Delta_{\text{unnorm}}f)(i) = d(i)f(i) - \frac{1}{n}\sum_{j=1}^{n}w_{ij}f(j), \quad \Delta_{unorm}f = (D - W)f. \tag{2.8}$$

We have the following conditions on χ, γ and ϕ in order to get these Laplacians:

$$\text{norm}: \quad \frac{1}{\chi(d_i)}\frac{1}{n}\sum_{j=1}\gamma^2(w_{ij})\phi(w_{ij}) = 1,\ \forall\, i, \qquad \text{and} \quad \frac{\gamma^2(w_{ij})\phi(w_{ij})}{\chi(d_i)} = \frac{w_{ij}}{d_i},$$

$$\text{unnorm}: \quad \frac{1}{\chi(d_i)}\frac{1}{n}\sum_{j=1}\gamma^2(w_{ij})\phi(w_{ij}) = d_i,\ \forall\, i, \qquad \text{and} \quad \frac{\gamma^2(w_{ij})\phi(w_{ij})}{\chi(d_i)} = w_{ij}.$$

We observe that there exist several choices of χ,γ and ϕ which result in Δ_{norm} or Δ_{unnorm}. Therefore it can cause confusion if one speaks of the 'normalized' or 'unnormalized' graph Laplacian without explicitly defining the corresponding Hilbert

spaces and the difference operator. If one fixes ϕ and γ to be only functions of w_{ij}, then it is easy to see that $\chi(d_i) = d_i$ and $\chi(d_i) = 1$ are the only possible choices for χ for the normalized respectively unnormalized graph Laplacian. In general there is no natural way to fix all the structure. However for the special setting we consider where the points are sampled from a probability measure on a submanifold in $\mathbb{R}^d$, we will show in Section 2.3.6 that the choices of $\chi(d_i) = d_i$ for the normalized and $\chi(d_i) = 1$ for the unnormalized graph Laplacian are the only choices for $\mathcal{H}_V$ which will lead in general to mutual consistency of the limits of $\mathcal{H}_V$ and Δ.

2.1.5 Smoothness functionals for regularization on undirected graphs

The Laplacian can be used to define a smoothness functional $S : \mathbb{R}^V \to \mathbb{R}_+$ by

$$S(f) = \langle df, df \rangle_{\mathcal{H}_E} = \langle f, \Delta f \rangle_{\mathcal{H}_V} .$$

Note that the same smoothness functional can be defined also for directed graphs. Using our general definition of the graph Laplacian for undirected graphs we arrive at:

$$\begin{aligned} S(f) &= \frac{1}{n} \sum_{l=1}^{n} f(l) \left[f(l) \frac{1}{n} \sum_{i=1}^{n} \gamma(w_{il})^2 \phi(w_{il}) - \frac{1}{n} \sum_{i=1}^{n} f(i) \gamma(w_{il})^2 \phi(w_{il}) \right] \\ &= \frac{1}{2n^2} \sum_{i,l=1}^{n} (f(l) - f(i))^2 \gamma(w_{il})^2 \phi(w_{il}). \end{aligned} \tag{2.9}$$

From the explicit expression of $S(f)$ we deduce the following result:

Proposition 2.5 *The smoothness functional $S(f) = \langle df, df \rangle_{\mathcal{H}_E} = \langle f, \Delta f \rangle_{\mathcal{H}_V}$ induced by the graph Laplacian Δ is independent of the choice of the inner product in $\mathcal{H}_V$. Moreover $S(f)$ depends only on the product $\gamma(w_{il})^2 \phi(w_{il})$.*

This implies that in a learning algorithm where one uses as the loss function $l(y, f) = \|y - f\|^2_{\mathcal{H}_V}$ and $S(f)$ as the regularizer:

$$\min_{f \in \mathcal{H}_V} \|f - y\|^2_{\mathcal{H}_V} + \lambda S(f), \qquad \lambda > 0$$

one can choose the norm used to measure the loss of f independent from the smoothness functional $S(f)$.
The smoothness functional $S(f)$ penalizes a discrete version of the first derivative of f. Similarly to regularization in Euclidean space one can extend this to higher-order derivatives by considering powers of the Laplacian Δ. That is for all $k \in \mathbb{N}$. We define

$$S^k(f) := \left\langle f, \Delta^k f \right\rangle_{\mathcal{H}_V} .$$

Then one has similar to the Euclidean case for $k \in \mathbb{N}$,

$$S^{2k}(f) = \left\langle \Delta^k f, \Delta^k f \right\rangle_{\mathcal{H}_V} , \quad \text{and} \quad S^{2k+1}(f) = \left\langle d\Delta^k f, d\Delta^k f \right\rangle_{\mathcal{H}_E} .$$

Note however that for these higher-order derivatives Proposition 2.5 no longer holds. The most general form of such a regularizer is then given by

$$\Omega(f) = \sum_{k=0}^{\infty} a_k S_k(f) = \sum_{k=0}^{\infty} a_k \left\langle f, \Delta^k f \right\rangle_{\mathcal{H}_V} ,$$

where $\sum_{k=0}^{\infty} |a_k| < \infty$. Such regularizers have been studied by Kondor, Lafferty and Smola in [55, 85]. There a slightly different point of view was developed for the special case of the unnormalized Laplacian. Namely they used the spectral decomposition of $\Delta_{\text{unnorm}} = D - W$. It is straightforward to do the same analysis for general graph Laplacians Δ. Since Δ is self-adjoint in $\mathcal{H}(V, \chi)$ and $|V| = n < \infty$, there exists an orthonormal basis for $\mathcal{H}(V, \chi)$ consisting of eigenvectors of Δ. We denote by λ_i, $i = 1, \ldots, n$ the eigenvalues in increasing order and by u_i, $i = 1, \ldots, n$ the set of normalized[3] eigenvectors of the graph Laplacian Δ, that is

$$\Delta u_i = \lambda_i u_i.$$

Note that Δ is positive semi-definite and therefore $\lambda_i \geq 0$. Then the spectral decomposition of Δ is as follows

$$\Delta = \sum_{i=1}^{n} \lambda_i \, u_i \otimes u_i.$$

One could get here the wrong impression that Δ is a symmetric matrix, however the spectral decomposition has always to be understood with respect to the inner product in $\mathcal{H}_V$, so that

$$\Delta f = \sum_{i=1}^{n} \lambda_i \, \langle u_i, f \rangle_V \, u_i.$$

The general smoothness functional $S_k(f)$ can then be written as

$$S_k(f) = \left\langle f, \Delta^k f \right\rangle_V = \sum_{i=1}^{n} \lambda_i^k \, \langle u_i, f \rangle_V^2 \,,$$

and the general regularizer $\Omega(f)$ becomes

$$\Omega(f) = \sum_{k=0}^{\infty} \sum_{i=1}^{n} a_k \lambda_i^k \, \langle u_i, f \rangle_V^2 = \sum_{i=1}^{n} \rho(\lambda_i) \, \langle u_i, f \rangle_V^2 \,,$$

where we have introduced the function $\rho(\lambda) = \sum_{k=0}^{\infty} a_k \lambda^k$. One can see the spectral decomposition of functions in $\mathcal{H}_V$ as a discrete analogue to the Fourier spectral decomposition of functions in $\mathbb{R}^d$ where also the eigenfunctions of the Laplacian are used.

In the last part of this section we want to clarify the difference between the 'normalized' graph Laplacian introduced in spectral graph theory and what we call 'normalized' graph Laplacian which has its origin in discrete potential theory. In spectral graph theory [24] the following matrix is introduced as 'normalized' graph Laplacian

$$\Delta'' = \mathbb{1} - D^{-1/2} W D^{-1/2}$$

This matrix has the advantage over the normalized graph Laplacian $\Delta_{\text{norm}} = \mathbb{1} - D^{-1}W$ that it is symmetric with respect to the standard inner product, that is $\chi(d_i) = 1$. In [107] it was shown that it can be derived as the graph Laplacian $\Delta'' = d^* d$ with $\mathcal{H}(V, 1)$ and $\mathcal{H}(E, 1)$ and the difference operator d described in Equation 2.3. It is easy to see that

$$\Delta'' = D^{1/2} \Delta_{\text{norm}} D^{-1/2},$$

[3] $\langle u_i, u_i \rangle_{\mathcal{H}(V,\chi)} = 1$.

from which one can deduce that Δ'' and Δ_{norm} have the same eigenvalues. Moreover if $\Delta'' v = \lambda v$, then

$$D^{1/2}\Delta_{\text{norm}}D^{-1/2}v = \lambda\, v \quad \Longrightarrow \quad \Delta_{\text{norm}}D^{-1/2}v = D^{-1/2}v,$$

so that $u = D^{-1/2}v$ is a right eigenvector of Δ_{norm}. Therefore from the point of view of spectral theory they are equivalent. Nevertheless since the eigenvectors are different the induced smoothness functional will also be different. Furthermore, as discussed before, Δ'' is not zero on the constant functions on the graph and therefore its limit will not be a variant of a continuous Laplacian. The last point is the main reason why we will not consider the limit of Δ''.

2.2 Submanifolds in $\mathbb{R}^d$ and associated operators

In this section we introduce the basics of the differential geometry of submanifolds in $\mathbb{R}^d$. In particular we treat Riemannian manifolds with measure. In standard Riemannian geometry one usually considers only the standard measure called volume element[4] which is induced from the Lebesgue measure. Effects of a non-uniform measure, in our case the probability measure generating the data, are rarely considered. Only in the last years in the framework of the so-called metric-measure spaces, that are metric spaces with a measure, also interest arose in Riemannian manifolds with measure.

Remark: In the rest of this chapter we use the Einstein summation convention that is over indices occurring twice has to be summed. We use the conventions regarding the definitions of curvature etc. of Lee in [60].

2.2.1 Basics of submanifolds

Submanifolds

We first give the general definition of a submanifold taken from [23].

Definition 2.6 (Submanifold) *A subset M of an n-dimensional manifold X is an m-dimensional submanifold M if every point $x \in M$ is in the domain of a chart (U, ϕ) of X such that*

$$\phi : U \cap M \to \mathbb{R}^m \times a, \quad \phi(x) = (x^1, \ldots, x^m, a^1, \ldots, a^{n-m})$$

where a is a fixed element in $\mathbb{R}^{n-m}$.

Note that this definition excludes irregular cases like intersecting submanifolds or self-approaching submanifolds. In the following it is more appropriate to take the following point of view. Let M be an m-dimensional manifold. The mapping $i : M \to X$ is said to be an immersion if i is differentiable and the differential of i has rank m everywhere. An injective immersion is called embedding. Then $i(M)$ is a manifold. A regular embedding is an embedding where the manifold structure of $i(M)$ is equivalent to the submanifold structure in X. This is not the case e.g. if $i(M)$ is self-approaching.

Let $i : M \to X$ be a regular embedding so that M is a submanifold of X. Now

[4] Sometimes it is also called volume form since it is an m-form where m is the dimension of the manifold M. In coordinates x^i it is explicitly given as $dV(x) = \sqrt{\det g}\, dx^1 \wedge \ldots \wedge dx^m$.

let X be a Riemannian manifold, that is X has a Riemannian metric h. Then one can induce a metric g on M using the mapping i, namely $g = i^*h$, where $i^* : T^*_{i(x)}X \to T^*_x M$ is the pull-back[5] of the differentiable mapping i. In this case i trivially is an isometric embedding.
In our setting we will study the case $X = \mathbb{R}^d$ and we will always assume that the submanifold M is equipped with the metric induced from $\mathbb{R}^d$. Often this natural assumption is made implicit in machine learning papers.

Normal coordinates

In the proofs we often use normal coordinates. Therefore we give here a very short description, what normal coordinates are. Intuitively normal coordinates around a point p of an m-dimensional Riemannian manifold M are coordinates chosen such that M looks around p like $\mathbb{R}^m$ in the best possible way. This is achieved by adapting the coordinate lines to geodesics through the point p. The reference for the following material is [52].
First we define the exponential map of M at p. For that reason denote by c_v the unique geodesic starting at $c(0) = x$ with tangent vector $\dot{c}(0) = v$ (c_v depends smoothly on p and v).

Definition 2.7 (Exponential Map) *Let M be a Riemannian manifold, $p \in M$, $V_p = \{v \in T_pM,\ c_v \text{ defined on } [0,1]\}$, then*

$$\exp_p : V_p \to M, \quad v \to c_v(1),$$

is called the exponential map of M at p.

It can be shown that $\exp_p$ maps a neighborhood of $0 \in T_pM$ diffeomorphically onto a neighborhood U of $p \in M$. This justifies the definition of normal coordinates.

Definition 2.8 (Normal coordinates) *Let U be a neighborhood of p in M such that $\exp_p$ is a diffeomorphism. The local coordinates defined by the chart $(U, \exp_p^{-1})$ are called normal coordinates with center p.*

Note that in $\mathbb{R}^m \supset \exp_p^{-1}(U)$ we use always an orthonormal basis. The following concept called injectivity radius describes the largest ball around a point p such that normal coordinates can be introduced.

Definition 2.9 (Injectivity radius) *Let M be a Riemannian manifold, $p \in M$. Then the injectivity radius of p is*

$$\text{inj}(p) = \sup\{r > 0, \exp_p \text{ is defined on } \overline{B_{\mathbb{R}^m}(0,\rho)} \text{ and injective}\}.$$

It can be shown that $\text{inj}(p) > 0, \forall p \in M \backslash \partial M$. Moreover for compact manifolds without boundary there exists a lower bound $\text{inj}_{\min} > 0$ such that $\text{inj}(p) \geq \text{inj}_{\min}, \forall p \in M$. However for manifolds with boundary one has $\text{inj}(p_n) \to 0$ for any sequence of points p_n with limit on the boundary.
The motivation for introducing normal coordinates is that the geometry is especially simple in these coordinates. The following Theorem makes this more precise.

[5] $T^*_x M$ is the dual of the tangent space T_xM. Every differentiable mapping $i : M \to X$ induces a pull-back $i^* : T^*_{i(x)}X \to T^*_x M$. Let $u \in T_xM$, $w \in T^*_{i(x)}X$ and denote by i' the differential of i. Then i^* is defined by $(i^*w)(u) = w(i'u)$.

Theorem 2.10 *In normal coordinates around p one has for the Riemannian metric g and the Christoffel symbols $\Gamma^i{}_{jk}\partial_i^a = \partial_j^b \nabla_b \partial_k^a$ at $p = \exp^{-1}(0)$,*

$$g_{ij}(0) = \delta_{ij}, \quad g_{ij,k}(0) = 0, \quad \Gamma^i{}_{jk}(0) = 0.$$

The second derivative of the metric cannot be made to vanish in general. There curvature effects come into play which cannot be deleted by a coordinate transformation. To summarize, normal coordinates with center p achieve that up to first order the geometry of M at point p looks like that of $\mathbb{R}^m$.

The second fundamental form

Let M be an isometrically embedded submanifold of a manifold X. At each point $p \in M$ one can decompose the tangent space T_pX into a subspace T_pM, which is the tangent space to M, and the orthogonal normal space N_pM. In the same way one can split the covariant derivative of X at p, $\nabla_U V$ into a component tangent $(\nabla_U V)^\top$ and normal $(\nabla_U V)^\perp$ to M.

Definition 2.11 *The second fundamental form Π of an isometrically embedded submanifold M of X is defined as*

$$\Pi : T_pM \otimes T_pM \to N_pM, \quad \Pi(U,V) = (\nabla_U V)^\perp$$

The following theorem, see [60], then shows that the covariant derivative of M at p is nothing else than the projection of the covariant derivative of X at p onto T_pM.

Theorem 2.12 (Gauss Formula) *Let U, V be vector fields on M which are arbitrarily extended to X, then the following holds along M*

$$\tilde{\nabla}_U V = \nabla_U V + \Pi(U,V)$$

where $\tilde{\nabla}$ is the covariant derivative of X and ∇ the covariant derivative of M.

The second fundamental form provides also a connection between the curvature tensors of X and M.

Theorem 2.13 (The Gauss equation) *For any $U, V, W, Z \in T_pM$ the following equation holds*

$$\tilde{R}(U,V,W,Z) = R(U,V,W,Z) - \langle \Pi(U,Z), \Pi(V,W) \rangle + \langle \Pi(U,W), \Pi(V,Z) \rangle ,$$

where $\tilde{R}$ is the Riemann curvature[6] tensor of X and R the curvature tensor of M.

Later on we are interested to connect distances in M to the corresponding distances in X. Since Riemannian manifolds are length spaces and therefore the distance is induced by minimizing curves (locally the geodesics), it is of special interest to connect properties of curves of M with respect to X. Applying the Gauss Formula to a curve $c(t) : (a,b) \to M$ yields the following

$$\tilde{D}_t V = D_t V + \Pi(V, \dot{c}),$$

[6]The Riemann curvature tensor of a Riemannian manifold M is defined as $R : T_pM \otimes T_pM \otimes T_pM \to T_p^*M$,

$$R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z.$$

In local coordinates x^i, $R_{ijk}{}^l \partial_l = R(\partial_i, \partial_j)\partial_k$ and $R_{ijkm} = g_{lm} R_{ijk}{}^l$.

where $\tilde{D}_t = \dot{c}^a \tilde{\nabla}_a$ and $\dot{c}$ is the tangent vector field to the curve $c(t)$. Now let $c(t)$ be a geodesic parameterized by arc-length, that is with unit-speed, then its acceleration fulfills $D_t \dot{c} = \dot{c}^a \nabla_a \dot{c}^b = 0$ (however that is only true locally in the interior of M, globally if M has boundary length minimizing curves may behave differently especially if a length minimizing curve goes along the boundary its acceleration can be non-zero), and one gets for the acceleration in the ambient space

$$\tilde{D}_t \dot{c} = \Pi(\dot{c}, \dot{c}).$$

In our setting where $X = \mathbb{R}^d$ the term $\tilde{D}_t \dot{c}$ is just the ordinary acceleration $\ddot{c}$ in $\mathbb{R}^d$. Remember that the norm of the acceleration vector is inverse to the curvature of the curve at that point (if c is parameterized by arc-length[7]). Due to this connection it becomes more apparent why the second fundamental form is often called the extrinsic curvature (with respect to X).
The following Lemma gives an explicit expression of the second fundamental form Π in normal coordinates in the case where M is an isometrically embedded submanifold of $\mathbb{R}^d$. It turns out that in normal coordinates Π is just given by the Hessian of i.

Lemma 2.14 *Let $e_\alpha, \alpha = 1, \ldots, d$ denote an orthonormal basis of $T_{i(x)}\mathbb{R}^d$ then the second fundamental form of M in normal coordinates y is given as:*

$$\Pi(\partial_{y^i}, \partial_{y^j})\Big|_0 = \frac{\partial^2 i^\alpha}{\partial y^i \partial y^j} e_\alpha$$

Proof: Let $\tilde{\nabla}$ be the flat connection of $\mathbb{R}^d$ and ∇ the connection of M. Then by Theorem 2.12,

$$\Pi(\partial_{y^i}, \partial_{y^j}) = \tilde{\nabla}_{i^* \partial_{y^i}} (i^* \partial_{y^j}) - \nabla_{\partial_{y^i}} \partial_{y^j} = \partial_{y^i} \left(\frac{\partial i^\alpha}{\partial_{y^j}} \right) e_\alpha = \frac{\partial^2 i^\alpha}{\partial y^i \partial y^j} e_\alpha,$$

where the second equality follows from the flatness of $\tilde{\nabla}$ and $\Gamma^i_{jk}\Big|_0 = 0$ in normal coordinates. □

Manifolds with boundary of bounded geometry

We will consider in general non-compact submanifolds with boundary. In textbooks on Riemannian geometry one usually only finds material for the case where the manifold has no boundary. Also the analysis e.g. definition of Sobolev spaces on non-compact Riemannian manifolds seems to be non-standard. We profit here very much from the thesis and an accompanying paper of Schick [77, 78] which introduces manifolds with boundary of bounded geometry. All material of this section is taken from these articles. Naturally this plus of generality leads also to a slightly larger technical overload. Nevertheless we think that it is worth this effort since the class of manifolds with boundary of bounded geometry includes almost any kind of submanifold one could have in mind. Moreover, to our knowledge, it is the most general setting where one can still introduce a notion of Sobolev spaces with the usual properties.
Note here that the boundary ∂M is an isometric submanifold of M. As introduced

[7] Note that if c is parameterized by arc-length $\dot{c}$ is tangent to M, that is in particular $\|\dot{c}\|_{T_x X} = \|\dot{c}\|_{T_x M}$

in the last section, it therefore has a second fundamental form $\overline{\Pi}$ which should not be mixed up with the second fundamental form Π of M which is with respect to the ambient space $\mathbb{R}^d$. We denote by $\overline{\nabla}$ the connection and by $\overline{R}$ the curvature of ∂M. Moreover let ν be the normal inward vector field at ∂M and let K be the normal geodesic flow defined as $K : \partial M \times [0,\infty) \to M : (x',t) \to \exp^M_{x'}(t\nu_{x'})$. Then the collar set $N(s)$ is defined as $N(s) := K(\partial M \times [0,s])$ for $s \geq 0$.

Definition 2.15 (Manifold with boundary of bounded geometry) *Suppose M is a manifold with boundary ∂M (possibly empty). It is of bounded geometry if the following holds:*

- *(N) Normal Collar: there exists $r_C > 0$ so that the geodesic collar*

$$\partial M \times [0, r_C) \to M : (t,x) \to \exp_x(t\nu_x)$$

  is a diffeomorphism onto its image (ν_x is the inward normal vector).

- *(IC) The injectivity radius $r_{\text{inj}}(\partial M)$ of ∂M is positive.*

- *(I) Injectivity radius of M: There is $r_i > 0$ so that if $r \leq r_i$ then for $x \in M \backslash N(r)$ the exponential map is a diffeomorphism on $B_M(0,r) \subset T_xM$ so that normal coordinates are defined on every ball $B_M(x,r)$ for $x \in M \backslash N(r)$.*

- *(B) Curvature bounds: For every $k \in \mathbb{N}$ there is C_k so that $|\nabla^i R| \leq C_k$ and $\overline{\nabla}^i \overline{\Pi} \leq C_k$ for $0 \leq i \leq k$.*

Note that (B) imposes bounds on all orders of the derivatives of the curvatures. One could also restrict the definition to the order of derivatives needed for the goals one pursues. But this would require even more notational effort, therefore we skip this. In particular in [77] it is argued that boundedness of all derivatives of the curvature is very close to the boundedness of the curvature alone.
The lower bound on the injectivity radius of M and the bound on the curvature are standard to define manifolds of bounded geometry without boundary. Now the problem of the injectivity radius of M is that at the boundary it somehow makes only partially sense since $\text{inj}(x) \to 0$ as $d(x, \partial M) \to 0$. Therefore one replaces next to the boundary standard normal coordinates with normal collar coordinates.

Definition 2.16 (normal collar coordinates) *Let M be a Riemannian manifold with boundary ∂M. Fix $x' \in \partial M$ and an orthonormal basis of $T_{x'}\partial M$ to identify $T_{x'}\partial M$ with $\mathbb{R}^{m-1}$. For $r_1, r_2 > 0$ sufficiently small (such that the following map is injective) define normal collar coordinates,*

$$n_{x'} : B_{\mathbb{R}^{m-1}}(0,r_1) \times [0,r_2] \to M : (v,t) \to \exp^M_{\exp^{\partial M}_{x'}(v)}(t\nu).$$

The tuple (r_1, r_2) is called the width of the normal collar chart $n_{x'}$.

The following proposition shows why manifolds of bounded geometry are especially interesting.

Proposition 2.17 *Assume that conditions $(N), (IC), (I)$ of Definition 2.15 hold.*

- *(B1) There exist $0 < R_1 \leq r_{\text{inj}}(\partial M)$, $0 < R_2 \leq r_C$ and $0 < R_3 \leq r_i$ and constants $C_K > 0$ for each $K \in \mathbb{N}$ such that whenever we have normal boundary*

coordinates of with (r_1, r_2) with $r_1 \leq R_1$ and $r_2 \leq R_2$ or normal coordinates of radius $r_3 \leq r_i$ then in these coordinates

$$|D^\alpha g_{ij}| \leq C_K \quad \text{and} \quad |D^\alpha g^{ij}| \leq C_K \quad \text{forall} \quad |\alpha| \leq K.$$

The condition (B) in Definition 2.15 holds if and only if $(B1)$ holds. The constants C_K can be chosen to depend only on $r_i, r_C, r_{\text{inj}}(\partial M)$ and C_k.

Note that due to $g^{ij}g_{jk} = \delta^i_k$ one gets upper and lower bounds on g_{ij} resp. g^{ij} which result in upper and lower bounds for $\sqrt{\det g}$. This implies that we have upper and lower bounds on the volume form $dV(x) = \sqrt{\det g}\, dx$ which will be used in the following lemma.

Lemma 2.18 *Let (M, g) be a Riemannian manifold with boundary of bounded geometry of dimension m. Then there exists $R_0 > 0$ and constants $S_1 > 0$ and S_2 such that for all $x \in M$ and $r \leq R_0$ one has*

$$S_1 r^m \leq \text{vol}(B_M(x, r)) \leq S_2 r^m$$

Another important tool for analysis on manifolds are appropriate function spaces. In order to define a Sobolev norm one first has to fix a family of charts U_i with $M \subset \cup_i U_i$ and then define the Sobolev norm with respect to these charts. The resulting norm will depend on the choice of the charts U_i. Since in differential geometry the choice of the charts should not matter, the natural question arises how the Sobolev norm corresponding to a different choice of charts V_i is related to that for the choice U_i. In general, the Sobolev norms will not be the same but if one assumes that the transition maps are smooth and the manifold M is compact then the resulting norms will be equivalent and therefore define the same topology. Now if one has a non-compact manifold this argumentation does not work anymore. This problem is solved by Schick in [77] by defining a partition of unity on M based on normal coordinate charts. Then it can be shown that the change of coordinates between these normal coordinate charts is well-behaved due to the bounded geometry of M. In that way it is possible to establish a well-defined notion of Sobolev spaces on manifolds with boundary of bounded geometry. In particular due to the uniform bounds on the transition maps between normal coordinate charts derived in [77] the following norm on $C^k(M)$ makes sense[8]:

$$\|f\|_{C^k(M)} = \sup_{\sum_{i=1}^m l_i \leq k,\, x \in M} \left| \frac{\partial^{|\sum_{i=1}^m l_i|}}{\partial (x^1)^{l_1} \dots \partial (x^m)^{l_m}} f(x) \right|,$$

where x^i are coordinate functions with respect to some normal chart. In the following we will denote with $C^k(M)$ the space of C^k-functions on M together with the norm $\|\cdot\|_{C^k(M)}$.

Intrinsic versus extrinsic properties

Most of the proofs for the continuous part will work with Taylor expansions in normal coordinates. It is then of special interest to have a connection between intrinsic and extrinsic distances. Since the distance on M is induced from $\mathbb{R}^d$, it is obvious

[8] Any other norm with respect to a different choice of normal coordinate charts will be equivalent.

that one has $\|x-y\|_{\mathbb{R}^d} \sim d_M(x,y)$ for all $x, y \in M$ which are sufficiently close. The next proposition proven by Smolyanov, Weizsäcker and Wittich in [86] provides an asymptotic expression of geometric quantities of the submanifold M in the neighborhood of a point $x \in M$. Particularly it gives a third-order approximation of the intrinsic distance $d_M(x,y)$ in M in terms of the extrinsic distance in the ambient space X which is in our case just the Euclidean distance in $\mathbb{R}^d$.

Proposition 2.19 *Let $i : M \to \mathbb{R}^d$ be an isometric embedding of the smooth m-dimensional Riemannian manifold M into $\mathbb{R}^d$. Let $x \in M$ and V be a neighborhood of 0 in $\mathbb{R}^m$ and let $\Psi : V \to U$ provide normal coordinates of a neighborhood U of x, that is $\Psi(0) = x$. Then for all $y \in V$:*

$$\|y\|^2_{\mathbb{R}^m} = d^2_M(x, \Psi(y)) = \|(i \circ \Psi)(y) - i(x)\|^2_{\mathbb{R}^d} + \frac{1}{12} \|\Pi(\dot\gamma, \dot\gamma)\|^2_{T_x\mathbb{R}^d} + O(\|x\|^5_{\mathbb{R}^m}),$$

where Π is the second fundamental form of M and γ the unique geodesic from x to $\Psi(y)$ such that $\dot\gamma = y^i \partial_{y^i}$.
The volume form $dV = \sqrt{\det\, g_{ij}(y)}dy$ of M satisfies in normal coordinates,

$$dV = \left(1 + \frac{1}{6} R_{iuvi}\, y^u y^v + O(\|y\|^3_{\mathbb{R}^m})\right) dy,$$

in particular

$$(\Delta \sqrt{\det\, g_{ij}})(0) = -\frac{1}{3} R,$$

where R is the scalar curvature (i.e., $R = g^{ik} g^{jl} R_{ijkl}$).

We would like to note that in [87] this proposition was formulated for general ambient spaces X, that is arbitrary Riemannian manifolds X. Using the more general form of this proposition one could extend the whole setting to submanifolds of other ambient spaces X. However in order to use the scheme, one needs an explicit expression of the distances in X which is usually not available for general Riemannian manifolds. Nevertheless for some special cases, like the sphere, one knows the geodesic distance. Submanifolds of the sphere could be of interest e.g. in geophysics or astronomy.
The previous proposition is very helpful since it gives an asymptotic expression of the geodesic distance $d_M(x,y)$ on M in terms of the extrinsic Euclidean distance. The following lemma is a non-asymptotic statement taken from [15] which we present in a slightly different form. However we will first establish a connection between what they call the 'minimum radius of curvature' and upper bounds on the extrinsic curvatures of M and ∂M. Let

$$\Pi_{\max} = \sup_{x \in M} \sup_{v \in T_x M, \|v\|=1} \|\Pi(v,v)\|,$$

and let

$$\overline{\Pi}_{\max} = \sup_{x \in \partial M} \sup_{v \in T_x \partial M, \|v\|=1} \left\|\overline{\Pi}(v,v)\right\|,$$

where $\overline{\Pi}$ is the second fundamental form of ∂M as a submanifold of M. We set $\overline{\Pi}_{\max} = 0$ if the boundary ∂M is empty.
Using the relation between the acceleration in the ambient space and the second fundamental form for unit-speed curves γ with no acceleration in M ($D_t \dot c = 0$)

established in section 2.2.1, we get for the Euclidean acceleration of the curve γ in $\mathbb{R}^d$:

$$\|\ddot{\gamma}\| = \|\Pi(\dot{\gamma}, \dot{\gamma})\|.$$

Now if one has a non-empty boundary ∂M it can happen that one has a length-minimizing curve which goes (partially) along the boundary (imagine $\mathbb{R}^d$ with a ball at the origin cut out). Then the segment c along the boundary will be a geodesic of the submanifold ∂M, see [2], that is $\overline{D}_t\dot{c} = \overline{\nabla}_{\dot{c}}\dot{c} = 0$ where $\overline{\nabla}$ is the connection of ∂M induced by M. However c will not be a geodesic in M (in the sense of a curve with no acceleration) since by the Gauss-Formula 2.12

$$D_t\dot{c} = \overline{D}_t\dot{c} + \overline{\Pi}(\dot{c}, \dot{c}) = \overline{\Pi}(\dot{c}, \dot{c}).$$

In general therefore the upper bound on the Euclidean acceleration of a length-minimizing curve γ in M is given by

$$\|\ddot{\gamma}\| = \left\|\overline{\Pi}(\dot{\gamma}, \dot{\gamma}) + \Pi(\dot{\gamma}, \dot{\gamma})\right\| \leq \overline{\Pi}_{\max} + \Pi_{\max}.$$

Using this inequality, one can derive a lower bound on the 'minimum radius of curvature' ρ defined in [15] as $\rho = \inf\{1/\|\ddot{\gamma}\|_{\mathbb{R}^d}\}$ where the infimum is taken over all unit-speed geodesics γ of M (in the sense of length-minimizing curves):

$$\rho \geq \frac{1}{\overline{\Pi}_{\max} + \Pi_{\max}}.$$

Finally we can formulate the Lemma from [15].

Lemma 2.20 *Let $x, y \in M$ with $d_M(x, y) \leq \pi\rho$ then*

$$2\rho \sin(d_M(x, y)/(2\rho)) \leq \|x - y\|_{\mathbb{R}^d} \leq d_M(x, y).$$

Noting that $\sin(x) \geq x/2$ for $0 \leq x \leq \pi/2$, we get as an easier to handle corollary:

Corollary 2.21 *Let $x, y \in M$ with $d_M(x, y) \leq \pi\rho$ then*

$$\frac{1}{2} d_M(x, y) \leq \|x - y\|_{\mathbb{R}^d} \leq d_M(x, y).$$

This corollary is nice however quite useless for our purposes since we will have only the Euclidean distances between points and therefore we have no possibility to check the condition $d_M(x, y) \leq \pi\rho$. In general small Euclidean distance does not imply small intrinsic distance. Imagine a circle where one has cut out a very small segment and consider now a point near to one end. Then the Euclidean distance between the two ends is very small however the geodesic distance is very large. We will now show that under an additional assumption one can transform the above corollary so that one can use it when one has only knowledge about Euclidean distances.

Lemma 2.22 *Let M have a finite radius of curvature $\rho > 0$. We further assume that*

$$\kappa := \inf_{x \in M} \inf_{y \in M \setminus B_M(x, \pi\rho)} \|x - y\|$$

is non-zero. Then $B_{\mathbb{R}^d}(x, \kappa/2) \cap M \subset B_M(x, \kappa) \subset B_M(x, \pi\rho)$. Particularly, if $x, y \in M$ and $\|x - y\| \leq \kappa/2$,

$$\frac{1}{2} d_M(x, y) \leq \|x - y\|_{\mathbb{R}^d} \leq d_M(x, y) \leq \kappa.$$

Proof: By definition κ is at most the infimum of $\|x-y\|$ where y satisfies $d_M(x,y) = \pi\rho$. Therefore the set $B_{\mathbb{R}^d}(x,\kappa/2) \cap M$ is a subset of $B_M(x,\pi\rho)$. The rest of the lemma follows then by Corollary 2.21. □

The Figure 2.2 illustrates this construction:

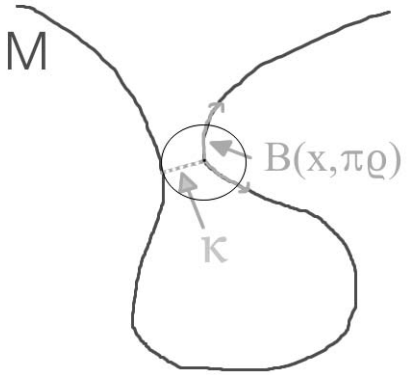


Abbildung 2.2: κ is the Euclidean distance of $x \in M$ to $M \backslash B_M(x,\pi\rho)$.

2.2.2 The weighted Laplacian and the continuous smoothness functional

The Laplacian is one of the most prominent operators in mathematics. The following general properties are taken from the books of Rosenberg [74] and Bérard [13]. It occurs in many partial differential equations governing physics. Mainly because it is in Euclidean space the only second-order differential operator which is translation and rotation invariant. In Euclidean space $\mathbb{R}^d$ it is defined as

$$\Delta_{\mathbb{R}^d} f = \operatorname{div}(\operatorname{grad} f) = \sum_{i=1}^{d} \partial_i^2 f.$$

Moreover for any domain $\Omega \subseteq \mathbb{R}^d$ it is a negative-semidefinite symmetric operator on $C_c^\infty(\Omega)$ which is a dense subset of $L_2(\Omega)$ (formally self-adjoint),

$$\int_\Omega f \Delta g \, dx = -\int_\Omega \langle \nabla f, \nabla g \rangle \, dx,$$

and can be extended to a self-adjoint operator on $L_2(\Omega)$ in several ways depending on the choice of boundary conditions. For any compact domain Ω (with suitable boundary conditions) it can be shown that Δ has a pure point spectrum and the eigenfunctions are smooth and form a complete orthonormal basis of $L_2(\Omega)$, see e.g. [13].

The Laplace-Beltrami operator on a manifold M is the natural equivalent of the Laplacian in $\mathbb{R}^d$, defined as

$$\Delta_M f = \operatorname{div}(\operatorname{grad} f) = \nabla^a \nabla_a f$$

However the more natural definition is the following. Let $f, g \in C_c^\infty(M)$ then

$$\int_M f \Delta g \, dV(x) = -\int_M \langle \nabla f, \nabla g \rangle \, dV(x),$$

where $dV = \sqrt{\det g}\, dx$ is the natural volume element of M. This definition allows easily an extension to the case where we have a Riemannian manifold M with a measure P where P will be in our case the probability measure generating the data. We assume in the following that P is absolutely continuous wrt the natural volume element dV of the manifold and its density[9] is denoted by p. The notion of weighted Laplacians seems not to be standard in differential geometry. Only quite recently with the emerging interest in metric-measure spaces also interest in Riemannian manifolds equipped with a measure arose.

Definition 2.23 (Weighted Laplacian) *Let (M, g_{ab}) be a Riemannian manifold with measure P where P has a differentiable density p with respect to the natural volume element $dV = \sqrt{\det g}\, dx$, and let Δ_M be the Laplace-Beltrami operator on M. Then we define the s-th weighted Laplacian Δ_s as*

$$\Delta_s := \Delta_M + \frac{s}{p} g^{ab}(\nabla_a p)\nabla_b = \frac{1}{p^s} g^{ab}\nabla_a(p^s \nabla_b) = \frac{1}{p^s}\operatorname{div}(p^s \operatorname{grad}). \tag{2.10}$$

This definition is motivated by the following equality

$$\int_M f(\Delta_s g)\, p^s dV = \int_M f\big(\Delta g + \frac{s}{p}\langle \nabla p, \nabla g\rangle\big) p^s dV = -\int_M \langle \nabla f, \nabla g\rangle\, p^s dV, \tag{2.11}$$

where $f, g \in C_c^\infty(M)$. The family of weighted Laplacians contains two cases which are particularly interesting. The first one, $s = 0$, corresponds to the standard Laplace-Beltrami operator. This case is interesting if one only wants to use properties of the geometry of the manifold but not of the data generating probability measure. The second case, $s = 1$, corresponds to the standard weighted Laplacian $\Delta_1 = \frac{1}{p}\nabla^a(p\nabla_a)$.
In the next section it will turn out that through a data-dependent change of the weights of the graph we can get a subfamily of the just defined weighted Laplacians as the limit operators of the graph Laplacian. The rest of this section will be used to motivate the importance of the understanding of this limit in different applications. Three different however very much connected properties of the Laplacian are used in machine learning

- The Laplacian generates the diffusion process. In semi-supervised learning algorithms with a small number of labeled points one would like to propagate the labels along regions of high-density. The limit operator Δ_s shows the influence of a non-uniform density p. The second term $\frac{s}{p}\langle \nabla p, \nabla f\rangle$ leads to an anisotropy in the diffusion. If $s < 0$ this term enforces diffusion in the direction of the maximum of the density wheras diffusion in the direction away from the maximum of the density is weakenend. If $s > 0$ this is just the other way round.

- The smoothness functional induced by the weighted Laplacian Δ_s, see equation 2.11, is given by
$$S(f) = \int_M \langle \nabla f, \nabla f\rangle\, p^s dV.$$
This smoothness functional prefers for $s > 0$ functions which are smooth in high-density regions whereas unsmooth behavior in low-density is penalized

[9] Note that the case when the probability measure is absolutely continuous wrt the Lebesgue measure on $\mathbb{R}^d$ is a special case of our setting.

less. This property can also be interesting in semi-supervised learning where one assumes especially if only a few labeled points are known that the classifier should be constant in high-density regions whereas changes of the classifier are allowed in low-density regions, see [18] for some discussion of this point. In Figure 2.3 this is illustrated by mapping a density profile in $\mathbb{R}^2$ onto a two-dimensional manifold. However also the case $s < 0$ can be interesting.

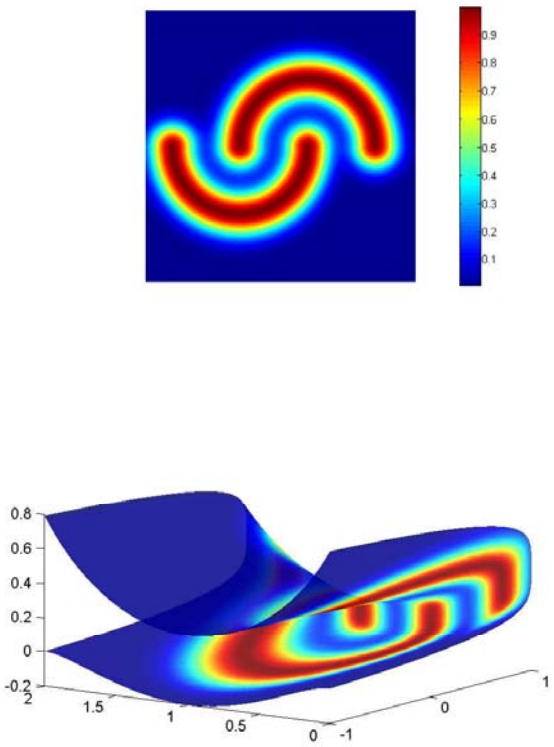


Abbildung 2.3: A density profile mapped onto a 2-dimensional submanifold in $\mathbb{R}^3$ with two clusters.

Minimizing then the smoothness functional $S(f)$, implies that one enforces smoothness of the function f where one has little data, and one allows the function to vary a lot where one has sampled a lot of data points. Such a penalization is more appropriate for regression and has been considered by Canu and Elisseeff in [20].

- The eigenfunctions of the Laplacian Δ_s can be seen as the limit partioning of spectral clustering for the normalized graph Laplacian (however a rigorous mathematical proof has not been given yet). If $s = 0$ one gets just a geometric clustering in the sense that irrespectively of the probability measure generating the data the clustering is determined by the geometry of the submanifold. If $s > 0$ the eigenfunction correponding to the first non-zero eigenvalue is likely to change its sign in a low-density region. This argument follows from the previous discussion on the smoothness functional $S(f)$ and the Rayleigh-Ritz principle. Let us assume for a moment that M is compact without boundary and that $p(x) > 0, \forall\, x \in M$, then the eigenspace corresponding to the first eigenvalue

$\lambda_0 = 0$ is given by the constant functions. The first non-zero eigenvalue can then be determined by the Rayleigh-Ritz variational principle

$$\lambda_1 = \inf_{u \in C^\infty(M)} \left\{ \frac{\int_M \|\nabla u\|^2 p(x)^s dV(x)}{\int_M u^2(x) p(x)^s dV(x)} \;\middle|\; \int_M u(x)\, p(x)^s dV(x) = 0 \right\}.$$

Since the first eigenfunction has to be orthogonal to the constant functions, it has to change its sign. However since $\|\nabla u\|^2$ is weighted by the power of the density p^s it is obvious that for $s > 0$ the function will change its sign in a region of low density.

2.3 Continuum limit of the graph structure

In this section we will treat the continuum limit of certain neighborhood graphs built from random samples of a probability measure P. In particular the case where P has support on a submanifold M will be considered. This is especially interesting if one is given data X where the dimension d of the feature space $\mathbb{R}^d$ is much higher than the number m of intrinsic parameters resp. degrees of freedom of the data.
Let us now introduce the setting more precisely. We assume to have points X_i, $i = 1, \ldots, n$ drawn i.i.d. from the probability measure P which has support on a submanifold M. We see the points X_i as vertices of a graph. We further assume that we are given a kernel function $k : \mathbb{R}_+ \to \mathbb{R}_+$ (see 2.3.1 for the assumptions on this function) and the neighborhood parameter $h \in \mathbb{R}_+^*$. As proposed by Coifman and Lafon in [58, 25], we use this kernel function k to define the following family of data-dependent kernel functions $\tilde{k}_{\lambda,h}$ parameterized by λ as:

$$\tilde{k}_{\lambda,h}(X_i, X_j) = \frac{1}{h^m} \frac{k(\|X_i - X_j\|^2 / h^2)}{[d(X_i) d(X_j)]^\lambda},\ \lambda \geq 0,$$

where $d(X_i) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h^m} k(\|X_i - X_j\|^2 / h^2)$ is the degree function introduced in Section 2.1 with respect to edge-weights $\frac{1}{h^m} k(\|X_i - X_j\|^2 / h^2)$. Finally we use $\tilde{k}$ to define the weight $w_{ij} = w(X_i, X_j)$ of the edge between the points X_i and X_j as

$$w_{\lambda,h}(X_i, X_j) = \tilde{k}_{\lambda,h}(X_i, X_j).$$

Note that the case $\lambda = 0$ corresponds to weights with no data-dependent modifications. The parameter $h \in \mathbb{R}_+^*$ determines the neighborhood of a point since we will assume that k is decreasing or even has compact support. The goal of this section is to study the limits of several graph structures especially of the graph Laplacian as the sample size n goes to infinity and the neighborhood parameter h goes to zero. The limit respectively the pointwise consistency of the normalized graph Laplacian has been published in [40].
The setting and the assumptions on the submanifold M are described in Section 2.3.1. In the proofs we often use convolutions of functions on the submanifold M with respect to the extrinsic Euclidean distance. The asymptotic limit of such convolutions is discussed in Section 2.3.2. Then in Section 2.3.3 we study the limit of the degree function. As a byproduct we get results for kernel density estimation on submanifolds of $\mathbb{R}^d$. In Section 2.3.4 we derive the pointwise limit of the so called normalized and unnormalized graph Laplacians with general data-dependent

edge-weights $\tilde{k}_{\lambda,h}$. Finally in Section 2.3.5 we show the limit of the inner product in $\mathcal{H}(V,\chi)$ for $\chi(d) = d$ and the limit of the smoothness functional $S(f) = \langle f, \Delta f\rangle$. In Section 2.3.6 we summarize the results and show that one can partially fix the graph structure by requiring mutual consistency of the limits of $\mathcal{H}_V$ and Δ.

2.3.1 Notations and assumptions

In general we work on complete non-compact manifolds with boundary. Compared to a setting where one considers only compact manifolds one needs a slightly larger technical overhead. However we will indicate how the technical assumptions simplify if one has a compact submanifold with boundary or even only a compact manifold without boundary.
We impose the following assumptions on the submanifold M:

Assumption 2.24
- *$i : M \to \mathbb{R}^d$ is a smooth, isometric embedding,*
- *M is a smooth manifold with boundary of bounded geometry (the boundary ∂M can be empty),*
- *M has bounded second fundamental form,*
- *$\kappa := \inf_{x\in M} \inf_{y\in M\setminus B_M(x,\pi\rho)} \|i(x) - i(y)\| > 0$, where ρ is the radius of curvature defined in Section 2.2.1. It holds $\rho > 0$ since the second fundamental form of M as well as of ∂M are bounded,*
- *for any $x \in M\setminus\partial M$,*

$$\delta(x) := \inf_{y\in M\setminus B_M(x,\frac{1}{3}\min\{\mathrm{inj}(x),\pi\rho\})} \|i(x) - i(y)\|_{\mathbb{R}^d} > 0,$$

where $\mathrm{inj}(x)$ is the injectivity radius[10] at x and $\rho > 0$ is the radius of curvature.

The first condition ensures that M is a smooth isometric submanifold of $\mathbb{R}^d$ as introduced in Section 2.2.1. As discussed in section 2.2.1, manifolds of bounded geometry are in general non-compact, complete Riemannian manifolds with boundary where one has uniform control over all intrinsic curvatures. The uniform control allows one to still do reasonable analysis in this general setting. Compact Riemannian submanifolds (with or without boundary) are always of bounded geometry. The third condition ensures that M also has well-behaved extrinsic geometry. This condition together with the fourth condition enables us to get global upper and lower bounds of the intrinsic distance on M in terms of the extrinsic distance in $\mathbb{R}^d$ and vice versa, see Lemma 2.22. The fourth condition is only necessary in the case of non-compact submanifolds. It prevents the manifold from self-approaching. More precisely it ensures that if parts of M are far away from x in the geometry of M they do not come too close to x in the geometry of $\mathbb{R}^d$. Assuming a regular submanifold, this assumption is already included implicitly. However for non-compact submanifolds the self-approaching could happen at infinity. Therefore we exclude it explicitly. Moreover note that for submanifolds with boundary one has $\mathrm{inj}(x) \to 0$ as x approaches the boundary ∂M[11]. Therefore also $\delta(x) \to 0$ as $d(x,\partial M) \to 0$. However for pointwise convergence proofs in the interior of M this behavior of $\delta(x)$ at the boundary

[10] Note that the injectivity radius $\mathrm{inj}(x)$ is always positive.
[11] This is the reason why one replaces normal coordinates in the neighborhood of the boundary with normal collar coordinates.

does not matter.
In order to emphasize the distinction between extrinsic and intrinsic properties of the manifold we always use the slightly cumbersome notations $x \in M$ (intrinsic) and $i(x) \in \mathbb{R}^d$ (extrinsic). The reader who is not familiar with Riemannian geometry should keep in mind that locally, a submanifold of dimension m looks like $\mathbb{R}^m$. This becomes apparent if one uses normal coordinates. Also the following dictionary between terms of the manifold M and the case when one has only an open set in $\mathbb{R}^d$ (i is then the identity mapping) might be useful.

Manifold M	open set in $\mathbb{R}^d$
g_{ij} , $\sqrt{\det g}$	δ_{ij} , 1
natural volume element	Lebesgue measure
Δ_s	$\Delta_s = \sum_{i=1}^d \frac{\partial^2 f}{\partial (z_i)^2} + \frac{s}{p} \sum_{i=1}^d \frac{\partial p}{\partial z^i} \frac{\partial f}{\partial z^i}$

The kernel functions which are used to define the weights of the graph are always functions of the squared norm in $\mathbb{R}^d$. Furthermore, we make the following assumptions on the kernel function k:

Assumption 2.25
- $k : \mathbb{R}_+^* \to \mathbb{R}$ *is measurable, non-negative and non-increasing,*
- $k \in C^2(\mathbb{R}_+^*)$*, that is in particular* k*,* $\frac{\partial k}{\partial x}$ *and* $\frac{\partial^2 k}{\partial x^2}$ *are bounded,*
- k*,* $|\frac{\partial k}{\partial x}|$ *and* $|\frac{\partial^2 k}{\partial x^2}|$ *have exponential decay:* $\exists c, \alpha, A \in \mathbb{R}_+$ *such that for any* $t \geq A$*,* $f(t) \leq ce^{-\alpha t}$*, where* $f(t) = \max\{k(t), |\frac{\partial k}{\partial x}|(t), |\frac{\partial^2 k}{\partial x^2}|(t)\}$*,*
- $k(0) = 0$*,*
- $\exists r_k > 0$ *such that* $k(x) \geq \frac{\|k\|_\infty}{2}$ *for* $x \in]0, r_k]$*.*

The third condition implies that the graph will have no loops[12]. In particular the kernel is not continuous at the origin. One could prove all statements also without this condition. The advantage of this condition is that some estimators become unbiased by imposing this condition. Also let us introduce the helpful notation[13], $k_h(t) = \frac{1}{h^m} k\left(\frac{t}{h^2}\right)$ where we call h the bandwidth of the kernel. Moreover we define the following two constants related to the kernel function k,

$$C_1 = \int_{\mathbb{R}^m} k(\|y\|^2) dy < \infty, \quad C_2 = \int_{\mathbb{R}^m} k(\|y\|^2) y_1^2 dy < \infty. \tag{2.12}$$

We also have some assumptions on the probability measure P.

Assumption 2.26
- P *is absolutely continuous with respect to the natural volume element* dV *on* M*,*
- *the density* p *fulfills:* $p \in C^3(M)$ *and* $p(x) > 0, \forall\, x \in M$*,*
- *the sample* $X_i, i = 1, \ldots, n$ *is drawn i.i.d. from* P*.*

We will call Assumptions 2.24 on the submanifold, Assumptions 2.25 on the kernel function, and the Assumptions 2.26 on the probability measure P together the *standard assumptions*.

[12] An edge from a vertex to itself is called a loop.
[13] In order to avoid problems with differentiation the argument of the kernel function will be the squared norm.

2.3.2 Asymptotics of Euclidean convolutions on the submanifold M

The following proposition describes the asymptotic expression of the convolution of a function f on M with a kernel function depending on the Euclidean distance $\|x-y\|$ on the submanifold M with respect to the probability measure P on M. This result is interesting since it shows how the use of the Euclidean distance introduces a curvature effect if one averages a function locally. A similar result was presented by Lafon in [58]. However the analysis there is only correct if the submanifold is a hypersurface (a submanifold of codimension 1). Moreover in [58] the density p is not invariantly defined with respect to the natural volume element. Since we integrate with respect to the natural volume element we therefore get an additional factor. Our proof is similar to that of Smolyanov, Weizsäcker and Wittich in [87] where under stronger conditions a similar result was proven for the Gaussian kernel. The more general setting and the use of general kernel functions makes the proof a little bit more complicated.

Proposition 2.27 *Let M and k satisfy Assumptions 2.24 and 2.25. Furthermore let P have a density p with respect to the natural volume element and $p \in C^3(M)$. Then for any $x \in M \backslash \partial M$, there exists an $h_0(x) > 0$ such that for all $h < h_0(x)$ and any $f \in C^3(M)$,*

$$\begin{aligned}&\int_M k_h\left(\|i(x)-i(y)\|^2_{\mathbb{R}^d}\right) f(y)p(y)\sqrt{\det g}\,dy\\ &=C_1 p(x)f(x) + \frac{h^2}{2}\,C_2\Big(p(x)f(x)S(x) + (\Delta_M(pf))(x)\Big) + O(h^3),\end{aligned}$$

where

$$S(x) = \frac{1}{2}\Big[-R\big|_x + \frac{1}{2}\Big\|\sum\nolimits_a \Pi(\partial_a,\partial_a)\Big\|^2_{T_{i(x)}\mathbb{R}^d}\Big],$$

and $O(h^3)$ is a function depending on x, $\|f\|_{C^3}$ and $\|p\|_{C^3}$.

The following lemmas are needed in the proof.

Lemma 2.28 *If the kernel $k : \mathbb{R} \to \mathbb{R}$ satisfies the assumptions in Assumption 2.25,*

$$\int_{\mathbb{R}^m} \frac{\partial k}{\partial x}(\|u\|^2)u^i u^j u^k u^l du = -\frac{1}{2}C_2\left[\delta^{ij}\delta^{kl} + \delta^{ik}\delta^{jl} + \delta^{il}\delta^{jk}\right]. \tag{2.13}$$

Proof: Note first that for a function $f(\|u\|^2)$ one has $\frac{\partial f}{\partial \|u\|^2} = \frac{\partial f}{\partial u_i^2}$. The rest follows from partial integration.

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{\partial k}{\partial u^2}(u^2)\,u^2\,du &= \int_0^{\infty} \frac{\partial k}{\partial v}(v)\,\sqrt{v}\,dv = \left[k(v)\,\sqrt{v}\right]_0^{\infty} - \int_0^{\infty} k(v)\,\frac{1}{2\sqrt{v}}\,dv\\ &= -\frac{1}{2}\int_{-\infty}^{\infty} k(u^2)du,\end{aligned}$$

where $[k(v)\,\sqrt{v}]_0^{\infty} = 0$ due to the boundedness and exponential decay of k.
In the same way one can derive

$$\int_{-\infty}^{\infty} \frac{\partial k}{\partial u^2}(u^2)\,u^4\,du = -\frac{3}{2}\int_{-\infty}^{\infty} k(u^2)u^2\,du$$

With these results and the fact that since k is an even function only integration over even powers of coordinates will be non-zero the proof is finished. □

Lemma 2.29 *Let k satisfy Assumption 2.25 and let V_{ijkl} be a given tensor. Assume now $\|z\|^2 \geq \|z\|^2 + V_{ijkl}z^i z^j z^k z^l + \beta(z)\|z\|^5 \geq \frac{1}{4}\|z\|^2$ on $B(0, r_{\min}) \subset \mathbb{R}^m$, where $\beta(z)$ is continuous and $\beta(z) \sim O(1)$ as $z \to 0$. Then there exists a constant C and a $h_0 > 0$ such that for all $h < h_0$ and all $f \in C^3(B(0, r_{\min}))$,*

$$\Big| \int_{B(0,r_{\min})} k_h\left(\frac{\|z\|^2 + V_{ijkl}z^i z^j z^k z^l + \beta(z)\|z\|^5)}{h^2}\right) f(z)dz$$
$$- \Big(C_1 f(0) + C_2\frac{h^2}{2}\Big[(\Delta f)(0) - f(0)\sum_{i,k}^{m} V_{iikk} + V_{ikik} + V_{ikki}\Big]\Big)\Big| \leq Ch^3.$$

where C is a constant depending on k, $r_{\min}$, V_{ijkl} and $\|f\|_{C^3}$.

Proof: As a first step we do a Taylor expansion of the kernel around $\|z\|^2/h^2$:

$$k_h\left(\frac{\|z\|^2 + \eta}{h^2}\right) = k_h\left(\frac{\|z\|^2}{h^2}\right) + \frac{\partial k_h(x)}{\partial x}\Big|_{\frac{\|z\|^2}{h^2}}\frac{\eta}{h^2} + \frac{\partial^2 k_h(x)}{\partial x^2}\Big|_{\frac{\|z\|^2(1-\theta)+\theta\,\eta}{h^2}}\frac{\eta^2}{h^4},$$

where in the last term $0 \leq \theta(z) \leq 1$. We then decompose the integral:

$$\int_{B(0,r_{\min})} k_h\left(\frac{\|z\|^2 + V_{ijkl}z^i z^j z^k z^l + \beta(z)\|z\|^5}{h^2}\right) f(z)dz$$
$$= \int_{\mathbb{R}^m}\left(k_h\left(\frac{\|z\|^2}{h^2}\right) + \frac{\partial k_h(x)}{\partial x}\Big|_{\frac{\|z\|^2}{h^2}}\frac{V_{ijkl}\,z^i z^j z^k z^l}{h^2}\right)$$
$$\left(f(0) + \langle \nabla f\big|_0, z\rangle + \frac{1}{2}\frac{\partial^2 f}{\partial z^i \partial z^j}\Big|_0 z^i z^j\right) dz + \sum_{i=0}^{4}\alpha_i$$

where we define the five error terms α_i as:

$$\alpha_0 = \int_{B(0,r_{\min})} \frac{\partial k_h}{\partial x}\Big|_{\frac{\|z\|^2}{h^2}}\frac{\beta(z)\|z\|^5}{h^2} f(z)\,dz,$$

$$\alpha_1 = \int_{B(0,r_{\min})} \frac{\partial^2 k_h}{\partial x^2}\Big|_{\frac{\|z\|^2(1-\theta)+\theta\eta}{h^2}}\frac{\left(V_{ijkl}z^i z^j z^k z^l + \beta(z)\|z\|^5\right)^2}{h^4} f(z)dz,$$

$$\alpha_2 = \int_{B(0,r_{\min})} k_h\left(\frac{\|z\|^2 + V_{ijkl}z^i z^j z^k z^l + \beta(z)\|z\|^5}{h^2}\right)\frac{1}{6}\frac{\partial^3 f}{\partial z^i \partial z^j \partial z^k}(\theta z) z^i z^j z^k dz,$$

$$\alpha_3 = \int_{\mathbb{R}^m \setminus B(0,r_{\min})} k_h\left(\frac{\|z\|^2}{h^2}\right)\left(f(0) + \langle \nabla f\big|_0, z\rangle + \frac{1}{2}\frac{\partial^2 f}{\partial z^i \partial z^j}\Big|_0 z^i z^j\right) dz,$$

$$\alpha_4 = \int_{\mathbb{R}^m \setminus B(0,r_{\min})} \frac{\partial k_h}{\partial x}\Big|_{\frac{\|z\|^2}{h^2}}\frac{V_{ijkl}\,z^i z^j z^k z^l}{h^2}$$
$$\left(f(0) + \langle \nabla f\big|_0, z\rangle + \frac{1}{2}\frac{\partial^2 f}{\partial z^i \partial z^j}\Big|_0 z^i z^j\right) dz,$$

where in α_1, $\eta = V_{ijkl} z^i z^j z^k z^l + \beta(z) \|z\|^5$.
With $\int_{\mathbb{R}^m} k(\|z\|^2)\, z_i\, dz = 0$, $\forall\, i$, $\int_{\mathbb{R}^m} k(\|z\|^2)\, z_i\, z_j dz = 0$ if $i \neq j$, and Lemma 2.28 the main term simplifies to:

$$\begin{aligned}
&\int_{\mathbb{R}^m} \left(k_h\left(\frac{\|z\|^2}{h^2}\right) + \frac{\partial k_h(x)}{\partial x}\Big|_{\frac{\|z\|^2}{h^2}} \frac{V_{ijkl}\, z^i z^j z^k z^l}{h^2} \right) \left(f(0) + \frac{1}{2} \frac{\partial^2 f}{\partial z^i \partial z^j}\Big|_0 z^i z^j \right) dz \\
&= \int_{\mathbb{R}^m} \left(k\left(\|u\|^2\right) + h^2 \frac{\partial k(x)}{\partial x}\Big|_{\|u\|^2} V_{ijkl}\, u^i u^j u^k u^l \right) \left(f(0) + \frac{h^2}{2} \frac{\partial^2 f}{\partial z^i \partial z^j}\Big|_0 u^i u^j \right) du \\
&= C_1 f(0) - \frac{h^2}{2} C_2 f(0) V_{ijkl} \left[\delta^{ij}\delta^{kl} + \delta^{ik}\delta^{jl} + \delta^{il}\delta^{jk} \right] + \frac{h^2}{2} C_2 \sum_{i=1}^{m} \frac{\partial^2 f}{\partial (z^i)^2}\Big|_0 + O(h^4)
\end{aligned}$$

where the $O(h^4)$ term is finite due to the exponential decay of k and depends on k, $r_{\min}$, V_{ijkl} and $\|f\|_{C^3}$. Now we can upper bound the remaining error terms α_i, $i = 0, \ldots, 4$. For the argument of the kernel in α_1 and α_2 we have by our assumptions on $B(0, r_{\min})$:

$$\frac{\|z\|^2}{h^2} \geq \frac{\|z\|^2 + V_{ijkl} z^i z^j z^k z^l + \beta(z) \|z\|^5}{h^2} \geq \frac{\|z\|^2}{4h^2}.$$

Note that this inequality implies that β is uniformly bounded on $B(0, r_{\min})$ in terms of $r_{\min}$ and V_{ijkl}. Moreover for small enough h we have $\frac{r_{\min}}{h} \geq \sqrt{A}$ (see Assumptions 2.25 for the definition of A) so that we can use the exponential decay of k for α_3 and α_4.

$$|\alpha_0| \leq h^3 \|f\|_{C^3} \int_{B(0, \frac{r_{\min}}{h})} \frac{\partial k_h}{\partial x}\Big|_{\|u\|^2} |\beta(hu)| \|u\|^5 \, du$$

Since $\frac{\partial k_h}{\partial x}$ is bounded and has exponential decay, one has $|\alpha_0| \leq K_0 h^3$ where K_0 depends on k, $r_{\min}$ and $\|f\|_{C^3}$.

$$\begin{aligned}
|\alpha_1| &\leq \int_{B(0, r_{\min})} \left| \frac{\partial^2 k_h}{\partial x^2} \left(\frac{\|z\|^2 (1-\theta) + \theta\eta}{h^2} \right) \right| \frac{\left(V_{ijkl} z^i z^j z^k z^l + \beta(z) \|z\|^5 \right)^2}{h^4} f(z) dz \\
&\leq \|f\|_{C^3} h^4 \int_{B(0, \frac{r_{\min}}{h})} \left| \frac{\partial^2 k}{\partial x^2} \left(\|u\|^2 (1 - \theta(hu)) + \theta(hu)\eta(hu) \right) \right| \\
&\qquad \left(m^2 \max_{i,j,k,l} |V_{ijkl}| \|u\|^4 + h \|\beta\|_\infty \|u\|^5 \right)^2 du
\end{aligned}$$

First suppose $\frac{r_{\min}}{h} \leq 2\sqrt{A}$ then the integral is bounded since the integrands are bounded on $B(0, \frac{r_{\min}}{h})$. Now suppose $\frac{r_{\min}}{h} \geq 2\sqrt{A}$ and decompose $B(0, \frac{r_{\min}}{h})$ as $B(0, \frac{r_{\min}}{h}) = B(0, 2\sqrt{A}) \cup B(0, \frac{r_{\min}}{h}) \backslash B(0, 2\sqrt{A})$. On $B(0, 2\sqrt{A})$ the integral is finite since $\left| \frac{\partial^2 k}{\partial x^2} \right|$ is bounded and on the complement the integral is also finite since $\left| \frac{\partial^2 k}{\partial x^2} \right|$ has exponential decay since by assumption

$$\|u\|^2 (1 - \theta(hu)) + \theta(hu)\eta(hu) \geq \frac{1}{4} \|u\|^2 \geq A.$$

Therefore there exists a constant K_1 such that $|\alpha_1| \leq K_1\, h^4$.

$$
\begin{aligned}
|\alpha_2| &\leq \int_{B(0,r_{\min})} k_h\left(\frac{\|z\|^2}{4h^2}\right)\frac{1}{6}\frac{\partial^3 f}{\partial z^i \partial z^j \partial z^k}(\theta z) z^i z^j z^k dz \\
&\leq \frac{m^{3/2}}{6}\,\|f\|_{C^3}\, h^3 \int_{\mathbb{R}^m} k\left(\frac{\|u\|^2}{4}\right)\|u\|^3\, du \leq K_2\, h^3,
\end{aligned}
$$

$$
\begin{aligned}
|\alpha_3| &\leq \int_{\mathbb{R}^m\setminus B(0,r_{\min})} k_h\left(\frac{\|z\|^2}{h^2}\right)\left(f(0) + \left\langle \nabla f\big|_0, z\right\rangle + \frac{1}{2}\frac{\partial^2 f}{\partial z^i \partial z^j}\Big|_0 z^i z^j\right) dz \\
&\leq c\,\|f\|_{C^3}\int_{\mathbb{R}^m\setminus B(0,r_{\min})} \exp\left(-\frac{\alpha\,\|z\|^2}{h^2}\right)(1 + m\,h^2\,\|z\|^2) dz \\
&\leq c\exp\left(-\alpha\frac{r_{\min}^2}{2h^2}\right)\sqrt{\left(\frac{2\pi}{\alpha}\right)^m}(1 + m\,h^2\frac{m}{\alpha}),
\end{aligned}
$$

$$
\begin{aligned}
|\alpha_4| &\leq \|f\|_{C^3}\, K_4 \int_{\mathbb{R}^m\setminus B(0,r_{\min})}\left|\frac{\partial k_h}{\partial x}\left(\frac{\|z\|^2}{h^2}\right)\right|\frac{\|z\|^4}{h^2}(1+\|z\|^2) dz \\
&\leq c\,K_4\,h^2\,\|f\|_{C^3}\exp\left(-\alpha\frac{r_{\min}^2}{2h^2}\right)\int_{\mathbb{R}^m} e^{-\alpha\|u\|^2}\,\|u\|^4\,(1+h^2\,\|u\|^2) du,
\end{aligned}
$$

where K_4 in the error term α_4 is a constant depending on $\max_{i,j,k,l}|V_{ijkl}|$. Now one has[14]: $e^{-\frac{\xi^2}{h^2}} \leq \frac{h^s}{\xi^s}$ for $h \leq \xi/s$. In particular it holds $h^3 \geq \exp\left(-\alpha\frac{r_{\min}^2}{2h^2}\right)$ for $h \leq \frac{1}{3}\sqrt{\frac{\alpha}{2}}\, r_{\min}$, so that for $h < \min\{\frac{1}{3}\sqrt{\frac{\alpha}{2}}\, r_{\min}, \frac{r_{\min}}{\sqrt{A}}\} = h_0$ all error terms α_i, $i = 0,\ldots,4$ are smaller than a constant times h^3 where the constant depends on k, $r_{\min}$, V_{ijkl} and $\|f\|_{C^3}$. This finishes the proof. □

Now we are ready to prove Proposition 2.27,

Proof: Let $\epsilon = \frac{1}{3}\min\{\mathrm{inj}(x), \pi\rho\}$[15] where ϵ is positive by the assumptions on M. Then we decompose M as $M = B(x,\epsilon) \cup (M\setminus B(x,\epsilon))$ and integrate separately. The integral over $M\setminus B(x,\epsilon)$ can be upper bounded by using the definition of $\delta(x)$ (see Assumption 2.24) and the fact that k is non-increasing:

$$
\begin{aligned}
\int_M k_h\left(\|i(x)-i(y)\|^2_{\mathbb{R}^d}\right) f(y)p(y)\sqrt{\det g}\, dy &= \int_{B(x,\epsilon)} k_h\left(\|i(x)-i(y)\|^2_{\mathbb{R}^d}\right) f(y)p(y)\sqrt{\det g}\, dy \\
&\quad + \int_{M\setminus B(x,\epsilon)} k_h\left(\|i(x)-i(y)\|^2_{\mathbb{R}^d}\right) f(y)p(y)\sqrt{\det g}\, dy
\end{aligned}
\tag{2.14}
$$

Since k is non-increasing, we have the following inequality for the second term in (2.14):

$$
\int_{M\setminus B(x,\epsilon)} k_h\left(\|i(x)-i(y)\|^2_{\mathbb{R}^d}\right) f(y)p(y)\sqrt{\det g}\, dy \leq \frac{1}{h^m} k\left(\frac{\delta(x)^2}{h^2}\right)\|f\|_\infty
$$

[14] This inequality can be deduced from $e^x \geq x^n$ for all $x \geq 4n^2$.

[15] The factor 1/3 is needed in Theorem 2.35

Since $\delta(x)$ is positive by the assumptions on M and k decays exponentially, we can make the upper bound smaller than h^3 for small enough h. Now we deal with the integral over $B(x,\epsilon)$. Since ϵ is smaller than the injectivity radius $\mathrm{inj}(x)$, we can introduce normal coordinates $z = \exp^{-1}(y)$ with origin $0 = \exp^{-1}(x)$ on $B(x,\epsilon)$, so that we can write the integral over $B(x,\epsilon)$ using Proposition 2.19 as:

$$\int_{B(0,\epsilon)} k_h \left(\frac{\|z\|^2 - \frac{1}{12}\sum_{\alpha=1}^{d} \frac{\partial^2 i^\alpha}{\partial z^a \partial z^b} \frac{\partial^2 i^\alpha}{\partial z^u \partial z^v} z^a z^b z^u z^v + O(\|z\|^5)}{h^2} \right) p(z) f(z) \sqrt{\det g}\, dz$$

Using our assumptions, we see that $pf\sqrt{\det g}$ is in $C^3(B(0,\epsilon))$. Moreover, by Corollary 2.21 one has for $d_M(x,y) \leq \pi\rho$, $\frac{1}{2} d_M(x,y) \leq \|x-y\| \leq d_M(x,y)$. Therefore we can apply Lemma 2.29 and compute the integral in (2.3.2) which results in:

$$\begin{aligned} &\Big[p(0)f(0)\Big(C_1 + C_2 \frac{h^2}{24} \sum_{\alpha=1}^{d} \frac{\partial^2 i^\alpha}{\partial z^a \partial z^b} \frac{\partial^2 i^\alpha}{\partial z^c \partial z^d} \left[\delta^{ab}\delta^{cd} + \delta^{ac}\delta^{bd} + \delta^{ad}\delta^{bc} \right] \Big) \\ &+ C_2 \frac{h^2}{2} \Delta_M (pf\sqrt{\det g})\Big|_0 + O(h^3) \Big], \end{aligned} \tag{2.15}$$

where we have used that in normal coordinates z^i at 0 the Laplace-Beltrami operator Δ_M is given as $\Delta_M f\Big|_x = \sum_{i=1}^{m} \frac{\partial^2 f}{\partial (z^i)^2}\Big|_0$. The second term in the above equation can be evaluated using the Gauss equations, see [87, Proposition 6].

$$\begin{aligned} &\sum_{a,b=1}^{m} \sum_{\alpha=1}^{d} \frac{\partial^2 i^\alpha}{\partial z^a \partial z^b} \frac{\partial^2 i^\alpha}{\partial z^c \partial z^d} \left[\delta^{ab}\delta^{cd} + \delta^{ac}\delta^{bd} + \delta^{ad}\delta^{bc} \right] \\ =& \sum_{a,b=1}^{m} \sum_{\alpha=1}^{d} \left(\frac{\partial^2 i^\alpha}{\partial (z^a)^2} \frac{\partial^2 i^\alpha}{\partial (z^b)^2} + 2 \frac{\partial^2 i^\alpha}{\partial z^a \partial z^b} \frac{\partial^2 i^\alpha}{\partial z^a \partial z^b} \right) \\ =& 2 \sum_{a,b=1}^{m} \sum_{\alpha=1}^{d} \left(\frac{\partial^2 i^\alpha}{\partial z^a \partial z^b} \frac{\partial^2 i^\alpha}{\partial z^a \partial z^b} - \frac{\partial^2 i^\alpha}{\partial (z^a)^2} \frac{\partial^2 i^\alpha}{\partial (z^b)^2} \right) + 3 \sum_{a,b=1}^{m} \sum_{\alpha=1}^{d} \frac{\partial^2 i^\alpha}{\partial (z^a)^2} \frac{\partial^2 i^\alpha}{\partial (z^b)^2} \\ =& 2 \sum_{a,b=1}^{m} \langle \Pi(\partial_{z^a}, \partial_{z^b}), \Pi(\partial_{z^a}, \partial_{z^b}) \rangle - \langle \Pi(\partial_{z^a}, \partial_{z^a}), \Pi(\partial_{z^b}, \partial_{z^b}) \rangle \\ &+ 3 \left\| \sum_{a=1}^{m} \Pi(\partial_{z^a}, \partial_{z^a}) \right\|^2_{T_{i(x)}\mathbb{R}^d} = -2R + 3 \left\| \sum_{j=1}^{m} \Pi(\partial_{z^j}, \partial_{z^j}) \right\|^2_{T_{i(x)}\mathbb{R}^d}, \end{aligned}$$

where R is the scalar curvature. Plugging this result into (2.15) and using from Proposition 2.19, $\Delta_M \sqrt{\det g}\big|_0 = -\frac{1}{3}R$, the proof is finished. □

The following Lemma is an application of Bernstein's inequality. Together with the previous proposition it will be the main ingredient for proving consistency statements for the graph structure.

Lemma 2.30 *Let $X_1, \ldots, X_n$ be n i.i.d. random vectors in $\mathbb{R}^d$ with law P which is absolutely continuous with respect to the natural volume element dV of a submanifold $M \subset \mathbb{R}^d$ satisfying Assumption 2.24. Let p denote its density which is bounded, continuous, and positive $p(x) > 0$, for any $x \in M$. Furthermore, let k be a kernel*

with compact support on $[0, R_k^2]$ *satisfying Assumption 2.25. Let* $x \in M \backslash \partial M$ *and define* $b_1 = \|k\|_\infty \|f\|_\infty$, $b_2 = K \|f\|_\infty^2$ *where* K *is a constant depending on* $\|p\|_\infty$, $\|k\|_\infty$ *and* R_k. *Then for any bounded function* f,

$$\mathrm{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} k_h(\|i(x)-i(X_i)\|^2)f(X_i) - \mathbb{E}\, k_h(\|i(x)-i(Z)\|^2)f(Z)\Big| > \epsilon\Big) \leq 2\exp\left(-\frac{nh^m\varepsilon^2}{2b_2+2b_1\varepsilon/3}\right).$$

Proof: Since by assumption $\kappa > 0$, we have by Lemma 2.22 for any $x, y \in M$ with $\|x-y\| \leq \kappa/2$, $d_M(x,y) \leq 2\|i(x)-i(y)\|$. This implies $\forall a \leq \kappa/2$, $B_{\mathbb{R}^d}(x,a) \cap M \subset B_M(x,2a)$.
Let $W_i := k_h(\|i(x)-i(X_i)\|^2)f(X_i)$. We have

$$|W_i| \leq \frac{\|k\|_\infty}{h^m} \sup_{y \in B_{\mathbb{R}^d}(x,hR_k)\cap M} |f(y)| \leq \frac{\|k\|_\infty}{h^m}\|f\|_\infty := \frac{b_1}{h^m}.$$

For the variance of W we distinguish two cases. First let $hR_k < s := \min\{\kappa/2, R_0/2\}$ then we get

$$\begin{aligned}
\operatorname{Var} W &\leq \mathbb{E}_Z\, k_h^2(\|i(x)-i(Z)\|^2)f^2(Z) \leq \frac{\|k\|_\infty}{h^m}\mathbb{E}_Z\, k_h(\|i(x)-i(Z)\|^2)f^2(Z)\\
&= \frac{\|k\|_\infty}{h^m}\int_{B_{\mathbb{R}^d}(x,hR_k)\cap M} k_h(\|i(x)-i(y)\|^2)f^2(y)p(y)\sqrt{\det g}\,dy\\
&\leq \frac{\|k\|_\infty}{h^m}\int_{B_M(x,2hR_k)} k_h(\frac{1}{2}d_M(x,y)^2)f^2(y)p(y)\sqrt{\det g}\,dy\\
&\leq \frac{\|k\|_\infty^2}{h^{2m}}\|f\|_\infty^2\|p\|_\infty\int_{B_{\mathbb{R}^m}(0,2hR_k)}\sqrt{\det g}\,dz\\
&\leq \frac{\|k\|_\infty^2}{h^{2m}}\|f\|_\infty^2\|p\|_\infty\, S_2\, h^m R_k^m 2^m \leq C\frac{\|k\|_\infty^2}{h^m}\|p\|_\infty\|f\|_\infty^2,
\end{aligned}$$

where we have used Lemma 2.18 and C is given as $C = 2^m S_2 R_k^m$ Now consider $hR_k \geq s$, then

$$\operatorname{Var} W \leq \frac{\|k\|_\infty^2}{h^{2m}}\|f\|_\infty^2 \leq \frac{R_k^m\|k\|_\infty^2}{s^m h^m}\|f\|_\infty^2$$

Therefore we define $b_2 = K\|f\|_\infty^2$ with $K = \max\{C\|k\|_\infty^2\|p\|_\infty, \frac{R_k^m\|k\|_\infty^2}{s^m}\}$. By Bernstein's inequality we finally get

$$\mathrm{P}\Big(\big|\tfrac{1}{n}\textstyle\sum_{i=1}^n W_i - \mathbb{E}W\big| > \varepsilon\Big) \;\leq\; 2e^{-\frac{nh^m\varepsilon^2}{2b_2+2b_1\varepsilon/3}}$$

Both constants b_2 and b_1 are independent of x. □

Note that $\mathbb{E}_Z\, k_h(\|i(x)-i(Z)\|^2)f(Z) = \int\limits_M k_h(\|i(x)-i(y)\|^2)f(y)p(y)\sqrt{\det g}\,dy$.

2.3.3 Pointwise consistency of the degree function d or kernel density estimation on a submanifold in $\mathbb{R}^d$

In this section we will establish the asymptotic limit of the degree functions $d_{\lambda,h,n}$ corresponding to the weights $w_{\lambda,h}$. We will consider first the case $\lambda = 0$ which follows

easily from the results in the previous section. As a first step we extend the degree function from the graph to all points $x \in M$ using the kernel function by

$$d_{h,n}(x) = \frac{1}{n}\sum_{i=1}^{n} k_h(\|i(x) - i(X_i)\|).$$

It turns out that $d_{h,n}(x)$ is nothing else than a kernel density estimator on the submanifold M. For any fixed h it converges towards the h-averaged density

$$p_h(x) = \mathbb{E}_Z\, k_h(\|i(x) - i(Z)\|^2) = \int_M k_h(\|i(x) - i(y)\|^2) p(y)\sqrt{\det g}\, dy.$$

However in the following n and h are varied at the same time. This corresponds more to what is done in practice: as one gets more points, one shrinks the neighborhood.

Proposition 2.31 (Pointwise consistency of $d_{h,n}(x)$) *Suppose the standard assumptions hold. Furthermore let k be a kernel with compact support on $[0, R_k^2]$. Let $x \in M/\partial M$, then if $h \to 0$ and $nh^m \to \infty$,*

$$\lim_{n\to\infty} d_{h,n}(x) = C_1\, p(x) \quad \textit{in probability.}$$

If $nh^m/\log n \to \infty$, then almost sure convergence holds.

Proof: Using Proposition 2.27, we have for $x \in M/\partial M$ and $h \le h_0(x)$,

$$p_h(x) = C_1 p(x) + \frac{h^2}{4} C_2 \Big(p(x) \Big[-R + \frac{1}{2} \Big\| \sum_a \Pi(\partial_a, \partial_a) \Big\|^2_{T_{i(x)}\mathbb{R}^d} \Big] + 2(\Delta_M p)(x) \Big) + O(h^3).$$

so that $\lim_{h\to 0} p_h(x) = C_1 p(x)$. Using Lemma 2.30, we get,

$$\mathrm{P}(|d_{h,n}(x) - p_h(x)| > \epsilon) \le 2\exp\left(-\frac{nh^m\varepsilon^2}{2b_2 + 2b_1\varepsilon/3} \right),$$

where b_1 and b_2 are constants only depending on the kernel k. Therefore convergence in probability for $h \to 0$ follows under the condition $nh^m \to \infty$. Complete convergence follows if $\sum_{n=1}^{\infty} \mathrm{P}(|d_{h,n}(x) - p_h(x)| > \epsilon) < \infty$ which holds if $nh^m/\log n \to \infty$. Since complete convergence implies almost sure convergence the proof is complete. □

Another result on density estimation on submanifolds in $\mathbb{R}^d$ has been derived by Hendricks, Janssen and Ruymgaart in [47]. They construct a density estimator for compact submanifolds in $\mathbb{R}^d$ of the form

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}_{\|i(x)-i(X_i)\|\le\rho_n}}{\mathrm{vol}_M(B_{\mathbb{R}^d}(x,\rho_n)\cap M)},$$

using also the Euclidean distance for the balls. However they need the volume function of M for the normalization which is in general not available since we have no other information about M than the sample points X_i. In [47] then uniform strong convergence is shown where the conditions on the sequence ρ_n are very similar to

the conditions on h in Proposition 2.31. One could extend the strong pointwise convergence in our result to strong uniform convergence for a compact submanifold. However since our main concern is the more general structure of the continuum limit, we will not prove the more general result for $d_{h,n}(x)$. Related to our result is also the the recent work of Pelletier [72] on kernel density estimation for compact Riemannian manifolds without boundary. However apart from that we work in the more general setting of manifolds with bounded geometry, the main difference is that in [72] it is assumed that one knows the intrinsic distance function $d_M(x,y)$ on M. We cannot make such an assumption since we do not know the submanifold M beforehand. Instead we use the Euclidean distance of the ambient space $\mathbb{R}^d$. Using a normalization of the standard kernel density estimator in $\mathbb{R}^d$, Pelletier proves the appealing feature that his proposed kernel density estimator indeed produces a density even in the non-asymptotic regime. However the knowledge of this normalization factor requires full knowledge of the geometry of M which we do not have beforehand. In [72] then convergence in $L_2(M)$ is shown, whereas we show pointwise convergence.
We now give the limit of the general degree function $d_{\lambda,h,n}$ extended to all $x \in M$,

$$d_{\lambda,h,n}(x) = \frac{1}{n}\sum_{i=1}^{n} \tilde{k}_{\lambda,h}(x, X_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{k_h(\|i(x) - i(X_i)\|^2)}{[d_{h,n}(x) d_{h,n}(X_i)]^\lambda}.$$

In correspondence to the averaged density $p_h(x)$ where the convolution was done with the kernel k we can introduce a function $d_{\lambda,h}(x)$ where the convolution is done with the kernel $\tilde{k}$:

$$\tilde{k}(x,y) = \frac{k_h(\|i(x) - i(y)\|^2)}{[p_h(x) p_h(y)]^\lambda}.$$

This yields

$$\begin{aligned} d_{\lambda,h} &= \int_M \tilde{k}_{\lambda,h}(x,y) p(y) \sqrt{\det g}\, dy \\ &= \frac{1}{(p_h(x))^\lambda} \int_M k_h(\|i(x) - i(y)\|^2) \frac{p(y)}{(p_h(y))^\lambda} \sqrt{\det g}\, dy. \end{aligned}$$

The following lemma will be helpful in the rest of this chapter.

Lemma 2.32 *Let k have compact support on $[0, R_k^2]$ and let $0 < h \leq h_{\max}$. Then for any $x \in M$ there exist constants $D_1, D_2 > 0$ independent of h such that for any $y \in B_{\mathbb{R}^d}(x, hR_k) \cap M$ one has*

$$0 < D_1 \leq p_h(y) \leq D_2.$$

Proof: Suppose first that $hR_k < s := \min\{\kappa/2, R_0/2\}$. Since $\|y - z\| \leq hR_k \leq \kappa/2$ we have by Lemma 2.22: $\frac{1}{2} d_M(y,z) \leq \|y - z\| \leq d_M(y,z)$. Moreover since $p(x) > 0$ on M and p is bounded and continuous, there exist lower and upper bounds $p_{\min}$ and $p_{\max}$ on the density on $B_M(x, 4hR_k)$. That implies

$$p_h(y) \leq \frac{\|k\|_\infty}{h^m} p_{\max} \int_{B_M(y, 2hR_k)} \sqrt{\det g}\, dz \leq \|k\|_\infty\, p_{\max}\, S_2\, 2^m R_k^m,$$

where the last inequality follows from Lemma 2.18. Note further that $d_M(x,y) \leq 2hR_k$ and $d_M(y,z) \leq 2hR_k$ implies $d_M(x,z) \leq 4hR_k$. Using the assumption on the kernel function that $k(x) \geq \|k\|_\infty /2$ for $0 < x \leq r_k$, we get

$$\begin{aligned} p_h(y) &\geq \frac{\|k\|_\infty}{2h^m} \int_{B_{\mathbb{R}^d}(x,h\,r_k)\cap M} p(z)\sqrt{\det g}\,dz \geq \frac{\|k\|_\infty}{2h^m} p_{\min}\,\mathrm{vol}_M(B_M(x,h\,r_k)) \\ &\geq \frac{\|k\|_\infty}{2} p_{\min}\,S_1\,r_k^m. \end{aligned}$$

Now suppose $s \leq hR_k$ and $h \leq h_{\max}$. Then $p_h(y) \leq \frac{\|k\|_\infty}{h^m} \leq \|k\|_\infty \left(\frac{R_k}{s}\right)^m$. For the lower bound we get

$$\begin{aligned} p_h(y) &\geq \int_M k_h(d_M(y,z))p(z)\sqrt{\det g}\,dz \geq \int_{B_M(y,h\,r_k)} k_h(d_M(y,z))p(z)\sqrt{\det g}\,dz \\ &\geq \frac{\|k\|_\infty}{2h^m}\,\mathrm{P}\Big(B_M(y,h\,r_k)\Big) \geq \frac{\|k\|_\infty}{2h_{\max}^m}\,\mathrm{P}\Big(B_M(y,s\,\frac{r_k}{R_k})\Big) \end{aligned}$$

Since p is continuous and $p > 0$, the function $y \to \mathrm{P}\Big(B_M(y,s\,\frac{r_k}{R_k})\Big)$ is continuous and positive and therefore has a lower bound greater zero on the ball $B_{\mathbb{R}^d}(x,hR_k)\cap M$. □

Proposition 2.33 (Pointwise consistency of $d_{\lambda,h,n}$) *Suppose the standard assumptions hold. Furthermore let k be a kernel with compact support on $[0,R_k^2]$. Let $x \in M/\partial M$ and $\lambda \geq 0$. Then there exists a constant C such that for any $\frac{2\|k\|_\infty}{nh^m} < \epsilon < 1/C$, $0 < h < h_{\max}$ with probability at least $1 - C\,ne^{\frac{-nh^m\epsilon^2}{C}}$,*

$$|d_{\lambda,h,n}(x) - d_{\lambda,h}(x)| \leq \epsilon.$$

In particular if $h \to 0$ and $nh^m/\log n \to \infty$,

$$\lim_{n\to\infty} d_{h,\lambda,n}(x) = C_1^{1-2\lambda} p(x)^{1-2\lambda} \quad \textit{almost surely.}$$

Proof: The idea of this proof is to show that several empirical quantities which can be expressed as a sum of i.i.d. random variables are close to their expectation. Then one can deduce that also $d_{\lambda,h,n}$ will be close to $d_{\lambda,h}$. Consider the event $\mathcal{E}$ for which one has

$$\begin{cases} \text{for any } j \in \{1,\dots,n\},\ \Big|d_{h,n}(X_j) - p_h(X_j)\Big| \leq \varepsilon \\ \Big|d_{h,n}(x) - p_h(x)\Big| \leq \varepsilon \\ \Big|\frac{1}{n}\sum_{j=1}^n \frac{k_h(\|i(x)-i(X_j)\|^2)}{[p_h(X_j)]^{-\lambda}} - \int\limits_M k_h(\|i(x)-i(y)\|^2)\frac{p(y)}{[p_h(y)]^{-\lambda}}\sqrt{\det g}\,dy\Big| \leq \varepsilon \end{cases}$$

We will now prove that for sufficiently large C the event $\mathcal{E}$ holds with probability at least $1 - Cne^{-\frac{nh^m\varepsilon^2}{C}}$. For the second assertion defining $\mathcal{E}$, we use Lemma 2.30

$$\mathrm{P}(|d_{h,n}(x) - p_h(x)| > \epsilon) \leq 2\exp\left(-\frac{nh^m\varepsilon^2}{2b_2 + 2b_1\varepsilon/3}\right),$$

where b_1 and b_2 are constants depending on the kernel k and p. For the first term in the event $\mathcal{E}$ remember that $k(0) = 0$. We get for $\frac{\|k\|_\infty}{nh^m} < \varepsilon/2$ and $1 \le j \le n$,

$$\mathrm{P}\Big(\Big|\tfrac{1}{n}\sum_{i=1}^n k_h(\|i(X_j) - i(X_i)\|^2) - p_h(x)\Big| > \varepsilon \Big| X_j\Big) \le 2\exp\Big(-\tfrac{(n-1)h^m\varepsilon^2}{8b_2+4b_1\varepsilon/3}\Big).$$

This follows by

$$\begin{aligned}\Big|\frac{1}{n}\sum_{i=1}^n k_h(\|i(X_j) - i(X_i)\|^2) - p_h(X_j)\Big| \le & \Big|\frac{1}{n(n-1)}\sum_{i=1}^n k_h(\|i(X_j) - i(X_i)\|^2)\Big| \\ & + \Big|\frac{1}{n-1}\sum_{i\neq j} k_h(\|i(X_j) - i(X_i)\|^2) - p_h(X_j)\Big|\end{aligned}$$

where the first term is upper bounded by $\frac{\|k\|_\infty}{nh^m}$. First integrating wrt to the law of X_j (the right hand side of the bound is independent of X_j) and then using a union bound, we get

$$\mathrm{P}\Big(\text{for any } j \in \{1,\dots,n\},\ \Big|d_{h,n}(X_j) - p_h(X_j)\Big| \le \varepsilon\Big) > 1 - 2n\exp\Big(-\tfrac{(n-1)h^m\varepsilon^2}{8b_2+4b_1\varepsilon/3}\Big).$$

Noting that $\frac{1}{p_h(y)}$ is bounded for $y \in B_{\mathbb{R}^d}(x, hR_k) \cap M$ for $0 < h \le h_{\max}$ by Lemma 2.32, we get by Lemma 2.30 a Bernstein type bound for the probability of the third event in $\mathcal{E}$. Finally, combining all these results, we obtain that there exists a constant C such that for $\frac{2\|k\|_\infty}{nh^m} \le \varepsilon \le 1$, the event[16] $\mathcal{E}$ holds with probability at least $1 - Cne^{-\frac{nh^m\varepsilon^2}{C}}$. Let us define

$$\left\{\begin{array}{rcl}\mathcal{B} & := & \int_M k_h(\|i(x) - i(y)\|^2)[p_h(y)]^{-\lambda} p(y)\sqrt{\det g}dy \\ \hat{\mathcal{B}} & := & \frac{1}{n}\sum_{j=1}^n k_h(\|i(x) - i(X_j)\|^2)\big[d_{h,n}(X_j)\big]^{-\lambda}\end{array}\right.$$

then $d_{\lambda,h,n}(x) = \frac{\hat{\mathcal{B}}(x)}{d^\lambda_{h,n}(x)}$ and $d_{\lambda,h}(x) = \frac{\mathcal{B}(x)}{p^\lambda_h(x)}$. Let us now work only on the event $\mathcal{E}$. By Lemma 2.32 for any $y \in B_{\mathbb{R}^d}(x, hR_k) \cap M$ there exist constants D_1, D_2 such that $0 < D_1 \le p_h(y) \le D_2$. Using the first order Taylor formula of $[x \mapsto x^{-\lambda}]$, we obtain that for any $\lambda \ge 0$ and $a, b > \beta$, $\big|a^{-\lambda} - b^{-\lambda}\big| \le \lambda\beta^{-\lambda-1}|a-b|$. So we can write for $\varepsilon < D_1/2$,

$$\begin{array}{rcl}\big|\hat{\mathcal{B}} - \mathcal{B}\big| & \le & \Big|\frac{1}{n}\sum_{j=1}^n k_h(\|i(x) - i(X_j)\|^2)\Big(\big[d_{h,n}(X_j)\big]^{-\lambda} - [p_h(X_j)]^{-\lambda}\Big)\Big| \\ & & + \Big|\frac{1}{n}\sum_{j=1}^n k_h(\|i(x) - i(X_j)\|^2)[p_h(X_j)]^{-\lambda} - \mathcal{B}\Big| \\ & \le & \big|d_{h,n}(x)\big|\lambda\,(D_1 - \varepsilon)^{-\lambda-1}\varepsilon + \varepsilon \\ & \le & (D_2 + \varepsilon)\,\lambda\,(D_1 - \varepsilon)^{-\lambda-1}\varepsilon + \varepsilon := C'\varepsilon.\end{array}$$

Using that for $\lambda \ge 0$, $|a-b|^\lambda \le \lambda\max\{|a|,|b|\}^{\lambda-1}|a-b|$ we get $|(d_{h,n}(x))^\lambda - (p_h(x))^\lambda| \le C''\varepsilon.$ with $C'' = \lambda(D_2+\varepsilon)^{\lambda-1}$. Let $\zeta := \frac{1}{2}D_1^\lambda$. We have $p_h(x)^\lambda \ge 2\zeta$. Let us introduce further $\varepsilon_0 := \min\{\frac{\zeta}{C''}, 1\}$. For $\frac{2\|k\|_\infty}{nh^m} \le \varepsilon \le \varepsilon_0$, we have also $d^\lambda_{h,n}(x) \ge \zeta$. Combining the last three results, we obtain that there exists $C''' > 0$ such that

$$\Big|\frac{\hat{\mathcal{B}}}{d^\lambda_{h,n}(x)} - \frac{\mathcal{B}}{p^\lambda_h(x)}\Big| \le \frac{|\hat{\mathcal{B}} - \mathcal{B}|}{d^\lambda_{h,n}(x)} + \mathcal{B}\frac{|d^\lambda_{h,n}(x) - p^\lambda_h(x)|}{p^\lambda_h(x)d^\lambda_{h,n}(x)} \le \frac{C'\varepsilon}{\zeta} + \frac{D_2}{D_1^\lambda}\frac{C''\varepsilon}{2\zeta^2} \le C'''\varepsilon.$$

[16] The upper bound on ε is here not necessary but allows to write the bound more compactly.

We have proven that there exists a constant $C > 1$ such that for any $\frac{2\|k\|_\infty}{nh^m} < \varepsilon < 1/C$,

$$|d_{\lambda,h,n}(x) - d_{\lambda,h}(x)| \leq C'''\varepsilon,$$

with probability at least $1 - Cne^{-\frac{nh^m\varepsilon^2}{C}}$.
To prove the second assertion of the proposition, we employ again Proposition 2.27 for

$$\begin{aligned} d_{\lambda,h}(x) &= \int_M \tilde{k}_{\lambda,h}(x,y)p(y)\sqrt{\det g}\,dy \\ &= \frac{1}{p_h^\lambda(x)}\int_M k_h(\|i(x)-i(y)\|^2)\frac{p(y)}{p_h(y)^\lambda}\sqrt{\det g}\,dy. \end{aligned}$$

Due to the compact support of k we only have to control the expansion of p_h on the set $B_{\mathbb{R}^d}(x, 2hR_k)\cap M$. For sufficiently small h we have $\overline{B_{\mathbb{R}^d}(x,2hR_k)\cap M}\cap\partial M = \emptyset$. Moreover it can be directly seen from the proof of Proposition 2.27 that the upper bound of the interval $[0, h_0(y)]$ for which the expansion holds depends continuously on $\delta(x)$ and $\epsilon(y)$, where $\epsilon(y) = \frac{1}{3}\min\{\pi\rho, \mathrm{inj}(y)\}$. Now $h_0(x)$ is continuous since $\mathrm{inj}(x)$ is continuous on compact subsets, see [54][Prop. 2.1.10], and $\delta(x)$ is continuous since the injectivity radius is continuous. Therefore we conclude that since $h_0(y)$ is continuous on $\overline{B(x,2hR_k)\cap M}$ and $h_0(y) > 0$, $h_1(x) = \inf_{y\in\overline{B_{\mathbb{R}^d}(x,2hR_k)\cap M}} h_0(y) > 0$. Then for the interval $(0, h_1(x))$ the expansion of $p_h(y)$ holds uniformly over the whole set $B(x, 2hR_k)\cap M$. Using $\frac{1}{(a+h^2b)^\lambda} = \frac{1}{a^\lambda} - \lambda\frac{h^2b}{a^{\lambda+1}} + O(h^4)$, we get for $h \in (0, h_1(x))$,

$$\begin{aligned} d_{\lambda,h}(x) &= \frac{1}{p_h^\lambda(x)}\int_M k_h(\|i(x)-i(y)\|^2) \\ &\qquad \Big[\frac{C_1p(y) - \lambda/2h^2C_2(p(y)S + \Delta p)}{C_1^{\lambda+1}p(y)^\lambda} + O(h^3)\Big]\sqrt{\det g}\,dy \\ &= \frac{1}{p_h^\lambda(x)}\Big[C_1^{1-\lambda}p(x)^{1-\lambda} + \frac{C_2h^2}{2C_1^\lambda}\Big((1-\lambda)\,S\,p(x)^{1-\lambda} + \Delta p^{1-\lambda} - p^{-\lambda}\Delta p\Big) \\ &\qquad + O(h^3)\Big], \end{aligned}$$

where we have introduced the abbreviation $S = \frac{1}{2}[-R + \frac{1}{2}\|\sum_a \Pi(\partial_a,\partial_a)\|^2_{T_{i(x)}\mathbb{R}^d}]$. Now using again Proposition 2.27, we finally arrive at

$$\begin{aligned} d_{\lambda,h}(x) = (C_1\,p(x))^{1-2\lambda} + h^2\frac{C_2p(x)^{1-2\lambda}}{2C_1^{2\lambda}}\Big((1-2\lambda)\,S\, &- \lambda(1-\lambda)p(x)^{-2}\,\|\nabla p\|^2 \\ &+ (1-3\lambda)p(x)^{-1}\Delta p\Big) + O(h^3). \end{aligned}$$

From this asymptotic expression we derive $\lim_{h\to 0} d_{\lambda,h}(x) = (C_1\,p(x))^{1-2\lambda}$. Using the exponential inequality, one can derive almost sure convergence by the same argument as in Proposition 2.31. □

2.3.4 Pointwise consistency of the normalized and unnormalized graph Laplacian

In Section 2.1.4 we introduced the normalized graph Laplacian:

$$(\Delta_{\text{norm}} f)(i) = f(i) - \frac{1}{d(i)} \frac{1}{n} \sum_{j=1}^{n} w_{ij} f(j), \quad \Delta_{norm} f = (\mathbb{1} - D^{-1}W)f, \tag{2.16}$$

as well as the unnormalized graph Laplacian

$$(\Delta_{\text{unnorm}} f)(i) = d(i)f(i) - \frac{1}{n} \sum_{j=1}^{n} w_{ij} f(j), \quad \Delta_{unorm} f = (D - W)f. \tag{2.17}$$

Note that $\Delta_{unorm} = D\Delta_{norm}$. We extend the graph Laplacians to all $x \in M$ as it was done in the last section using the data-dependent kernel $k_{\lambda,h}$.

$$\begin{aligned}
(\Delta_{\lambda,h,n} f)(x) &= \frac{1}{h^2} \left(f - \frac{1}{d_{\lambda,h,n}} A_{\lambda,h,n} f \right)(x), \qquad \text{normalized} \\
&:= \frac{1}{h^2} \left(f(x) - \frac{1}{d_{\lambda,h,n}(x)} \frac{1}{n} \sum_{j=1}^{n} \tilde{k}_{\lambda,h}(x, X_j) f(X_j) \right) \\
(\Delta'_{\lambda,h,n} f)(x) &= \frac{1}{h^2} \Big(d_{\lambda,h,n} f - A_{\lambda,h,n} f \Big)(x), \qquad \text{unnormalized} \\
&:= \frac{1}{h^2} \left(d_{\lambda,h,n}(x) f(x) - \frac{1}{n} \sum_{j=1}^{n} \tilde{k}_{\lambda,h}(x, X_j) f(X_j) \right)
\end{aligned} \tag{2.18}$$

The factor $1/h^2$ arises by introducing a $1/h$-factor in the weights γ of the derivative operator d of the graph. The introduction of this factor is necessary since d approximates a derivative.

We would like to note that for the normalized graph Laplacian $\Delta_{\lambda,h,n}$ the normalization with $1/h^m$ in the weights cancels out, whereas it does not cancel for the unnormalized graph Laplacian $\Delta'_{\lambda,h,n}$ except in the case $\lambda = 1/2$. The problem here is that in general the intrinsic dimension m of the manifold is unknown. Therefore a normalization with the correct factor $\frac{1}{h^m}$ is not possible, and in the limit $h \to 0$ the estimate $\Delta'_{\lambda,h,n}$ will generally either vanish or blow up. The easy way to circumvent this is just to rescale the whole estimate such that $\frac{1}{n}\sum_{i=1}^{n} d_{\lambda,h,n}(X_i)$ equals a fixed constant for every n. The disadvantage is that this method of rescaling introduces a global factor in the limit. A more elegant way might be to estimate simultaneously the dimension m of the submanifold and use the dimension estimate to calculate the correct normalization factor. In Section 2.4.1 a scheme to estimate the intrinsic dimension of a submanifold from random samples is suggested. However in this work we assume for simplicity for the unnormalized graph Laplacian that the intrinsic dimension m of the submanifold is known. It might be interesting to consider both estimates simultaneously, but we leave this as an open problem.

The rest of this section is organized as follows. First we introduce the continuous operators $\Delta_{\lambda,h}$ resp. $\Delta'_{\lambda,h}$ corresponding to the extended graph Laplacians $\Delta_{\lambda,h,n}$ and $\Delta'_{\lambda,h,n}$. We then derive the limit of $\Delta_{\lambda,h}$ and $\Delta'_{\lambda,h}$ as $h \to 0$. Second we show that with high probability $\Delta_{\lambda,h,n}$ and $\Delta'_{\lambda,h,n}$ are close to $\Delta_{\lambda,h}$ resp. $\Delta'_{\lambda,h}$. Combining both results we finally arrive at the desired consistency results.

The following continuous approximation $\Delta_{\lambda,h}$ was similarly introduced by Coifman and Lafon [58, 25]. Continuous approximations of the unnormalized Laplacian $\Delta'_{\lambda,h}$ were considered by Belkin for the uniform probability measure on a compact manifold without boundary in his thesis [11].

Definition 2.34 (Kernel-based approximation of the Laplacian) *We introduce the following kernel-based averaging operator $A_{\lambda,h}$:*

$$(A_{\lambda,h}f)(x) = \int_M \tilde{k}_{\lambda,h}(x,y)f(y)p(y)\sqrt{\det g}\,dy, \tag{2.19}$$

and the following operator $\Delta_{\lambda,h}$ corresponding to the normalized graph Laplacian

$$\Delta_{\lambda,h}f := \frac{1}{h^2}\left(f - \frac{1}{d_{\lambda,h}}A_{\lambda,h}f\right),$$

and $\Delta'_{\lambda,h}$ corresponding to the unnormalized graph Laplacian

$$\Delta'_{\lambda,h}f := \frac{1}{h^2}\Big(d_{\lambda,h}f - A_{\lambda,h}f\Big) = d_{\lambda,h}\Delta_{\lambda,h}.$$

At least the definition of the normalized approximation $\Delta_{\lambda,h}$ can be justified by the alternative definition of the Laplacian in $\mathbb{R}^d$ sometimes made in physics textbooks:

$$(\Delta f)(x) = \lim_{r\to 0} -\frac{1}{C_d\, r^2}\left(f(x) - \frac{1}{\mathrm{vol}(B(x,r))}\int_{B(x,r)} f(y)dy\right),$$

where C_d is a constant depending on the dimension d.
Approximations of the Laplace-Beltrami operator based on averaging with the Gaussian kernel have been studied in the special case of the uniform measure on a compact submanifold without boundary by Smolyanov, Weizsäcker and Wittich in [86, 87] and Belkin [11]. Belkin's result was then generalized by Lafon [58] to general densities and to a wider class of isotropic, positive definite kernels for compact submanifolds with boundary. However the proof given in [58] is only correct for compact hypersurfaces[17] in $\mathbb{R}^d$, a proof for the general case of compact submanifolds with boundary using boundary conditions is announced in [25]. In this section we will prove the pointwise convergence of the continuous approximation for general submanifolds M with boundary of bounded geometry with the additional Assumptions 2.24. This includes the case where M is not compact. Moreover, no assumptions of positive definiteness of the kernel are made nor any boundary condition on the function f is imposed. Almost any submanifold occuring in practice should be covered in this very general setting.
For pointwise convergence boundary conditions on f are not necessary. However for uniform convergence there is no way around them. Then the problem lies not in the proof that the continuous approximation still converges in the right way but in the transfer of the boundary condition to the discrete graph. The main problem is that since we have no information about M apart from the random samples the boundary will be hard to locate. Moreover since the boundary is a set of measure zero, we will actually almost surely never sample any point from the boundary. The rigorous treatment of the approximation of the boundary respectively the boundary

[17] A hypersurface is a submanifold of codimension 1.

conditions of a function on a randomly sampled graph remains as an open problem. Especially for dimensionality reduction the case of low-dimensional submanifolds in $\mathbb{R}^d$ is important. Notably, the analysis below also includes the case where due to noise the data is only concentrated around a submanifold. Now we are ready to formulate the asymptotic result for the operator $\Delta_{\lambda,h}$ which extends the result of Lafon mentioned before.

Theorem 2.35 *Suppose the standard assumptions hold. Furthermore let k be a kernel with compact support on $[0, R_k^2]$. Let $\lambda \geq 0$, and $x \in M \backslash \partial M$, then there exists an $h_1(x) > 0$ such that for all $h < h_1(x)$ and any $f \in C^3(M)$,*

$$\begin{aligned}(\Delta_{\lambda,h} f)(x) =& -\frac{C_2}{2\,C_1}\left((\Delta_M f)(x) + \frac{s}{p(x)}\left\langle \nabla p, \nabla f\right\rangle_{T_x M}\right) + O(h)\\ =& -\frac{C_2}{2\,C_1}(\Delta_s f)(x) + O(h),\end{aligned} \tag{2.20}$$

where Δ_M is the Laplace-Beltrami operator of M and $s = 2(1-\lambda)$.

Proof: The need for compactness of the kernel k comes from the fact that the modified kernel $\tilde{k}$ depends on $p_h(y)$. Now for a non-compact manifold it is not possible to have a lower bound on $p(x)$. Therefore also an upper bound of the truncated version of the integral with respect to an exponentially decaying kernel function is not possible without additional assumptions on p.
Moreover the Taylor expansion of Proposition 2.27 for $p_h(y)$ can only be used for h in the interval $(0, h_0(y))$. Obviously it can happen that $h_0(y) \to 0$ when we approach the boundary. Therefore, when we have to control $h_0(y)$ over the whole space M, the infimum could be zero, so that the estimate holds for no h. As argued in the proof of Proposition 2.33 for sufficiently small h, $\overline{B_{\mathbb{R}^d}(x, 2hR_k) \cap M} \cap \partial M = \emptyset$. Then h_0 is continuous and positive on $\overline{B_{\mathbb{R}^d}(x, 2hR_k) \cap M}$ and therefore has a lower bound $h_1(x) = \inf_{y \in \overline{B_{\mathbb{R}^d}(x,2hR_k)\cap M}} h_0(y) > 0$. Then for the interval $(0, h_1(x))$ the asymptotic expansion of $p_h(y)$ holds uniformly over the set $B_{\mathbb{R}^d}(x, 2hR_k) \cap M$. That is, using the definition of $\tilde{k}$ as well as Proposition 2.27 and the expansion $\frac{1}{(a+h^2 b)^\lambda} = \frac{1}{a^\lambda} - \lambda \frac{h^2 b}{a^{\lambda+1}} + O(h^4)$, we get for $h \in (0, h_1(x))$ that

$$\begin{aligned}&\int_M \tilde{k}_{\lambda,h}\left(\|i(x) - i(y)\|^2\right) f(y) p(y) \sqrt{\det g}\, dy\\ =&\frac{1}{p_h^\lambda(x)} \int_{B_{\mathbb{R}^d}(x,hR_k)\cap M} k_h\left(\|i(x) - i(y)\|^2\right) f(y)\\ &\left[\frac{C_1 p(y) - \lambda/2 C_2 h^2 (p(y) S + \Delta p)}{C_1^{\lambda+1} p(y)^\lambda} + O(h^3)\right] \sqrt{\det g}\, dy,\end{aligned}$$

where the $O(h^3)$-term is continuous on $B_{\mathbb{R}^d}(x, hR_k)$ and we have introduced the abbreviation $S = \frac{1}{2}[-R + \frac{1}{2}\left\|\sum_a \Pi(\partial_a, \partial_a)\right\|^2_{T_{i(x)}\mathbb{R}^d}]$. Using $f(y) = 1$ we get,

$$\begin{aligned}d_{\lambda,h}(x) =&\frac{1}{p_h^\lambda(x)} \int_{B_{\mathbb{R}^d}(x,hR_k)\cap M} k_h\left(\|i(x) - i(y)\|^2\right)\\ &\left[\frac{C_1 p(y) - \lambda/2 C_2 h^2 (p(y) S + \Delta p)}{C_1^{\lambda+1} p(y)^\lambda} + O(h^3)\right] \sqrt{\det g}\, dy,\end{aligned}$$

as an estimate for $d_{\lambda,h}(x)$. Now using Proposition 2.27 again, we arrive at:

$$\begin{aligned}\Delta_{\lambda,h} f =& \frac{1}{h^2}\left(f - \frac{A_{\lambda,h}}{d_{\lambda,h}}\right) = \frac{1}{h^2}\frac{d_{\lambda,h} f - A_{\lambda,h} f}{d_{\lambda,h}} \\ =& -\frac{C_2}{2\,C_1}\left(\Delta_M f + \frac{2(1-\lambda)}{p}\langle \nabla p, \nabla f\rangle\right) + O(h),\end{aligned}$$

where all $O(h)$-terms are finite on $B_{\mathbb{R}^d}(x, hR_k) \cap M$ since p is strictly positive. □

Note that the limit of $\Delta_{\lambda,h}$ has the opposite sign of Δ_s. This is due to the fact that the Laplace-Beltrami operator on manifolds is usually defined as a negative definite operator (in analogy to the Laplace operator in $\mathbb{R}^d$), whereas the graph Laplacian is positive definite. But this varies through the literature, thus the reader should be aware of the sign convention. With the relation $(\Delta'_{\lambda,h,n} f)(x) = d_{\lambda,h,n}(x)(\Delta_{\lambda,h,n} f)(x)$ one can easily adapt the last lines of the previous proof to derive the following corollary.

Corollary 2.36 *Under the assumptions of Theorem 2.35. Let $\lambda \geq 0$ and $x \in M \backslash \partial M$. Then there exists an $h_1(x) > 0$ such that for all $h < h_1(x)$ and any $f \in C^3(M)$,*

$$(\Delta'_{\lambda,h} f)(x) = -p(x)^{1-2\lambda}\frac{C_2}{2C_1^{2\lambda}}(\Delta_s f)(x) + O(h), \quad \text{where} \quad s = 2(1-\lambda). \tag{2.21}$$

Before we state the results for the general case with data-dependent weights we now treat the case $\lambda = 0$, that is we have non-data-dependent weights. There the proof is considerably simpler and much easier to follow. Moreover opposite to the general case here we get convergence in probability under slightly weaker conditions. Belkin and Niyogi have proven independently in [10] the weak consistency of the unnormalized graph Laplacian for compact submanifolds with the uniform probability measure using the Gaussian kernel for the weights. However their convergence rate is $nh^{2m+4} \to \infty$ which is suboptimal compared to our rate $nh^{m+4} \to \infty$. The difference arises since they use Hoeffding's inequality instead of Bernstein's inequality which results in a suboptimal behaviour in the constants. We prove here both the limit of the unnormalized and normalized graph Laplacian for general submanifolds with boundary of bounded geometry, with general probability measures P, and general kernel functions k as stated in our standard assumptions.

Theorem 2.37 (Weak and strong pointwise consistency for $\lambda = 0$) *Suppose the standard assumptions hold. Furthermore let k be a kernel with compact support on $[0, R_k^2]$. Let $x \in M/\partial M$ and $f \in C^3(M)$. Then if $h \to 0$ and $nh^{m+4} \to \infty$,*

$$\lim_{n\to\infty}(\Delta_{0,h,n} f)(x) = -\frac{C_2}{2\,C_1}(\Delta_2 f)(x) \quad \textit{in probability,}$$

and

$$\lim_{n\to\infty}(\Delta'_{0,h,n} f)(x) = -p(x)\frac{C_2}{2\,C_1}(\Delta_2 f)(x) \quad \textit{in probability.}$$

If even $nh^{m+4}/\log n \to \infty$, then almost sure convergence holds.

Proof: We give the proof for $\Delta_{0,h,n}$ since the proof for $\Delta'_{0,h,n}$ can be directly derived using Proposition 2.33 and Lemma 2.30 for the variance term together with

Corollary 2.36 for the bias term. Similar to the proof for the Nadaraya-Watson regression estimate of Greblicki et al. in [39], we rewrite the estimator $\Delta_{0,h,n} f$ in the following form

$$(\Delta_{0,h,n} f)(x) = \frac{1}{h^2}\left[f(x) - \frac{(A_{0,h}\, f)(x) + B_{1n}}{1 + B_{2n}}\right], \tag{2.22}$$

where

$$\begin{aligned}
(A_{0,h}\, f)(x) &= \frac{\mathbb{E}_Z\, k_h(\|i(x) - i(Z)\|^2) f(Z)}{\mathbb{E}_Z\, k_h(\|i(x) - i(Z)\|^2)}, \\
B_{1n} &= \frac{\frac{1}{n}\sum_{j=1}^n k_h(\|i(x) - i(X_j)\|^2) f(X_j) - \mathbb{E}_Z\, k_h(\|i(x) - i(Z)\|^2) f(Z)}{\mathbb{E}_Z\, k_h(\|i(x) - i(Z)\|^2)}, \\
B_{2n} &= \frac{\frac{1}{n}\sum_{j=1}^n k_h(\|i(x) - i(X_j)\|^2) - \mathbb{E}_Z\, k_h(\|i(x) - i(Z)\|^2)}{\mathbb{E}_Z\, k_h(\|i(x) - i(Z)\|^2)}.
\end{aligned}$$

In Theorem 2.35 we have shown that

$$\lim_{h\to 0}(\Delta_{0,h} f)(x) = \lim_{h\to 0}\frac{1}{h^2}\left[f(x) - (A_{0,h}\, f)(x)\right] = -\frac{C_2}{2\, C_1}(\Delta_2 f)(x). \tag{2.23}$$

Using the lower bound of $p_h(x) = \mathbb{E}_Z\, k_h(\|i(x) - i(Z)\|^2)$ derived in Lemma 2.32, we can directly apply Lemma 2.30. There exist constants d_1 and d_2 such that

$$\mathrm{P}(|B_{1n}| \geq h^2 t) \leq \exp\left(-\frac{n h^{m+4}\, t^2}{2\,\|k\|_\infty\,(d_2 + t h^2\, d_1/3)}\right).$$

The same analysis can be done for B_{2n} where we do not have to deal with the $1/h^2$-factor. This shows convergence in probability. Complete convergence (which implies almost sure convergence) can be shown by proving for all $t > 0$ the convergence of the series $\sum_{n=0}^\infty \mathrm{P}\left(|B_{1n}| \geq h^2 t\right) < \infty$. A sufficient condition for that is $n h^{m+4}/\log n \to \infty$ as $n \to \infty$. □

Now we will show that with high probability the continuous approximations $\Delta_{\lambda,h}$ resp. $\Delta'_{\lambda,h}$ are pointwise close to the extended graph Laplacians $\Delta_{\lambda,h,n}$ resp. $\Delta'_{\lambda,h,n}$ when applied to a function $f \in C^3(M)$. The following proposition will be helpful.

Proposition 2.38 *Suppose the standard assumptions hold. Furthermore let k be a kernel with compact support on $[0, R_k^2]$. Let $x \in M/\partial M$, $f \in C^3(M)$ and $\lambda \geq 0$. Then there exists a constant C such that for any $\frac{2\|k\|_\infty}{n h^m} < \epsilon < 1/C$, $0 < h < h_{\max}$, with probability at least $1 - C\, n\, e^{\frac{-n h^m \epsilon^2}{C}}$,*

$$|(A_{\lambda,h,n} f)(x) - (A_{\lambda,h} f)(x)| \leq \epsilon.$$

Proof: Note that $d_{\lambda,h,n} = (A_{\lambda,h,n} 1)$. Using the boundedness of f, the proof of Proposition 2.33 can be adapted in a straightforward way. □

This leads us to our main theorem for the normalized graph Laplacian.

Theorem 2.39 (Pointwise consistency of $\Delta_{\lambda,h,n}$) *Suppose the standard assumptions hold. Furthermore let k be a kernel with compact support on $[0, R_k^2]$. Let $x \in$*

$M/\partial M$, $\lambda \geq 0$. Then there exists for any $f \in C^3(M)$ a constant C such that for any $\frac{2\|k\|_\infty}{nh^{m+4}} < \epsilon < 1/C$, $0 < h < h_{\max}$ with probability at least $1 - C\,n\,e^{\frac{-nh^{m+4}\epsilon^2}{C}}$,

$$|(\Delta_{\lambda,h,n}f)(x) - (\Delta_{\lambda,h}f)(x)| \leq \epsilon.$$

Define $s = 2(1-\lambda)$ then if $h \to 0$ and $nh^{m+4}/\log n \to \infty$,

$$\lim_{n\to\infty}(\Delta_{\lambda,h,n}f)(x) = -\frac{C_2}{2\,C_1}(\Delta_s f)(x) \quad \textit{almost surely.}$$

Proof: First we note that

$$\begin{aligned}\left|(\Delta_{\lambda,h,n}f)(x) - (\Delta_{\lambda,h}f)(x)\right| &= \frac{1}{h^2}\left|\frac{(A_{\lambda,h,n}f)(x)}{d_{\lambda,h,n}(x)} - \frac{(A_{\lambda,h}f)(x)}{d_{\lambda,h}(x)}\right| \\ \leq \frac{1}{h^2}\left(\frac{|(A_{\lambda,h,n}f)(x) - (A_{\lambda,h}f)(x)|}{d_{\lambda,h,n}(x)} + (A_{\lambda,h}f)(x)\frac{|d_{\lambda,h,n}(x) - d_{\lambda,h}(x)|}{d_{\lambda,h,n}(x)d_{\lambda,h}(x)}\right).\end{aligned}$$

Using Lemma 2.32 we get the following upper and lower bounds for $d_{\lambda,h}(y)$:

$$\frac{D_1}{D_2^{2\lambda}} \leq d_{\lambda,h}(y) \leq \frac{D_2}{D_1^{2\lambda}}, \quad \forall\, y \in B_{\mathbb{R}^d}(x, hR_k) \cap M.$$

Moreover $(A_{\lambda,h}f)(x) \leq \|f\|_\infty d_{\lambda,h}(x) \leq \|f\|_\infty \frac{D_2}{D_1^{2\lambda}}$. Now set $\zeta = \frac{D_1}{2\,D_2^{2\lambda}}$, then $d_{\lambda,h} \geq 2\zeta$. Let us introduce $\varepsilon_0 := \min\{\zeta, 1\}$. For $\frac{2\|k\|_\infty}{(n-1)h^m} \leq \varepsilon \leq \varepsilon_0$ we have by Proposition 2.33, $d_{\lambda,h,n}(x) \geq \zeta$, with probability $1 - C\,n\,e^{\frac{-nh^m\epsilon^2}{C}}$. Combining Proposition 2.33 and Proposition 2.38, we see that there exists a constant C such that with probability $1 - C\,n\,e^{\frac{-nh^m\epsilon^2}{C}}$:

$$\left|(\Delta_{\lambda,h,n}f)(x) - (\Delta_{\lambda,h}f)(x)\right| \leq \frac{1}{h^2}\left(\frac{\epsilon}{\zeta} + \frac{D_2\,\|f\|_\infty}{D_1^{2\lambda}}\frac{\epsilon}{2\zeta^2}\right) \leq C'\frac{\epsilon}{h^2}.$$

This proves the first statement of the theorem. By Theorem 2.35 we know further that $\lim_{h\to 0}(\Delta_{\lambda,h}f)(x) = -\frac{C_2}{2\,C_1}(\Delta_s f)(x)$ with $s = 2(1-\lambda)$. Combining this with the first statement and using the argument we made in Proposition 2.31, we are done. □

Since the unnormalized graph Laplacian $\Delta'_{\lambda,h,n}$ is just a multiplication operator times the normalized graph Laplacian, we can directly formulate the following pointwise consistency result for the unnormalized graph Laplacian.

Corollary 2.40 (Pointwise consistency of $\Delta'_{\lambda,h,n}$) *Suppose the standard assumptions hold. Furthermore let k be a kernel with compact support on $[0, R_k^2]$. Let $x \in M/\partial M$, $\lambda \geq 0$. Then there exists for any $f \in C^3(M)$ a constant C such that for any $\frac{2\|k\|_\infty}{nh^{m+4}} < \epsilon < 1/C$, $0 < h < h_{\max}$ with probability at least $1 - C\,n\,e^{\frac{-nh^{m+4}\epsilon^2}{C}}$,*

$$\left|(\Delta'_{\lambda,h,n}f)(x) - (\Delta'_{\lambda,h}f)(x)\right| \leq \epsilon.$$

Define $s = 2(1-\lambda)$ then if $h \to 0$ and $nh^{m+4}/\log n \to \infty$,

$$\lim_{n\to\infty}(\Delta'_{\lambda,h,n}f)(x) = -\frac{C_2}{2\,C_1^{2\lambda}}\,p(x)^{1-2\lambda}\,(\Delta_s f)(x) \quad \textit{almost surely.}$$

Proof: We have

$$\begin{aligned}&\left|(\Delta'_{\lambda,h,n}f)(x) - (\Delta'_{\lambda,h}f)(x)\right| \\ &\quad\leq \frac{1}{h^2}\left[\|f\|_\infty \left|d_{\lambda,h,n}(x) - d_{\lambda,h}(x)\right| + \left|(A_{\lambda,h,n}f)(x) - (A_{\lambda,h}f)(x)\right|.\right]\end{aligned}$$

Using Propositions 2.33 and 2.38, we get the first statement. The limit for $\Delta_{\lambda,h}$ for $h \to 0$ has been derived in Corollary 2.36. Again using the arguments for almost sure convergence provided in the proof of Proposition 2.31, we are done. □

This result is quite interesting. We observe that in the case of a uniform density it does not make a difference whether we use the unnormalized or the normalized approximation of the Laplacian. However, as soon as we have a non-uniform density, the unnormalized one will converge only up to a function to the Laplacian, except in the case $\lambda = \frac{1}{2}$ where both the normalized and unnormalized approximation lead to the same result.
On the other hand in semi-supervised learning we not only want to have an anisotropic diffusion process which prefers directions along high-density regions but a second desired property is that one would like to have a faster diffusion where one has high-density and slower diffusion where one has low-density. So that labels in high density regions have a larger influence than labels in low-density regions. We think that labels of rarely occuring events can be misleading whereas it seems more reasonable to follow the labels of often occuring events. The limit operator of the unnormalized graph Laplacian has this kind of non-uniform diffusion constant which for $\lambda < \frac{1}{2}$ leads to the desired effect that diffusion in high-density regions is faster than diffusion in low-density regions. If $\lambda > 1/2$ the direction is reversed. Then diffusion in low-density regions is faster than diffusion in high-density regions.
This shows that the choice of the graph Laplacian depends on what kind of problem one wants to solve. In our opinion therefore from a machine learning point of view there is no universal best choice between the normalized and unnormalized graph Laplacian. However from a mathematical point of view the normalized graph Laplacian has the correct pointwise limit to the weighted Laplace-Beltrami operator.

2.3.5 Weak consistency of $\mathcal{H}_V$ and the smoothness functional $S(f)$

So far we have dealt with the asymptotic limit of the degree function and the graph Laplacian. Many semi-supervised and supervised learning algorithms use a regularization scheme which means that the algorithm makes a trade off between fit of the function measured in a norm on the function space $\mathbb{R}^V$ versus smoothness of f measured by the smoothness functional $\Omega(f)$:

$$\min_{f\in\mathcal{H}_V} \|y - f\|^2_{\mathcal{H}_V} + \lambda\,\Omega(f), \quad \lambda > 0.$$

In Proposition 2.5 we derived that the norm in $\mathcal{H}_V$ and the smoothness functional $S(f) = \langle f, \Delta f\rangle_{\mathcal{H}_V}$ can be chosen independently from each other. From this perspective it would be interesting to consider both limits in their general form. However sacrificing clarity we do not prove the more general results (which can be derived rather straightforwardly) but instead stick to a simple case that is we consider as weight function χ in the inner product in $\mathcal{H}_V$, $\chi(d) = d$. By weak consistency we mean that we will consider the asymptotic limit of $\langle f, g\rangle_V$ of two functions f, g in

$C^3(M)$ discretized to the vertices V via the random sample. The results on the asymptotic behaviour of $\mathcal{H}_V$ should make it easier to design norms on $\mathbb{R}_V$ which have the desired properties for the learning problem at hand. In particular they should help to achieve the desired influence of the density p.

Remark: In this section we add to the standard assumptions the assumption that M is compact. That implies in particular global upper and lower bounds on $p(x)$, since $p(x) > 0$ and p is continuous which in turn implies global upper and lower bounds on p_h.

Proposition 2.41 *Suppose the standard assumptions hold. Furthermore let k be a kernel with compact support on $[0, R_k^2]$. We choose the weighting function in $\mathcal{H}(V, \chi)$ as $\chi(d(X_i)) = d(X_i)$. Let f, g be bounded functions on M, then as $h \to 0$ and $nh^m / \log n \to \infty$*

$$\lim_{n\to\infty} \langle f, g\rangle_{\mathcal{H}_V} = C_1^{1-2\lambda} \int_M f(x)\, g(x)\, p(x)^{2-2\lambda} \sqrt{\det g}\, dx \quad \text{almost surely.}$$

Proof: We note

$$\begin{aligned}\Big|\, \langle f, g\rangle_{\mathcal{H}_V} - \mathbb{E}_Z\, f(Z)g(Z)d_{\lambda,h}(Z)\Big| \le \Big|\frac{1}{n}\sum_{i=1}^n f(X_i)g(X_i)\big|d_{\lambda,h,n}(X_i) - d_{\lambda,h}(X_i)\big|\Big| \\ + \Big|\frac{1}{n}\sum_{i=1}^n f(X_i)g(X_i)d_{\lambda,h}(X_i) - \mathbb{E}_Z\, f(Z)g(Z)d_{\lambda,h}(Z)\Big|\end{aligned}$$

In Proposition 2.33 we derived for $\frac{\|k\|_\infty}{nh^m} \le \epsilon \le 1/C$ that with probability $1 - Cne^{-\frac{nh^m\epsilon^2}{C}}$,

$$\big|d_{\lambda,h,n}(x) - d_{\lambda,h}(x)\big| \le \epsilon.$$

Since M is compact, we have global upper and lower bounds on $p_h(x)$. Therefore it is not hard to see from the proof of Proposition 2.33 that in this case C can be chosen independent of x. Therefore a union bound is easily possible so that with probability $1 - Cn^2e^{-\frac{nh^m\epsilon^2}{C}}$,

$$\max_{i=1,\dots,n} \big|d_{\lambda,h,n}(X_i) - d_{\lambda,h}(X_i)\big| \le \epsilon.$$

Now for the second part we know that the function $f(x)g(x)d_{\lambda,h}(x)$ is bounded. Therefore we can apply Hoeffding's inequality

$$\mathrm{P}\left(\Big|\frac{1}{n}\sum_{i=1}^n f(X_i)g(X_i)d_{\lambda,h}(X_i) - \mathbb{E}_Z\, f(Z)g(Z)d_{\lambda,h}(Z)\Big| \ge \epsilon\right) \le 2e^{-\frac{n\epsilon^2}{2K}},$$

where $K = \|f\, g\, d_{\lambda,h}\|_\infty^2$. Putting both results together, we know that there exists a constant C such that with probability $1 - Cn^2e^{-\frac{nh^m\epsilon^2}{C}}$,

$$\Big|\, \langle f, g\rangle_{\mathcal{H}_V} - \mathbb{E}_Z\, f(Z)g(Z)d_{\lambda,h}(Z)\Big| \le \epsilon.$$

Now $d_{\lambda,h}$ is uniformly bounded since M is compact and converges pointwise

$$\lim_{h\to 0} d_{\lambda,h}(x) = (C_1 p(x))^{1-2\lambda}.$$

Therefore by the dominated convergence theorem it follows that

$$\lim_{h\to 0} \mathbb{E}_Z\, f(Z)g(Z)d_{\lambda,h}(Z) = C_1^{1-2\lambda} \int_M f(x)\, g(x)\, p(x)^{2-2\lambda} \sqrt{\det g}\, dx.$$

□

In Section 2.1.5 we introduced the smoothness functional $S(f)$ associated to the Laplacian Δ as $S(f) = \langle f, \Delta f\rangle_{\mathcal{H}_V}$. Explicitly it is given by

$$S(f) = \frac{1}{2n^2} \sum_{i,l=1}^{n} (f(l) - f(i))^2 \gamma(w_{il})^2 \phi(w_{il}). \tag{2.24}$$

It should be again stressed that the smoothness functional only depends on the product $\gamma^2\,\phi$. In particular, it is independent of χ. Now for both the unnormalized and normalized graph Laplacian the product $\gamma(w_{il})^2\phi(w_{il}) = w_{il}$ is the same. Therefore both graph Laplacians induce the same smoothness functional.
We will restrict ourselves also to this case, that is $\gamma(w_{il})^2\phi(w_{il}) = w_{il}/h^2$ (The factor $1/h$ comes from $\gamma(w_{il})$). If we want to use the normalized graph Laplacian $\Delta_{\lambda,h,n}$, this amounts to fixing χ in $H(V,\chi)$ as $\chi(d_i) = d_i$ and for the unnormalized graph Laplacian $\Delta'_{\lambda,h,n}$ this implies $\chi(d_i) = 1$. (Note that this does not contradict Proposition 2.5 since the graph Laplacian depends on the choice of $\mathcal{H}_V$). With this choice and $w(x,y) = \tilde{k}_{\lambda,h}(x,y)$ the smoothness functional $S(f)$ simplifies to

$$S_{\lambda,h,n}(f) = \frac{1}{2(nh)^2} \sum_{i,j=1}^{n} (f(X_j) - f(X_i))^2 \tilde{k}_{\lambda,h}(X_i, X_j).$$

There are two ways to derive the limit of $S_{\lambda,h,n}(f)$. Since we know the limits of $\Delta_{n,\lambda,h}f$ and $d_{\lambda,h,n}$ we could directly work with $\langle f, \Delta_{\lambda,h,n}f\rangle_{H_V}$. However the proof becomes a little bit lengthy. In particular the generalization to uniform convergence of $S_n(f)$ over a function class $\mathcal{F}$ which is very important in the theoretical analysis of learning algorithms is less transparent. Therefore we use here a more direct approach based on (one-sample) U-statistics, see the Appendix 2.5.1 for the definition. In the following proposition we prove the essential part of the consistency result which follows afterwards.

Proposition 2.42 *Suppose the standard assumptions hold. Furthermore let k be a kernel with compact support on $[0, R_k^2]$. Let $f \in C^3(M)$ and define*

$$\tilde{S}_{\lambda,h,n}(f) = \frac{1}{2n(n-1)\,h^2} \sum_{i,j=1}^{n} (f(X_i) - f(X_j))^2 \frac{k_h(\|i(X_i) - i(X_j)\|)}{[p_h(X_i)p_h(X_j)]^\lambda}.$$

Then there exist constants K_1, K_2 such that

$$P\left(|\tilde{S}_{\lambda,h,n}(f) - \mathbb{E}\,\tilde{S}_{\lambda,h,n}(f)| \geq \epsilon\right) \leq 2\, e^{-\frac{[n/2]h^m\epsilon^2}{2K_1 + 2/3K_2\epsilon}}.$$

In particular, if $h \to 0$ and $nh^m \to \infty$. Then

$$\lim_{n\to\infty} \tilde{S}_{\lambda,h,n}(f) = \frac{C_2}{2C_1^\lambda} \int_M \langle \nabla f, \nabla f\rangle_{T_xM}\, p(x)^{2-2\lambda} \sqrt{\det g}\, dx, \quad \text{in probability.}$$

If $nh^m/\log n \to \infty$ almost sure convergence holds.

Proof: We first derive an upper bound for $\tilde{S}_{\lambda,h,n}(f)$. By assumption $|f(x)-f(y)| \leq L(f)d_M(x,y)$ where $L(f)$ is the Lipschitz constant of f. We distinguish two cases. First let $hR_k < \kappa/2$ then by Lemma 2.22 $d_M(x,y) \leq 2\|x-y\| \leq 2hR_k$. In this case using $p_h \geq K > 0$ since M is compact, we get:

$$\Big|\frac{1}{2h^2}(f(X_i)-f(X_j))^2\frac{k_h(\|i(X_i)-i(X_j)\|)}{[p_h(X_i)p_h(X_j)]^\lambda}\Big| \leq \frac{2L(f)^2\,R_k^2\,\|k\|_\infty}{h^m K^{2\lambda}}.$$

Now let $hR_k \geq \kappa/2$ then

$$\Big|\frac{1}{2h^2}(f(X_i)-f(X_j))^2\frac{k_h(\|i(X_i)-i(X_j)\|)}{[p_h(X_i)p_h(X_j)]^\lambda}\Big| \leq \frac{8R_k^2\,\|f\|_\infty\,\|k\|_\infty}{h^m\kappa^2 K^{2\lambda}}.$$

One can derive in the same way bounds for $|\mathbb{E}\,\tilde{S}_{\lambda,h,n}|$ using Lemma 2.18. For the variance we also distinguish two cases. First let $hR_k \leq \min\{\kappa/2, R_0/2\}$, then $d_M(x,y) \leq 2hR_k$ resp. $d_M(x,y) \leq R_0$ so that we can apply Lemma 2.18:

$$\begin{aligned}\operatorname{Var}\tilde{S}_{\lambda,h,n} &= \frac{1}{4h^4}\int_M\int_M \big(f(x)-f(y)\big)^4\frac{k_h^2(\|i(x)-i(y)\|)}{[p_h(x)p_h(y)]^{2\lambda}}p(x)p(y)\,dV(x)\,dV(y)\\ &\leq \frac{4R_k^4L(f)^4\,\|k\|_\infty^2\,\|p\|_\infty}{K^{4\lambda}}\frac{1}{h^{2m}}\int_M 2^m S_2 h^m R_k^m p(x)\,dV(x)\\ &= \frac{4\,2^m\,S_2\,R_k^{m+4}L(f)^4\,\|k\|_\infty^2\,\|p\|_\infty}{K^{4\lambda}h^m}.\end{aligned}$$

The second case is easy since we have a lower bound on h. Having derived these upper bounds we deduce that there exist two constants K_1 and K_2 independent of h such that with Theorem 2.53,

$$\mathrm{P}\left(|\tilde{S}_{\lambda,h,n}(f)-\mathbb{E}\,\tilde{S}_{\lambda,h,n}(f)| > \epsilon\right) \leq 2\,e^{-\frac{[n/2]h^m\epsilon^2}{2K_1+2/3\epsilon K_2}}. \tag{2.25}$$

For the last statement in the theorem we first derive the limit of $\mathbb{E}\,\tilde{S}_{\lambda,h,n}$ as $h \to 0$. Define

$$B_{\lambda,h}(x) = \frac{1}{2h^2}\int_M (f(x)^2+f(y)^2-2f(x)f(y))k_h(\|i(x)-i(y)\|)\frac{p(y)}{(p_h(x)p_h(y))^\lambda}dV(y).$$

Note that $\mathbb{E}\,\tilde{S}_{\lambda,h,n} = \int_M B_{\lambda,h}(x)\,p(x)\,dV(x)$. Then with Proposition 2.27 for any $x \in M\backslash\partial M$ there exists $h_0(x) > 0$ such that for any $0 < h < h_0(x)$:

$$B_{\lambda,h}(x) = \frac{C_2}{2C_1^\lambda}\,\langle\nabla f,\nabla f\rangle_{T_xM}\,p(x)^{1-2\lambda} + O(h).$$

In particular we have pointwise convergence on $M\backslash\partial M$,

$$\lim_{h\to 0} B_{\lambda,h}(x) = \frac{C_2}{2C_1^\lambda}\,\langle\nabla f,\nabla f\rangle_{T_xM}\,p(x)^{1-2\lambda},$$

with $P(M\backslash\partial M) = 1$. Moreover $B_{\lambda,h}$ is bounded since for $hR_k \leq \min\{\kappa/2, R_0/2\}$:

$$|B_{\lambda,h}(x)| \leq \frac{2L(f)^2R_k^2\,\|k\|_\infty\,\|p\|_\infty}{K^{2\lambda}}2^m S_2 R_k^m$$

A similar bound for $hR_k \geq \min\{\kappa/2, R_0/2\}$ can be easily derived. This constant lies in $L_1(M, P)$ so that by the dominated convergence theorem we get

$$\lim_{h\to 0} \mathbb{E}\, \tilde{S}_{\lambda,h,n} = \lim_{h\to 0} \int_M B_{\lambda,h}(x)\, p(x)\, dV(x) = \frac{C_2}{2C_1^\lambda} \int_M \langle \nabla f, \nabla f \rangle_{T_x M}\, p(x)^{1-2\lambda}\, p(x)\, dV(x).$$

The rest of the theorem follows using the exponential inequality and the standard argument used in Proposition 2.31. □

Basically the last proposition was one part of the proof of the following theorem which states the strong consistency of the graph smoothness functional $S_{\lambda,h,n}(f)$.

Theorem 2.43 (Strong consistency of the smoothness functional $S_{\lambda,h,n}$) *Suppose the standard assumptions hold. Furthermore let k be a kernel with compact support on $[0, R_k^2]$ and let $f \in C^3(M)$ Then there exist constants $C, C' > 0$ such that for all $\frac{C'}{nh^m} < \epsilon < 1/C$ and $0 < h < h_{\max}$ with probability at least $1 - Cne^{-\frac{nh^m\epsilon^2}{C}}$,*

$$\left| S_{\lambda,h,n}(f) - \mathbb{E}\, \tilde{S}_{\lambda,h,n}(f) \right| \leq \epsilon.$$

In particular, if $h \to 0$ and $nh^m/\log n \to \infty$. Then

$$\lim_{n\to\infty} S_{\lambda,h,n}(f) = \frac{C_2}{2C_1^\lambda} \int_M \langle \nabla f, \nabla f \rangle_{T_x M}\, p(x)^{2-2\lambda} \sqrt{\det g}\, dx, \quad \text{almost surely.}$$

Proof: From Proposition 2.31 one can deduce there exists a constant C such that

$$\max_{1\leq i\leq n} |d_{h,n}(X_i) - p_h(X_i)| \leq \epsilon$$

holds with probability $1 - Cne^{-\frac{nh^m\epsilon^2}{C}}$. Let us define

$$U_{n,h} = \frac{2R_k^2}{n(n-1)} \sum_{i,j=1}^{n} k_h(\|i(x) - i(y)\|),$$

then one can deduce from Equation 2.25 in the proof of Proposition 2.42, that there exist constants K_1, K_2 independent of h such that with probability greater than $1 - 2\, e^{-\frac{[n/2]h^m\epsilon^2}{2K_1 + 2/3\epsilon K_2}}$,

$$|U_{n,h} - \mathbb{E}\, U_{n,h}| \leq \epsilon$$

Note that for $hR_k \leq \min\{\kappa/2, R_0/2\}$ we have

$$\mathbb{E}\, U_{n,h} \leq S_2\, 2^m\, R_k^m\, \|p\|_\infty\, \|k\|_\infty .$$

Since M is compact, we have $\forall x \in M$, $0 < D_1 \leq p_h(x) \leq D_2$. Working on the event where the union bound for $d_{h,n}$ and the bound for $U_{n,h}$ hold and using a Taylor expansion of $x \to x^{-\lambda}$ with

$$\beta = \min\{d_{h,n}(X_i)d_{h,n}(X_j), p_h(X_i)p_h(X_j)\}^{-\lambda-1} \leq (D_1 - \epsilon)^{-2(\lambda+1)},$$

we get for $\epsilon < D_1/2$,

$$\begin{aligned}
&\left| \frac{1}{(d_{h,n}(X_i)d_{h,n}(X_j))^\lambda} - \frac{1}{(p_h(X_i)p_h(X_j))^\lambda} \right| \\
&\leq \lambda\, \beta\, |d_{h,n}(X_i)d_{h,n}(X_j) - p_h(X_i)p_h(X_j)| \\
&\leq \lambda\, \beta \left(|d_{h,n}(X_i) - p_h(X_i)| d_{h,n}(X_j) + p_h(X_i) |d_{h,n}(X_j) - p_h(X_j)| \right) \\
&\leq \lambda\, \beta\, [(D_2 + \epsilon)\epsilon + D_2\epsilon] \leq C'\epsilon.
\end{aligned}$$

Note that C' is independent of X_i and X_j. For $hR_k \leq \min\{\kappa/2, R_0/2\}$ and $\epsilon \leq \frac{1}{2}\mathbb{E}\, U_{n,h}$ we get

$$\begin{aligned}|S_{\lambda,h,n}(f) - \tilde{S}_{\lambda,h,n}(f)| &\leq \Big|\frac{1}{2\,n^2(n-1)h^2}\sum_{i,j=1}^{n}(f(X_i)-f(X_j))^2\frac{k_h(\|i(X_i)-i(X_j)\|)}{[d_{h,n}(X_i)d_{h,n}(X_j)]^\lambda}\Big| \\ &\quad + C'\epsilon\Big|\frac{1}{2\,n(n-1)h^2}\sum_{i,j=1}^{n}(f(X_i)-f(X_j))^2 k_h(\|i(X_i)-i(X_j)\|)\Big| \\ &\leq \frac{4}{nh^m}\frac{4^\lambda\, R_k^2\, L(f)^2\, \|k\|_\infty}{D_1^{2\lambda}} + C' L(f)^2\, S_2\, 2^m\, R_k^{2+m}\, \|p\|_\infty\, \|k\|_\infty\, \epsilon\end{aligned}$$

In the same way as shown in the proof of Proposition 2.42 a similar bound can be derived for $hR_k > \min\{\kappa/2, R_0/2\}$. Now applying Proposition 2.42 we are done. □

This proof applies only to one function. However uniform convergence for a suitable set of bounded functions with finite L_∞-covering numbers (this requires a certain smoothness of the considered set of functions) can either proven by a standard covering number approach, or one uses the more sophisticated uniform convergence bounds for U-statistics developed by Nolan and Pollard in [69].

2.3.6 Summary and fixation of $\mathcal{H}_V$ by mutual consistency requirement

When we introduced graph Laplacians for undirected graphs in Section 2.1.4, we noted that for both the normalized and unnormalized graph Laplacians there exists a whole family of choices of the graph structure $\mathcal{H}_V, \mathcal{H}_E$ and d which yield the same normalized resp. unnormalized graph Laplacian. We will show in this section that one can at least fix the weighting function χ in $\mathcal{H}_V$ (which then in turn implies that also the product $\gamma(w_{ij})^2\phi(w_{ij})$ is fixed). However fixing also the weights ϕ in $\mathcal{H}_E$ and γ in the difference operator d apart from the assumptions we imposed in 2.1.4, remains an open problem.

In Table 2.1 the results of the previous sections are summarized. Now we fix $\mathcal{H}_V$ for both the normalized and the unnormalized graph Laplacians by requiring mutual consistency of the limits. By construction every graph Laplacian $\Delta = d^*d$ is a self-adjoint operator in $\mathcal{H}(V,\chi)$. Now as we will show the limit operators of the normalized and unnormalized graph Laplacian are symmetric in some weighted L_2-space on M. Mutual consistency means now that we require that also $\langle f, g\rangle_{\mathcal{H}(V,\chi)}$ converges towards the corresponding weighted L_2-space.

Let us first consider the normalized graph Laplacian. In Theorem 2.39 it was shown (up to constants) that $\lim_{n\to\infty}(\Delta_{\lambda,h,n}f)(x) \sim (\Delta_s f)(x)$ almost surely as $h \to 0$ and $nh^{m+4}/\log n \to \infty$ with $s = 2 - 2\lambda$. Now in Equation 2.11 it was shown that Δ_s is symmetric on a dense subspace of $L_2(M, p^s\, dV)$. That yields the following scheme:

$\Delta_{\lambda,h,n}$	self-adjoint in	$\mathcal{H}(V,\chi)$
$\Downarrow$		$\Downarrow$?
Δ_s	symmetric in	$L_2(M, p^s\, dV)$

In order to satisfy mutual consistency, $\mathcal{H}(V,\chi)$ has to be asymptotically equal to $L_2(M, p^s\, dV)$ or said in another way: the limit operator Δ_s should be symmetric

Tabelle 2.1: Results of the previous sections. All statements hold (up to constants) pointwise for $h \to 0$ and $nh^{m+4}/\log n \to \infty$ (however some results require less restrictive conditions). On the left hand the graph objects are listed and on the right hand side their corresponding limit on the manifold for $n \to \infty$.

Graph	Submanifold M in $\mathbb{R}^d$
$d_{n,h}$	$p(x)$
$d_{\lambda,n,h}$	$(p(x))^{1-2\lambda}$
$\Delta_{\text{norm}} f$	$\Delta_s f,\ s = 2 - 2\lambda$
$\Delta_{\text{unnorm}} f$	$(p(x))^{1-2\lambda}\Delta_s f,\ s = 2 - 2\lambda$
$\langle f, g\rangle_{\mathcal{H}(V,d_{\lambda,h,n})}$	$\langle f, g\rangle_{\mathcal{H}}$ with $\mathcal{H} = L_2(M, p^s),\ s = 2 - 2\lambda$
$S(f) = \langle f, \Delta f\rangle_{\mathcal{H}_V}$	$\langle \nabla f, \nabla f\rangle_{\mathcal{H}}$ with $\mathcal{H} = L_2(M, p^s),\ s = 2 - 2\lambda$

with respect to the limit inner product of $\mathcal{H}_V$. We know by Proposition 2.41 that under the conditions stated there for $\chi(d_{\lambda,h,n}) = d_{\lambda,h,n}$ we have up to constants $\lim_{n\to\infty} \langle f, g\rangle_{\mathcal{H}_V} \sim \langle f, g\rangle_{L_2(M,p^s\,dV)}$ with $s = 2(1-\lambda)$. Since no other weight function $\chi(d_{\lambda,h,n})$ would yield the same limit, we conclude that for the specific weights on the graph we have chosen there exists no other choice of χ such that the inner product converges in this way[18]. Therefore the only choice of χ to fulfill the mutual consistency requirement is $\chi(d) = d$. Now the question is if we can also fix the weights ϕ in $\mathcal{H}(E, \phi)$ and the weights γ of d. Using the general form of the graph Laplacian for undirected graphs in Equation 2.6 and comparing it with that for the normalized one in Equation 2.7, we get only the condition:

$$\frac{1}{n}\sum_{i=1}^{n} \gamma(w_{ji})^2\phi(w_{ji}) = d_j \quad \longrightarrow \quad \gamma(w_{ji})^2\phi(w_{ji}) = w_{ji}. \tag{2.26}$$

Unfortunately this does neither fix γ nor ϕ. A possible solution might lie in studying also the limit of $\mathcal{H}(E, \phi)$. However at the moment we have little intuition about how this limit should look like. In particular the interpretation of functions in $\mathcal{H}(E, \phi)$ as discrete flows seems hard to transfer to the continuous setting.
Now the same analysis can be done for the unnormalized graph Laplacian $\Delta_{\lambda,h,n}$. For bounded functions f, g we have by the strong law of large numbers

$$\lim_{n\to\infty} \langle f, g\rangle_{\mathcal{H}(V,1)} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(X_i)g(X_i) = \int_M f(x)g(x)p(x)\,dV(x), \quad \text{a. s.}$$

Similar to the normalized case we get the following scheme:

[18] However this does not imply that there exists no other more general inner product on $\mathcal{H}_V$ which would have the same limit.

$$\begin{array}{ccc} \Delta'_{\lambda,h,n} & \text{self-adjoint in} & \mathcal{H}(V,\chi) \\ \Downarrow & & \Downarrow ? \\ p(x)^{1-2\lambda}\Delta_s & \text{symmetric in} & L_2(M, p\,dV) \end{array}$$

The only way how this convergence can happen (for all $\lambda \geq 0$) is $\chi(d) = 1$. Comparing the general form of the graph Laplacian for undirected graphs in Equation 2.6 with that for the normalized one in Equation 2.8, we get the same condition as in Equation 2.26 for the weights ϕ in $\mathcal{H}(E, \phi)$ and γ in the difference operator d.

2.4 Applications

2.4.1 Intrinsic dimensionality estimation of submanifolds in $\mathbb{R}^d$

The topic of intrinsic dimensionality estimation of submanifolds in $\mathbb{R}^d$ has a long history. In this work we consider the case where we have random samples from a probability distribution which has support on a submanifold in $\mathbb{R}^d$. In recent years there has been done a lot of work in estimating manifold structure from the data. However finding low-dimensional approximations of submanifolds is considerably harder than estimating their dimension, and the goal of what kind of the structure of the manifold should be preserved in the approximation differs from method to method.
The estimation of the intrinsic dimensionality is interesting in machine learning out of the following reasons:

- The intrinsic dimensionality is equal to the degrees of freedom or free parameters of the dataset which is an important qualitative description of the data,
- for some dimensionality reduction method one needs to know the intrinsic dimension, see also Section 2.3.4,
- one can use the intrinsic dimension as a feature for prediction (time series analysis),
- high intrinsic dimension of a dataset can explain bad learning performance (however high intrinsic dimension does not necessarily imply bad learning performance).

The goal of estimating the dimension of a submanifold is a well-defined mathematical problem. Let us first give a short description of different possibilities to define dimension. The most general one is the topological dimension:

Definition 2.44 (Topological dimension) *A topological space X has topological dimension m if every open covering C has an open refinement [19] C' in which every point occurs in at most $m+1$ sets in C'. m is the smallest such integer.*

The notion of Hausdorff dimension can deal with sets of non-integer dimension and is therefore suited for fractal geometry, see [32].

[19] A refinement C' of a covering C is a cover of X such that for every $S' \in C'$ there exists $S \in C$ such that $S' \subset S$.

Definition 2.45 (Hausdorff dimension) *Let F be a subset of $\mathbb{R}^d$, $s \in \mathbb{R}$, $s \geq 0$. For any $\delta > 0$ define*

$$H^s_\delta(F) := \inf\{\sum_{i\in I} \operatorname{diam}(U_i)^s \mid \{U_i\}_{i\in I} \text{ is a countable } \delta\text{-cover of } F\},$$

where $\operatorname{diam}(U) = \sup_{x,y\in U} \|x - y\|$. *Define* $H^s(F) = \lim_{\delta\to 0} H^s_\delta(F)$. *Then:*

$$\dim_{\text{Hausdorff}} F = \inf\{s \geq 0 \,|\, H^s(F) = 0\} = \sup\{s \,|\, H^s(F) = \infty\}.$$

The topological as well as the Hausdorff dimension agree for submanifolds in $\mathbb{R}^d$. Differences arise only if one considers more irregular sets like fractals, see [32]. The disadvantage of both is that they are very hard to compute.
The methods for dimensionality estimation up to now developed in statistics, statistical physics and machine learning can be roughly divided into two groups. The first one pioneered by Fukunaga [35] tries to determine the dimensionality by dividing the data in small subregions followed by a principal component analysis (PCA) of the points in each subregion. The averaged number of dominant eigenvalues (over the subregions) determines then the dimension, see [35]. This method has two drawbacks. First, one has to find a suitable scale for the size of the subregions. A too small scale will lead to a systematic underestimation of the dimension and a too large scale will lead to an overestimation due to the curvature of the submanifold. Second, one has to determine what one considers as dominant eigenvalues which is also a typical problem of standard PCA. The second type of estimators was originally designed in statistical physics to determine the dimension of the attractor of a chaotic dynamical system from samples of its time series, see [96] for a nice review. They are all based on the assumption that the volume of an m-dimensional set scales with its size s as s^m which implies that also the number of neighbors less than s apart will behave in the same way. This was the motivation for Grassberger and Procaccia in [38] to define the following notion of dimensionality which is easy to compute from given data.

Definition 2.46 (Correlation dimension) *Let X_i, $i = 1, \ldots, n$ be points drawn i.i.d. from a probability measure P where P has support on a submanifold in $\mathbb{R}^d$. Then define the correlation integral as*

$$C_n(s) = \frac{2}{n(n-1)} \sum_{i<j}^{n} \mathbb{1}_{\|X_i - X_j\| \leq s}. \tag{2.27}$$

The correlation dimension $\dim \mathrm{Corr}$ *is defined as*

$$\dim \mathrm{Corr} = \lim_{s\to 0} \lim_{n\to\infty} \frac{\log C_n(s)}{\log s}.$$

In practice one computes $C_n(s)$ for different s_i and then fits a line through the set of points $[\log s_i, \log C_n(s_i)]$ with least squares. Similar to the method of Fukunaga also for the correlation dimension one has the drawback that one has to choose the scales s_i. Note that this is a crucial step since the (discrete) data always looks 0-dimensional at a very small scale and is maybe even d-dimensional at a large scale so that one either under- or overestimates the dimension. In our analysis we will take the limits $s \to 0$ and $n \to \infty$ simultaneously since this corresponds also to what one

does in practice: as one gets more sample points one examines the data at a smaller scale.
The quantity we estimate is essentially the correlation integral with $\mathbb{1}$ replaced by a general kernel function. However the way we estimate the dimension is based on the convergence rate of the modified correlation integral. The advantage is that we only have to choose once a kind of 'smallest' scale at which one examines the data, the others are then determined by the convergence rate. Also we examine to our knowledge for the first time the influence of using in the correlation integral the distance in $\mathbb{R}^d$ instead of the intrinsic distance of the manifold. The asymptotic analysis of the modified correlation integral shows how the intrinsic and extrinsic curvature of the submanifold as well as the smoothness of the density of the probability measure influence the asymptotics of the correlation integral. Both effects lead to a scaling of $C_n(s)$ which is different from s^m.
This section has been partially published in [41].

Theoretical Background

We assume that the probability measure P generating the data $X_i \in \mathbb{R}^d$ has support on a m-dimensional submanifold M of $\mathbb{R}^d$. This means we are not trying to separate possible noise in the data from the underlying ground truth. In fact we will argue later in an experiment that on the basis of a finite sample it is in principle impossible to judge whether one has noise in the data or a very curved manifold. Moreover we also exclude the case of probability distributions with support of fractal dimension. As in the case of noise it is in principle impossible to judge based on a finite sample whether the data has fractal dimension or just very high curvature.
The m-dimensional submanifold M is a Riemannian manifold if one considers the induced metric from $\mathbb{R}^d$. That means that the inclusion map $i : M \to \mathbb{R}^d$ is an isometry (in the sense of Riemannian manifolds). Note that we will again use in the following the somehow cumbersome notation $x \in M$ and $i(x) \in \mathbb{R}^d$ in order to make it more obvious when we are working on the manifold M and when on $\mathbb{R}^d$. As any Riemannian manifold, M is also a metric space with the path-metric. A key point in the following proof will be the relation of the distance $d(x, y)$ on M and the Euclidean distance $\|i(x) - i(y)\|$ in $\mathbb{R}^d$ of two points $x, y \in M$ given in Proposition 2.19.
Also in this section we will work in the general setting of non-compact Riemannian submanifolds of $\mathbb{R}^d$ with boundary and of bounded geometry and in general we assume that the standard assumptions given in Section 2.3.1 hold. We impose an additional condition on the density $p(x)$ of the data-generating measure P:

$$\int_M p^2(x) dV(x) < \infty.$$

Again we define

$$k_h(\|i(x) - i(y)\|^2) = \frac{1}{h^m} k(\|i(x) - i(y)\|^2 / h^2).$$

Now that we have stated the setting we are working in we can introduce our estimator. We denote by X the i.i.d. sample $X_i, i = 1, \ldots, n$ of size n drawn from P.

Then the U-statistic we use is defined as

$$U_{n,h}(k) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} k_h(\|i(X_i) - i(X_j)\|^2).$$

For a moment we assume that we know the correct dimension m of M since the neighborhood parameter h is taken to the power of m. This factor of $1/h^m$ is then also the only difference (except that we use a "nicer" kernel function) to the standard correlation dimension estimate in Equation (2.27). This apparently small change allows us to take the limits as $n \to \infty$ and $h \to 0$ simultaneously which we think is the more natural setting and also corresponds more to what people actually do in practice. The actual estimator will be based on the behavior of $U_{n,h}(k)$ as one uses another integer instead of the correct dimension m in the factor $1/h^m$.
The expectation of $U_{n,h}(k)$ is given as

$$\mathbb{E}\, U_{n,h}(k) = \mathbb{E}\, k_h(\|i(X) - i(Y)\|^2) = \int_M \int_M k_h(\|i(x) - i(y)\|^2) p(x) p(y)\, dV(x)\, dV(y).$$

The central point is how this U-statistic behaves as $n \to \infty$ and $h \to 0$. At first we study how the expectation behaves as $h \to 0$. For this purpose we use a modification of Proposition 2.27.

Proposition 2.47 *Let* $S_\epsilon = \{x \in M, d(x, \partial M) < \epsilon\}$ *and*

$$f_h(x) = \int_M k_h\left(\|i(x) - i(y)\|^2_{\mathbb{R}^d}\right) p(y)\, dV(y).$$

Assume furthermore that $\delta_\epsilon := \inf_{x \in M \setminus S_\epsilon} \delta(x) > 0$. *Then there exists an* $h_0 > 0$, *such that for all* $h < h_0$,

$$\int_{M \setminus S_\epsilon} \left| f_h(x) - \left(C_1 p(x) + C_2 \frac{h^2}{2}\left[S(x)p(x) + (\Delta_M\, p)(x)\right]\right)\right| p(x)\, dV(x)$$
$$\le h^3 \int_{M \setminus S_\epsilon} \Gamma(x)\, p(x)\, dV(x),$$

with $S(x) = \frac{1}{2}\left[-R + \frac{1}{2}\left\|\sum_i \Pi(\partial_i, \partial_i)\right\|^2 \right]$ *and where* $\Gamma(x)$ *is a bounded function on* $M \setminus S_\epsilon$.

Proof: The expansion of $f_h(x)$ is given in Proposition 2.27, where $h_0(x)$ depends on $\mathrm{inj}(x)$ and $\delta(x)$. Now the injectivity radius on $M \setminus S_\epsilon$ is lower bounded due to our assumption of bounded geometry and $\delta(x)$ is by assumption lower bounded by δ_ϵ on $M \setminus S_\epsilon$. Therefore there exists an $h_0 = \inf_{x \in M \setminus S_\epsilon} h_0(x)$ such that the expansion holds on all of $M \setminus S_\epsilon$ for $h < h_0$. The function $\Gamma(x)$ is bounded since it depends only on the bounded quantities. □

This proposition shows that $U_{n,h}(k)$ has only asymptotically the expected scaling behavior. There is a second order correction induced by the curvature of M and the possibly non-uniform probability measure P. We would like to notice that the curvature corrections would vanish if one would use the intrinsic distance of M instead of the Euclidean distance in $U_{n,h}(k)$.

Proposition 2.48 *Under the stated assumptions on M, P and k,*

$$\lim_{h\to 0} \mathbb{E}\, U_{n,h}(k) = C_1 \int_M p(x)^2 dV(x).$$

Proof: Let $f_h(x) = \int_M k_h(\|i(x) - i(y)\|^2) p(y)\, dV(y)$. Then by Proposition 2.27, $\lim_{h\to 0} f_h(x) = C_1 p(x)$. Moreover one can show that there exists a constant C such that $|f_h(x)| \le C$. Namely by our assumptions on M for $h \le \min\{\kappa/2, R_0/2\}$:

$$\begin{aligned} |f_h(x)| &\le \frac{\|k\|_\infty \, \|p\|_\infty}{h^m} \int_M \mathbb{1}_{\|i(x)-i(y)\|\le hR_k}\, dV(y) \\ &\le \frac{\|k\|_\infty \, \|p\|_\infty}{h^m} \, \mathrm{vol}(B_M(x, 2hR_k)) \le \|k\|_\infty \, \|p\|_\infty \, 2^m \, S_2 \, R_k^m, \end{aligned}$$

where we have used Lemma 2.22 followed by Lemma 2.18. The upper bound for $h > \min\{\kappa/2, R_0/2\}$ is given by $|f_h| \le \frac{\|k\|_\infty}{h^m}$. Now both upper bounds for $|f_h|$ are bounded and therefore integrable with respect to P. The proposition then follows by the dominated convergence theorem since by assumption $\int_M p(x)^2\, dV(x) < \infty$. □

The next step in the proof is to control the deviation of $U_{h,h}$ from its expectation. A straightforward application of the concentration inequality for U-statistics of Hoeffding in Theorem 2.53 yields the following theorem.

Theorem 2.49 *Let M, k, P fulfill the standard assumptions then*

$$\mathrm{P}\left(|U_{n,h} - \mathbb{E}\, U_{n,h}| \ge \epsilon\right) \le 2\, e^{-\frac{[n/2]h^m\epsilon^2}{2\|k\|_\infty \mathbb{E}\, U_{n,h} + 2/3\|k\|_\infty - h^m \mathbb{E}\, U_{n,h}|\epsilon}}.$$

Furthermore let $n \to \infty$ and $h \to 0$ then if $nh^m \to \infty$,

$$\lim_{n\to\infty} U_{n,h}(k) = C_1 \int_M p(x)^2\, dV(x), \quad \text{in probability.}$$

If the stronger condition $nh^m / \log n \to \infty$ holds then

$$\lim_{n\to\infty} U_{n,h}(k) = C_1 \int_M p(x)^2\, dV(x), \quad \text{almost surely.}$$

Proof: We have $\|k_h\|_\infty = \|k\|_\infty / h^m$ and it can be verified that $\mathrm{Var}\, U_{n,h} \le K/h^m \mathbb{E}\, U_{n,h}$. Since $\mathbb{E}\, U_{n,h}(k) < \infty$ we can apply Theorem 2.53. Using this concentration inequality, convergence in probability of $U_{n,h}$ towards its expectation follows immediately from the condition $nh^m \to \infty$. Moreover we know from Proposition 2.48 the form of $\mathbb{E}\, U_{h,h}$ as $h \to 0$. Complete convergence which implies almost sure convergence follows from $\sum_{n=1}^\infty \mathrm{P}(|U_{n,h} - \mathbb{E}\, U_{n,h}| \ge \epsilon) < \infty$. This follows if the stronger condition holds. □

The previous theorem together with the following corollary will be the cornerstones of our algorithm.

Corollary 2.50 *Let M, P and k fulfill the standard assumptions and define $k_h = \frac{1}{h^l} k(\|i(x) - i(y)\|^2 / h^2)$. Then if $h \to 0$ and $nh^l \to \infty$,*

$$\begin{aligned} &\lim_{n\to\infty} U_{n,h}(k) = \infty, && \text{if} \quad l > m \\ &\lim_{n\to\infty} U_{n,h}(k) = 0, \text{ in probability}, && \text{if} \quad l < m \end{aligned}$$

Proof: In Theorem 2.49 we have shown that for $l = m$, $U_{n,h}(k)$ converges to $C_1 \int_M p(x)^2 \, dV(x)$ in probability. Now with the different power of h in front of the kernel we have convergence towards $\frac{C_1}{h^{l-m}} \int_M p(x)^2 \, dV(x)$. Since the integral is finite, $U_{n,h}(k)$ diverges if $l > m$ and converges to zero if $l < m$. □

Note that we get convergence to a finite number if and only if $l = m$ since $0 < \int_M p(x)^2 \, dV(x) < \infty$.

The Algorithm

The algorithm is based on the convergence result in Theorem 2.49 and Corollary 2.50. Using these results, we know that in order to get convergence in probability the bandwidth h has to fulfill $nh^m \to \infty$. Otherwise the U-statistic either diverges or approaches zero. We will use this property by fixing a convergence rate for each dimension, that means we are fixing h as a function of the sample size n. Then we compute the U-statistic for subsamples of different sizes where h varies according to the function we have fixed. Finally the dimension is determined by the U-statistic which has the smallest slope as a function of h.

First Step: Fixing $h_l(n)$

As a first step we fix $h_l(n)$ as a function of the sample size n and the dimension l. We choose the function $h_l(n)$ in such a way that it is just sufficient for convergence in probability so that $h_l(n)$ approaches zero at the fastest allowed rate, that is

$$nh(n)^l = \frac{1}{c^l} \log n \;\Rightarrow\; h_l(n) = \frac{1}{c} \left(\frac{\log n}{n} \right)^{1/l},$$

where c is a constant. The crucial point of this procedure is that the scales at which we look at the data vary according to the dimension l so that $U_{n,h}(k)$ will depend as a function of the sample size n on the chosen dimension l. We fix the constant c in the algorithm by determining a certain nearest neighbor scale. Let N be the total number of samples of our data set and define $d(X_i)$ as the distance of the sample X_i to its nearest neighbor. We set:

$$h_l(N) = \frac{1}{N} \sum_{i=1}^{N} d(X_i) \;\Rightarrow\; c = \frac{1}{h_l(N)} \left(\frac{\log N}{N} \right)^{1/l}.$$

In total we get for the function $h_l(n)$:

$$h_l(n) = h_l(N) \left(\frac{N}{n} \frac{\log n}{\log N} \right)^{1/l}.$$

Note that $h_l(N)$ does not depend on the dimension that is we examine the full data (all N sample points) for each dimension at the same scale $h_l(N)$. The important point is now that as we consider subsamples of size n the scale $h_l(n)$ is different for each dimension.

Second step: Computing the dimension

The choice of the kernel seems not to influence the result much. We choose a kernel with compact support to save computational time, that is

$$k(x) = (1 - x)_+$$

Note that this is a very simple kernel which allows a fast evaluation. However this kernel function is not differentiable at $x = 1$ so that it violates our assumptions on the kernel function. Instead one could use the kernel function $k(x) = e^{\alpha/(1-x)}$ for $x \leq 1$ and $k(x) = 0$ elsewhere which fulfills this condition. This kernel function has the problem that it is very small already before $x = 1$ which results in an effective choice of a smaller scale. Therefore in order to compare the results of the two kernel functions, one has to take a slightly larger value of the minimal scale $h_l(N)$ for the second one (dependent on α).
We consider subsamples of size $\{[N/5], [N/4], [N/3], [N/2], N\}$. For each dimension $l \in \{1, \ldots, \mathrm{l}_{\max}\}$, where we put usually $l_{\max} = \min\{d, 20\}$[20], we compute the empirical estimate of $U_{[N/r],h_l([N/r])}(k)$, $r = 1, \ldots, 5$.
In particular for small sample sizes taking subsamples is usually not a good idea since the variance of these estimates is quite high. In order to improve the estimates in particular of the subsamples of small size, we consider not only one subsample but several ones by using the so called two-sample U-statistics which is defined as follows. Given two i.i.d. samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, one considers the following U-statistic

$$U = \frac{1}{n^2} \sum_{i,j=1}^{n} k(X_i, Y_j).$$

It was shown by Hoeffding [48] that also this form of the U-statistic converges as in Theorem 2.53. In our case we will take two samples of the same distribution so that the expectation and variance and therefore also the limit and the convergence rate are the same as for the one-sample U-statistic.
Let us explain how we use this result for the subsamples. Consider the size $[N/2]$. Using the full set of N data points, we can build three subsamples, namely (X_{2k}, X_{2k}), (X_{2k+1}, X_{2k}), and (X_{2k+1}, X_{2k+1}). The first and the last one lead to one-sample U-statistics and the second one to a two-sample U-statistic. For each of these subsamples we compute the estimates of $U_{[N/2],h([N/2])}(k)$. Obviously one gets for the subsample of size $[N/r]$, $r = 5, 4, 3, 2, 1$, $r(r + 1)/2$ such estimates, and for each r we take the mean of them. This method looks at first quite complicated. However the implementation is straightforward and solves the problem that especially for small sample sizes N taking subsamples leads to high variances in the estimates. Using instead a set of subsamples with the described method, we can minimize the variance of the estimates corresponding to the subsamples.
The estimation of the U-statistics can be done for all dimensions and for all subsample sizes simultaneously. Especially for high-dimensional data which is potentially the most interesting one the main computational cost lies in the computation of the distances and not in the calculations of $U_{[N/r],h([N/r])}(k)$.
Finally in order to determine the dimension we fit for each dimension l a line through

[20]The choice of 20 as an upper bound is completely arbitrary and can be changed according to the problem at hand.

the five points

$$\left[\log h_l([N/r]),\, \log U_{[N/r],h_l([N/r])}(k)\right],\ r = 1, \ldots, 5,$$

with weighted least squares with weights $w(r) = 1/r$. This can be easily done in closed form. The dimension is then determined by the line with the smallest absolute value of the slope of the line. This is justified since by Theorem 2.49 and Corollary 2.50 the slope of $\log U_{n,h_l(n)}(k)$ behaves as $(m-l)\log h_l(n)$ as $n \to \infty$ and $h \to 0$. We use weighted least squares since for smaller subsamples we look at the data at a larger scale. Therefore if one has high curvature these estimates are less reliable.

Experiments

The experiments we perform are only partially based on datasets which have been previously used for dimensionality estimation. The reason for this is that these datasets do not have high extrinsic and intrinsic curvatures. In our experiments based on artificial datasets we study the influence of high curvature as well as noise on our estimator. Later on we will evaluate the estimator on two real world datasets. The first one is the face database used in the study of ISOMAP [95] and the MNIST database. For the MNIST database we actually do not know the intrinsic dimensionality. Therefore we study first for the digit 1 an artificial dataset where we can control the number of dimensions. This study gives then a hint how well our estimator performs. We compare the results of our method to that of the correlation dimension estimator described in the introduction and the estimator of Takens defined in [94] as

$$\dim_{\mathrm{Corr}}^{-1} = -\,<\log(\,\|i(X_i) - i(X_j)\|\ /h_{\mathrm{Takens}})>,$$

where $<>$ is the mean over all distances smaller than h_{Takens}. In order to do a fair comparison, we tried to optimize the scales s_i for the correlation dimension estimator as well as the 'maximal' scale h_{Takens} for the Takens estimator over all the datasets. Then we fixed them to $s_i = d + 0.2\,r\,\sigma$, $r = 1, \ldots, 5$, for the correlation dimension estimator and the maximal scale of the Takens estimator to, $h_{\mathrm{Takens}} = d + \sigma$, where d is the mean and σ the standard deviation of the nearest neighbor distances. We would like to note that also for the Takens estimator one has to determine only one scale however since it is a kind of 'maximal scale' it is more difficult to choose then a minimal scale as for our method.

Sinusoid on the circle In this example our one-dimensional submanifold is a strongly oscillating sinusoid on the circle in $\mathbb{R}^3$, see Figure 2.4.

$$s(t) : [0, 2\pi) \to \mathbb{R}^3, \quad s(t) \to (\sin t, \cos t, \frac{1}{10}\sin 150t).$$

We sample straightforward in our coordinate expression which yields a non-uniform probability measure on this manifold where more points appear at the extreme points of the sinusoid. We compare this submanifold to a circle with uniform noise of height 0.1 in the z-direction, see Figure 2.5, which results in a strip of the cylinder which is 2-dimensional. The results are shown in Table 2.2 for $400, 500$ and 600 sample points. Two conclusions can be drawn. The first rather obvious one is that very curved submanifolds require a large number of samples so that their dimension can

estimated correctly since the high curvature of the sinusoid is misinterpreted as a second dimension for small sample sizes. The second one is that for small sample sizes it is impossible to distinguish between noise and high curvature. The rather surprising fact is that already for a sample size of 600 we have an almost perfect distinction between the one-dimensional sinusoid and the two-dimensional strip of the cylinder.

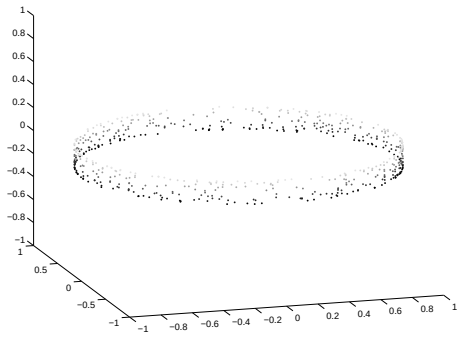


Abbildung 2.4: 600 samples of the sinusoid

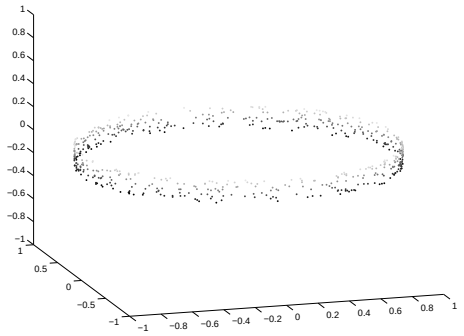


Abbildung 2.5: 600 samples of the circle with uniform noise of height 0.1 in the z-direction

The m-sphere In this experiment we study the m-dimensional spheres S^m embedded in $\mathbb{R}^{m+1}$. The $n = 600, 800, 1000, 1200$ data points are sampled in 90 trials uniformly from the sphere S^m. The number of successful trials is given in Table 2.3. For S^7 and S^9 the number of samples is no longer sufficient (curse of dimensionality), most of the time the dimension is underestimated by one.

The Gaussian distribution In this experiment the data is drawn from an isotropic Gaussian in $\mathbb{R}^d$ in order to show how the estimators can deal with a varying probability distribution. The results are shown in Table 2.4. In particular for dimension 6 we outperform the other estimators. This is the only dataset where our method has a clear advantage over the Takens estimator.

The 10-Möbius strip The k-Möbius strip is a submanifold in $\mathbb{R}^3$ which can be created by taking a rectangle, twisting it k-times, and then identifying the ends. If k

Tabelle 2.2: Correct estimates of dimension 1 for the sinusoid and dimension 2 for the noisy circle of 90 trials. $a/b/c$, a our method, b corr. dim. est., c Takens est.

	400	500	600
Sinuosid	15/0/12	49/57/49	86/88/90
Noisy Circle	90/90/90	90/90/90	90/90/90

Tabelle 2.3: Number of correct estimates of 90 trials for S^m. $a/b/c$, a our method, b corr. dim. est., c Takens est.

	600	800	1000	1200
S^3	90/89/90	90/90/90	90/90/90	90/90/90
S^5	83/80/88	87/81/90	89/86/90	90/89/90
S^7	68/57/65	73/66/79	78/66/78	79/72/84
S^9	30/36/32	47/30/43	50/33/47	58/45/50

is odd one gets a non-orientable manifold with surprising properties. It is obvious that this manifold has high extrinsic curvature increasing with the number of twists k. We considered a 10-Möbius strip, see Figure 2.6 for an illustration with 16000 points. The coordinate representation for $u \in [-1, 1], v \in [0, 2\pi)$, is as follows:

$$\begin{aligned} x_1(u,v) &= \big(1 + \frac{u}{2}\cos\big(\frac{k}{2}v\big)\big)\cos(v), \\ x_2(u,v) &= \big(1 + \frac{u}{2}\cos\big(\frac{k}{2}v\big)\big)\sin(v), \\ x_3(u,v) &= \frac{u}{2}\sin\big(\frac{k}{2}v\big). \end{aligned}$$

We sampled in this coordinate representation $20, 40, 80$ and 120 points. This example is done to illustrate that even for manifolds with high extrinsic curvature the intrinsic dimension can be estimated with a relatively small sample size, see Table 2.5.

A 12-dimensional manifold in $\mathbb{R}^{72}$ As the last artificial dataset we present a high-dimensional dataset, a 12-dimensional manifold in 72-dimensions. The submanifold

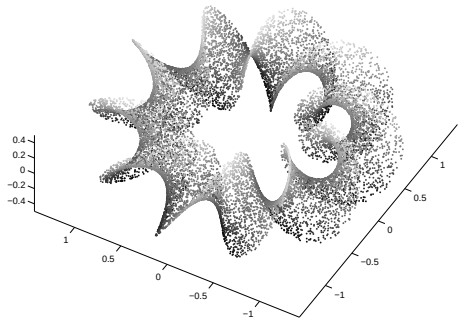


Abbildung 2.6: The 10-Möbius strip with 16000 points.

Tabelle 2.4: Number of correct estimates of 90 trials of a d-dimensional Gaussian. $a/b/c$, a our method, b corr. dim. est., c Takens est.

Dim	100	200	400	800
3	86/81/86	90/90/90	90/90/90	90/90/90
4	76/65/75	85/78/85	90/89/90	90/90/90
5	58/41/49	66/49/59	81/72/82	90/90/90
6	44/25/23	41/24/15	49/37/34	77/55/61

Tabelle 2.5: Correct estimates of 90 trials of the 10-Möbius strip. $a/b/c$, a our method, b corr. dim. est., c Takens est.

20	40	80	120
49/34/44	71/68/73	83/78/86	88/82/90

is generated by

$$\begin{aligned} x(\alpha) :&[0,1]^{12} \to \mathbb{R}^{72}, \\ x_{2i-1}(\alpha) &= \alpha_{i+1}\cos(2\pi\alpha_i), \quad i = 1,\dots,11, \\ x_{2i}(\alpha) &= \alpha_{i+1}\sin(2\pi\alpha_i), \quad i = 1,\dots,11, \\ x_{23}(\alpha) &= \alpha_1\cos(2\pi\alpha_{12}), \quad x_{24}(\alpha) = \alpha_1\sin(2\pi\alpha_{12}), \\ x_{j+24} &= x_{j+48} = x_j,\ j = 1,\dots,24. \end{aligned}$$

By construction the 12-dimensional manifold lies effectively in a 24-dimensional subspace. We sample directly in these coordinates which yields a non-uniform probability measure on the manifold which is concentrated around the origin. This leads to an interesting phenomenon when we try to estimate the dimension. The results shown in Table 2.6 illustrate the connection between high curvature and non-trivial probability measure effects on the manifold. We believe that they somehow cancel out in this case. Namely a highly curved manifold leads to an overestimation of the dimension whereas a concentrated probability measure leads to an underestimation. For a relatively small sample size of 800 we already get a quite good estimate of the dimension which is probably due to the high concentration around the origin.

The ISOMAP face database The ISOMAP face database consists of 698 images (256 gray levels) of size 64×64 of the face of a sculpture. This dataset has three parameters: the vertical and horizontal pose and the lighting direction (one-dimensional). All estimators get for this dataset in $\mathbb{R}^{4096}$ the correct intrinsic dimension of 3.

The MNIST dataset The MNIST dataset of handwritten digits consists of 70000 images (256 gray levels) of size 28×28. In the generation of the MNIST dataset the center of mass was computed for all images and then the image was translated such that the center of mass lies at the center of the image. However note that this does not mean that there are no translational degrees of freedom in this dataset since e.g. the digit 1 can be written with a line below or not and therefore the center of mass varies.

Tabelle 2.6: Correct est. of 90 trials on the 12-dim manifold. $a/b/c$, a our method, b corr. dim. est., c Takens est.

200	400	800	1600
46/42/43	60/51/61	64/70/68	84/85/85

The artificial 1-digit dataset

The intrinsic dimension of each digit is in principle unknown. In order to validate our experiment, we constructed an artificial dataset of the digit 1 where we can control the dimensionality. Namely we have 5 degrees of freedom: two for translations (T), one for rotation (R), one for line thickness (L), and one for having a small line at the bottom (V). The images are constructed by having an abstract 1 as a function on $[0,1]^2$ where the different transformations are applied and then this function on $[0,1]^2$ is discretized to an image of size 28×28. We constructed 5 datasets each of size 10000. The letter combination shows which transformations have been applied, see Figure 2.7 for samples of the TRLV dataset. The results of the estimators on this four datasets are shown in Table 2.7. In three cases we are able to estimate the correct intrinsic dimension whereas in one case we overestimate the dimension. Regarding these results on this artificial dataset, we have some confidence in the results on the real MNIST dataset.

Abbildung 2.7: Samples of the artificial 1-dataset T+R+L+V.

Intrinsic Dimensionality of the Digits in MNIST The estimated intrinsic dimensions are reported for each digit in Table 2.8 together with the number of samples of the digit in the MNIST database. Considering our result of the artificial dataset for the digit 1, we think that an estimated dimension 8 seems quite reasonable. Additional degrees of freedoms could be the length of the main line, the angle between the main line and the upper line, and the length of the upper line. The intrinsic dimensions of digit 2 and 3 were estimated in [27] for a subsample of size 1000 as 13 and 12, respectively 12 and 11, depending on the way they build their neighborhood graph. We estimate an intrinsic dimension of 13 for digit 2 and 14 for digit 3. In comparison the results roughly agree. The difference could arise since as we consider the whole dataset we look at the data at a smaller scale and therefore estimate a higher dimension.

Discussion and open problems

We have presented an algorithm for intrinsic dimensionality estimation of a submanifold in $\mathbb{R}^d$ from random samples. The assumptions we impose on the submanifold and the probability measure on this submanifold are not restrictive. Our theoretical analysis clarifies the influence of the curvature of the submanifold and smoothness of the density on the asymptotic behavior of our estimated quantity. Opposite to

Tabelle 2.7: Estimated Dimension of the artificial 1-data sets.

Art. Digit 1	T	TR	TRL	TRLV
int. dim.	2	3	4	5
est. int. dim.	2/1/2	3/4/4	5/4/4	5/5/5

Tabelle 2.8: Number of samples and estimated intrinsic dimensionality of the digits in MNIST.

1	2	3	4	5
7877	6990	7141	6824	6903
8/7/7	13/12/13	14/13/13	13/12/12	12/12/12
6	7	8	9	0
6876	7293	6825	6958	6903
11/11/11	10/10/10	14/13/13	12/11/11	12/11/11

the standard correlation dimension estimator we only have to choose once a scale at which we examine the data. The scales at which we examine subsamples are then fixed so that we have only one free parameter in our algorithm. Even more we fixed this parameter by choosing the somehow smallest scale at which it makes sense to look at the data. In that sense we have presented an algorithm without parameters which estimates the dimension for all kinds of submanifolds irrespectively of their intrinsic and extrinsic curvature and works well also for real world datasets. The experiments show that we are on average significantly better than the correlation dimension estimator and on a slightly better level than the Takens estimator (in particular for the Gaussian dataset).
The most apparent open problem is the intrinsic dimensionality estimation of submanifolds which are sampled noisy. Kegl proposes in [53] to estimate the correlation integral at different scales so that finally the dimension is no longer a fixed number but a function of the scale at which one examines the data. In principle this approach is appealing, however it does not solve the question under which conditions one can in prinicipal still estimate the dimension of a submanifold even if it is corrupted by noise. We have demonstrated by our first experiment that the estimation of the dimension of a highly curved submanifold leads to wrong results as long as the scale at which one looks at the data is larger than the radius of curvature of the submanifold. This implies that without bound on the curvature the estimation of the 'correct' dimension is impossible. However, even if one assumes that one has a bound on the radius of curvature, one needs also a bound on the noise level. Effectively the noise level has to be sufficiently smaller than the radius of curvature. Otherwise it can happen that the noise 'connects' parts of the submanifold and therefore the intrinsic dimensionality cannot be observed anymore. Then examining the data at a scale which lies in between the noise level and the radius of curvature should lead to a correct estimation of the dimension. A more formal mathematical justification of the just described scheme remains as an open problem.

2.5 Appendix

2.5.1 U-statistics

In this chapter we collect some basic facts about U-statistics. They are taken from the paper of Hoeffding [48] and the book of Serfling [84]. In particular we are interested in concentration inequalities for U-statistics.

Definition 2.51 (One-sample U-statistic) *Let $X_i, , i = 1, \dots, n$ be an i.i.d. sample from P on a set S and let $k : S \times S \to \mathbb{R}$ be a symmetric function. Then*

$$U_n(k) = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} k(X_i, X_j)$$

is called a (one-sample) U-statistic.

We just note that in a V-statistic it is summed also over the diagonal. In general if k is a bounded function on the diagonal all concentration inequalities for the U-statistic can be transferred to the corresponding V-statistic. We have introduced a U-statistic of degree 2, that is k is a function of 2 variables. However the general setting considers symmetric functions k of an arbitrary fixed number of variables. Another generalization are the so-called two-sample U-statistics where one has samples drawn from two possibly different probability measures.

Definition 2.52 (Two-sample U-statistic) *Let X_i, $i = 1, \dots, n$ and Y_j, $j = 1, \dots, m$ be two i.i.d. samples from P resp. Q on S and $k : S \times S \to \mathbb{R}$ a symmetric function. Then*

$$U_{nm}(k) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k(X_i, Y_j)$$

is called a (two-sample) U-statistic.

Most of the theory of one-sample U-statistics can be transferred to two-sample U-statistics. In particular the following Bernstein type concentration inequality can be proven to hold also for a two-sample U-statistic [48].

Theorem 2.53 *Let $\|k\|_\infty \leq b$, $\mathbb{E}\, k(X, Y) < \infty$, and $\sigma^2 = \operatorname{Var} k(X, Y) < \infty$ then*

$$P\left(|U_n(k) - \mathbb{E}\, U_n(k)| \geq \epsilon\right) \leq 2\, e^{-\frac{[n/2]\epsilon^2}{2\sigma^2 + 2/3|b - \mathbb{E}\, U_n(k)|\epsilon}},$$

where $[x]$ denotes the greatest integer smaller than x.

Kapitel 3

Kernels, Associated Structures and Generalizations

3.1 Introduction

Positive definite kernels are extremely powerful and versatile tools. They allow to construct spaces of functions on an arbitrary set with the convenient structure of a Hilbert space. Methods based on such kernels are usually very tractable because of the particular structure (reproducing property) of the space of functions. This has a large number of applications in particular for statistical learning, approximation, or interpolation where one has to manipulate functions defined on various types of data.

Our goal is to survey some of the results relevant for machine learning. Since the literature is scattered among various fields of mathematics, we believe that the learning community benefits from a unified exposition of the results and relationships between them. This chapter is an attempt to go into that direction where we concentrate on basic structures and properties and ways of generalizing the standard setting. Although the theory can be quite technical, we want to shed light on its essence and convey several important messages that anyone working with kernels and associated spaces should have in mind.

A first message is that there is an equivalence (in a strong) sense between several objects: positive definite kernels (which are specific functions of two variables), Hilbert spaces of functions with a certain topological property, Gaussian processes, and a class of positive operators. A second message is that the mysterious feature mapsässociated to kernels are not related to the Mercer property and they exist and can be defined in many different ways as soon as the kernel is positive definite. A third message is that the integral operator associated to a kernel has nice properties even if the kernel is not continuous. In particular it is tightly related to the covariance operator (i.e. the population limit of a covariance matrix) as they have the same spectrum. A fourth message is that most attempts to generalize kernels (e.g. to operator-valued or generalized functions) end up being special cases. This may seem surprising but it can be easily seen by changing the point of view one adopts, going from sets to functions on these sets. Finally, we recall that there exists a well-developed theory of indefinite kernels (i.e. kernels that are not positive definite) and their associated structures based on the notion of reproducing kernel Krein spaces.

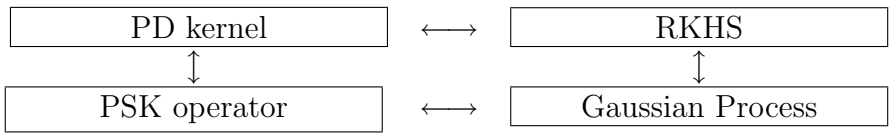
This chapter has been partially published in [44].

3.2 Positive Definite Kernels and Associated Structures

We restrict ourselves to the real-valued case and denote by $\mathbb{R}^{\mathcal{X}}$ the vector space of functions from $\mathcal{X}$ to $\mathbb{R}$ where $\mathcal{X}$ is an arbitrary *index* set[1] and by $\mathbb{R}^{[\mathcal{X}]}$ the vector space of finite linear combinations of evaluation functionals (i.e. of the form $\sum_{i=1}^{n} a_i \delta_{x_i}$). We define a bilinear map from $\mathbb{R}^{[\mathcal{X}]} \times \mathbb{R}^{\mathcal{X}}$ to $\mathbb{R}$ as

$$\left\langle \sum_{i=1}^{n} \alpha_i \delta_{x_i}, f \right\rangle_{\mathbb{R}^{[\mathcal{X}]}, \mathbb{R}^{\mathcal{X}}} := \sum_{i=1}^{n} \alpha_i f(x_i) \text{ where } x_1, \ldots, x_n \in \mathcal{X}.$$

In this first section we shortly review the notion of positive definite (PD) kernels and its associated structures. Indeed such a kernel can be associated to a space of functions, called reproducing kernel Hilbert space (RKHS), to a linear operator called positive symmetric kernel (PSK) operator and to a Gaussian process in a natural way. The following diagram illustrates the fact that all these notions are tightly related.



3.2.1 Definitions

We now give the definitions of the four objects in the preceding diagram.

Definition 3.1 *A real-valued symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a* **positive definite (PD) kernel** *if for all $n \geq 1$, $x_1, \ldots, x_n \in \mathcal{X}$, $c_1, \ldots, c_n \in \mathbb{R}$,*

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0. \tag{3.1}$$

The set of all real-valued positive definite kernels on $\mathcal{X}$ is denoted $\mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$.

Definition 3.2 *A* **positive symmetric kernel (PSK) operator** *K is a linear operator $K : \mathbb{R}^{[\mathcal{X}]} \to \mathbb{R}^{\mathcal{X}}$ which is symmetric*

$$\forall v', w' \in \mathbb{R}^{[\mathcal{X}]}, \quad \langle v', Kw' \rangle_{\mathbb{R}^{[\mathcal{X}]}, \mathbb{R}^{\mathcal{X}}} = \langle w', Kv' \rangle_{\mathbb{R}^{[\mathcal{X}]}, \mathbb{R}^{\mathcal{X}}},$$

and positive: $\forall v' \in \mathbb{R}^{[\mathcal{X}]}, \quad \langle v', Kv' \rangle_{\mathbb{R}^{[\mathcal{X}]}, \mathbb{R}^{\mathcal{X}}} \geq 0$.
The set of all such operators is denoted as $L_+(\mathbb{R}^{\mathcal{X}})$.

Definition 3.3 *A* **reproducing kernel Hilbert space (RKHS)** *$\mathcal{H}$ on $\mathcal{X}$ is a Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$ where all evaluation functionals $\delta_x : \mathcal{H} \to \mathbb{R}$, $\delta_x(f) = f(x)$ are continuous[2], equivalently for all $x \in \mathcal{X}$, there exists a $M_x < \infty$ such that*

$$\forall\, f \in \mathcal{H},\ |f(x)| \leq M_x \|f\|_{\mathcal{H}}\ .$$

The set of all such spaces is denoted as $\mathrm{Hilb}(\mathbb{R}^{\mathcal{X}})$.

[1] or also called *input space*.
[2] with respect to the topology induced by the norm of $\mathcal{H}$

This definition stresses the fact that an RKHS is a Hilbert space of pointwise defined functions where norm convergence implies pointwise convergence.

Definition 3.4 *A* **centered Gaussian process** *indexed by $\mathcal{X}$ is a family X_x, $x \in \mathcal{X}$, of jointly normal random variables, that is for each finite set $x_1, \ldots, x_n \in \mathcal{X}$, the vector $(X_{x_1}, \ldots, X_{x_n})$ is centered Gaussian*[3].
The set of all such processes is denoted by $G(\mathcal{X})$.

Note that we restrict ourselves to centered Gaussian random variables. In principle the results can be transferred to the non-centered case.

3.2.2 Properties and Connections

The fundamental and most important property of PD kernels is the relationship with inner product spaces. Often the use of kernel methods is justified by the implicit mapping of the input space $\mathcal{X}$ into a 'high-dimensional' feature space. As the next well-known proposition shows, such a mapping exists as soon as the kernel is positive definite and actually characterizes such kernels.

Proposition 3.5 *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a PD kernel if and only if there exists a Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, y \in \mathcal{X}\,, k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$.*

Note that this result has nothing to do with Mercer's theorem (we will come back to this issue in section 3.3.1). There exist many proofs of the above proposition and we will give one later.
We will now establish the connections between the four objects we have introduced in the previous section. It is well known (see e.g. [4]) that $\mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$ is closed under addition, multiplication by a non-negative number, and point-wise limits and has a partial order relationship ($k_1 \succeq k_2$ if $k_1 - k_2$ is PD). It is less known that all the other sets introduced above ($L_+(\mathbb{R}^{\mathcal{X}})$, $\mathrm{Hilb}(\mathbb{R}^{\mathcal{X}})$ and $G(\mathcal{X})$) have a similar structure. Actually, the following strong equivalence between these spaces and their structures holds.

Theorem 3.6 *[5] There exist bijections which preserve the structure of ordered, closed convex cones between each two of the following sets*

$$\mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}\,,\ L_+(\mathbb{R}^{\mathcal{X}})\,,\ \mathrm{Hilb}(\mathbb{R}^{\mathcal{X}})\,,\ G(\mathcal{X})\,.$$

An example how the order is transferred from $\mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$ to $\mathrm{Hilb}(\mathbb{R}^{\mathcal{X}})$ is the following.

Theorem 3.7 *[4] Let $k_1, k_2 \in \mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$ and $\mathcal{H}_1, \mathcal{H}_2$ their associated RKHS. Then $\mathcal{H}_1 \subset \mathcal{H}_2$, and $\|f_1\|_{\mathcal{H}_1} \geq \|f_1\|_{\mathcal{H}_2}\,, \forall\, f_1 \in \mathcal{H}_1$ if and only if $k_1 \preceq k_2$.*

In the remaining part of this section we will show some of these bijections. All other bijections can then be constructed by composing two of these maps. Additionally we introduce in the appendix several objects associated to a Gaussian Process. These objects become interesting if one is interested for example in sample path properties of a Gaussian Process.

[3] equivalently, all linear combinations $\sum \alpha_i X_{x_i}$ are real Gaussian random variables with zero mean.

PD Kernels and PSK Operators

The bijection between kernels and kernel operators is made explicit in the following lemma.

Lemma 3.8 *[83] Let $k \in \mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$. The linear operator $K : \mathbb{R}^{[\mathcal{X}]} \to \mathbb{R}^{\mathcal{X}}$ defined by $K(\delta_x) = k(x, \cdot)$ is a PSK operator. Conversely, given $K \in L_+(\mathcal{X})$, the function k defined as $k(x, y) = \langle \delta_x, K\delta_y \rangle_{\mathbb{R}[\mathcal{X}],\mathbb{R}^{\mathcal{X}}}$ is a PD kernel.*

The above lemma indicates the close correspondence between the kernel function and its associated operator. In particular, symmetry of one corresponds to symmetry of the other, while positive definiteness of the former one corresponds to positivity of the latter.

PD Kernels and RKHS

The following fundamental theorems illustrate the link between RKHS and PD kernels.

Theorem 3.9 *[4] Let $\mathcal{H}$ be a Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$, $\mathcal{H}$ is a RKHS if and only if there exists a map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that*

$$\begin{aligned} &\forall\, x \in \mathcal{X}, \quad k(x, \cdot) \in \mathcal{H}, \\ &\forall\, f \in \mathcal{H}, \quad \langle f(\cdot), k(x, \cdot) \rangle_{\mathcal{H}} = f(x)\,. \end{aligned}$$

If such a k exists, it is unique and it is a PD kernel.

The second property is called the *reproducing property* of the RKHS and k is called the (reproducing) kernel of $\mathcal{H}$.

Theorem 3.10 *(Moore, see [4]) If k is a positive definite kernel, then there exists a unique reproducing kernel Hilbert space $\mathcal{H}$ whose kernel is k.*

Proof: We give a sketch of the proof (of both theorems above) which involves an important construction. The proof proceeds in three steps. The first step is to consider the set of all finite linear combinations of the kernel: $\mathcal{G} = \mathrm{Span}\{k(x, .) : x \in \mathcal{X}\}$ and to endow it with the following inner product

$$\left\langle \sum_i a_i k(x_i, .), \sum_j b_j k(x_j, .) \right\rangle_{\mathcal{G}} = \sum_{i,j} a_i b_j k(x_i, x_j)\,. \tag{3.2}$$

It can be shown that this is indeed a well-defined inner product. At this point we already have the reproducing property on $\mathcal{G}$. The second step is to construct the semi-norm associated to this inner product and to show (thanks to the Cauchy-Schwarz inequality) that it is actually a norm. Hence, and this is the third step, $\mathcal{G}$ is a pre-Hilbert space which can be completed[4] into a Hilbert space $\mathcal{H}$ of functions. Finally, one has to check that the reproducing property carries over to the completion. It is then easy to show that any other Hilbert space with the same reproducing kernel has to be isometric isomorphic. Namely let $\mathcal{K}$ be another RKHS with reproducing

[4] i.e. we add to $\mathcal{G}$ the pointwise limits of all Cauchy sequences of elements of $\mathcal{G}$

kernel k. It is obvious that $\mathcal{H}$ has to be a closed subspace of $\mathcal{K}$. Then $\mathcal{K}$ can be decomposed into $\mathcal{K} = \mathcal{H} \oplus \mathcal{H}^\perp$. Now let $f \in \mathcal{K}$, but $f \notin \mathcal{H}$. Then for all $x \in \mathcal{X}$

$$f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{K}} = \left\langle f^{\|} + f^{\perp}, k(x,\cdot)\right\rangle_{\mathcal{K}} = f^{\|}(x)$$

Therefore $f \equiv f^{\|}$ which is a contradiction so that we get $\mathcal{K} = \mathcal{H}$. □

Hence $\mathcal{H}$ is simply the completion of the linear span (i.e. finite linear combinations) of the functions $k(x,\cdot)$ endowed with the inner product (3.2).

PD Kernels and Gaussian Processes

It is well-known that a centered Gaussian process $(X_x)_{x\in\mathcal{X}}$ is uniquely determined by its covariance function $\mathbb{E}[X_s X_t]$ which is a positive definite kernel. Conversely any positive definite kernel defines a covariance function and therefore a unique Gaussian process by Theorem 3.40.

3.3 Useful Properties

A quite useful relationship between $k \in \mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$ and the set $\mathcal{X}$ is that k induces a semi-metric on $\mathcal{X}$ by $d_k(x,y) = \|k(x,\cdot) - k(y,\cdot)\|_{\mathcal{H}}$. Many properties of the RKHS can be stated in terms of this (semi)-metric space $(\mathcal{X}, d_k)$ as we will later see in the study of the separability of the RKHS.

3.3.1 Feature Maps

Sometimes Mercer's theorem is mentioned as a necessary condition to have a feature map. The goal of this section is to show, that it is a sufficient condition but it requires additional assumptions on $\mathcal{X}$ and k. As we have seen in Proposition 3.5 a necessary and sufficient condition that such a feature map into a Hilbert space exists is that the kernel is positive definite. Two questions can then be raised: Can such a map be constructed explicitly ? What is the induced representation for the kernel ? The first question has an affirmative answer without any further assumptions on k as the following feature maps $\Phi : X \to \mathcal{H}$ show.

1. *Aronszajn map*
   $\phi : x \mapsto k(x,\cdot)$, $\mathcal{H}$ is the associated RKHS, $k(x,y) = \langle k(x,\cdot), k(y,\cdot)\rangle$
2. *Kolmogorov map*
   $\phi : x \mapsto X_x$, $\mathcal{H} = L_2(\mathbb{R}^{\mathcal{X}}, \mu)$ where μ is a Gaussian measure[5], $k(x,y) = \mathbb{E}[X_x X_y]$
3. *Integral map*
   There exists a set T and a measure μ on T such that one has $\phi : x \mapsto (\Gamma_x(t))_{t\in T}$, $\mathcal{H} = L_2(T,\mu)$[6], $k(x,y) = \int \Gamma(x,t)\Gamma(y,t)d\mu(t)$
4. *Basis map*
   given any orthornormal basis[7] $(f_\alpha)_{\alpha\in I}$ of the RKHS associated to $\mathcal{H}$, one has $\phi : x \mapsto (f_\alpha(x))_{\alpha\in I}$, $\mathcal{H} = \ell_2(I)$[8] and $k(x,y) = \sum_{\alpha\in I} f_\alpha(x) f_\alpha(y)$.

[5] see Section 3.7.1 for details.
[6] The Kolmogorov map shows that such a set T and a measure μ always exist.
[7] such a basis always exists but may be uncountable, in which case, only a countable subset of the coordinates of any vector are non-zero.
[8] space of square summable functions on I with countable support

When infinite sums are involved like in the last case, it is important to specify in which sense the sum converges. In general the convergence occurs for each pair (x, y). However, [68] shows one has stronger convergence, namely uniform on every set $A \times B \subset \mathcal{X} \times \mathcal{X}$, with A bounded and B compact (w.r.t. the topology induced by d_k).
The integral map seems at first to be redundant since also the Kolmogorov map is an integral map. We listed it as an extra feature map since for certain classes of kernels, e.g. translation-invariant kernels on $\mathbb{R}^d$, the set T needed to represent the kernel is much smaller than $\mathbb{R}^{\mathcal{X}}$, namely it is $\mathbb{R}^d$ instead of $\mathbb{R}^{\mathbb{R}^d}$ for the translation-invariant kernels on $\mathbb{R}^d$.
Given additional structure of the kernel resp. the corresponding RKHS, there exist other feature space interpretations. Mercer's theorem is a special case of the basis map. It gives stronger convergence properties of the kernel representation, but needs additional assumptions, namely $\mathcal{X}$ has to be compact and the kernel k continuous.

3.3.2 Boundedness and Continuity

Because of the PD property and Cauchy-Schwarz inequality, there are relationships between the function $x \mapsto k(x, x)$ and $(x, y) \mapsto k(x, y)$ when one considers boundedness or continuity properties of the kernel.

Lemma 3.11 *For a PD kernel k the following two statements are equivalent*

(i) $x \mapsto k(x, x)$ *is bounded;*

(ii) $(x, y) \mapsto k(x, y)$ *is bounded.*

Lemma 3.12 *[56] A PD kernel k is continuous on $\mathcal{X} \times \mathcal{X}$ if and only if the following two conditions are fulfilled*

(i) $x \mapsto k(x, x)$ *is continuous;*

(ii) for any fixed x the function $y \mapsto k(x, y)$ is continuous at $y = x$.

These conditions are equivalent to the continuity of the function $(x, y) \mapsto k(x, y)$ at every point of the diagonal $\{(x, y) : x = y\}$.

Corollary 3.13 *If k is continuous on $\mathcal{X} \times \mathcal{X}$ then the identity map $(\mathcal{X}, d) \to (\mathcal{X}, d_k)$ is continuous.*

Proof: Follows directly from $d_k^2(x, x_n) = k(x, x) - 2k(x_n, x) + k(x_n, x_n)$. □

A related question is: when does the RKHS consist of continuous functions ? Since $k(x, \cdot)$ belongs to the associated RKHS, this means that k has to be at least separately continuous. The following theorem provides necessary and sufficient conditions in a rather general setting.

Theorem 3.14 *[83] Let $\mathcal{X}$ be a locally compact space and $C(\mathcal{X})$ the space of continuous functions on $\mathcal{X}$ with the topology of uniform convergence on compact subsets. The canonical injection $i : \mathcal{H}_k \to C(\mathcal{X})$ is continuous if and only if $k(x, y)$ is separately continuous on $\mathcal{X} \times \mathcal{X}$ and locally bounded.*

3.3.3 When is a Function in a RKHS ?

Let us suppose we are given a function f and want to know if it is contained in the RKHS associated to a PD kernel k. We give a general result.

Lemma 3.15 *[56] The function f belongs to the RKHS $\mathcal{H}$ associated to k if and only if there exists $\epsilon > 0$ such that*

$$R_\epsilon(x, y) = k(x, y) - \epsilon f(x) f(y)$$

is a positive definite kernel. Equivalently this corresponds to the condition

$$\sup_{|I|<\infty, (a_i)_{i\in I}\in\mathbb{R}, (x_i)_{i\in I}\in\mathcal{X}} \frac{\sum_{i\in I} a_i f(x_i)}{\left(\sum_{i,j\in I} a_i a_j k(x_i, x_j)\right)^{1/2}} < \infty.$$

If this is satisfied, one can compute the norm of f as the value of the above supremum, or as $\|f\|_{\mathcal{H}} = \inf\{1/\sqrt{\epsilon} \,|\, \epsilon > 0, R_\epsilon \succeq 0\}$.

A simple consequence of this lemma is that the RKHS associated to any bounded kernel cannot contain unbounded functions[9].

3.3.4 Separability of the RKHS

Some convergence proofs of iterative algorithms require the separability of the RKHS. However, this is seldom made explicit in the Machine Learning literature. The first result gives a necessary and sufficient condition for separability.

Theorem 3.16 *[62] $\mathcal{H}_k$ is separable if and only if $(\mathcal{X}, d_k)$ is separable.*

Proof: Let $\mathcal{H}_k$ be separable, then since $\mathcal{H}_k$ is a metric space $\mathcal{H}_k$ and every subset of $\mathcal{H}_k$ is second countable[10]. Particularly the set $k(\mathcal{X}, \cdot) := \{k(x, \cdot) \,|\, x \in \mathcal{X}\}$ is second countable and therefore separable. Since $(\mathcal{X}, d_k)$ is isometric to the set $k(\mathcal{X}, \cdot)$, $(\mathcal{X}, d_k)$ is separable.
We sketch the proof of the other direction. Since $(\mathcal{X}, d_k)$ is separable, $k(\mathcal{X}, \cdot)$ is separable. Then it is easy to show that the span of $k(\mathcal{X}, \cdot)$ with rational numbers is dense in $\operatorname{Span} k(\mathcal{X}, \cdot)$ and since $\mathcal{H}_k = \overline{\operatorname{Span} k(\mathcal{X}, \cdot)}$ we are done. □

In the case of continuous kernels we get the following consequence

Theorem 3.17 *[56] Let $\mathcal{X}$ be a topological space, k a PD kernel which is continuous on $\mathcal{X} \times \mathcal{X}$, and $\mathcal{H}$ its associated RKHS. If $\mathcal{X}$ is separable, then $\mathcal{H}$ is separable.*

As a result any continuous kernel on $\mathbb{R}^n$ induces a separable RKHS e.g. the RKHS associated to the RBF kernel $k(x, y) = \exp(-\|x - y\|/\sigma^2)$ is separable. In the case where $\mathcal{H}_k$ is separable the basis feature map can be written with a countable sum. Again, this does not require anything like Mercer's theorem.

[9] This can also be seen directly by $\|f\|_\infty = \sup_{x\in\mathcal{X}} |\langle k(x, \cdot), f\rangle_{\mathcal{H}}| \le \sup_{x\in\mathcal{X}} \sqrt{k(x, x)}\, \|f\|_{\mathcal{H}}$

[10] A topological space is second countable if it has a countable topological basis.

3.4 Integral and Covariance Operators

In general we assume in statistical learning theory that the space $\mathcal{X}$ is endowed with a probability measure P. Then samples X_i are drawn according to this probability measure P. These define then the empirical measure $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$.
In kernel-algorithms one uses the so-called *(normalized) kernel matrix* $K_n : L_2(\mathcal{X}, P_n) \to L_2(\mathcal{X}, P_n)$ defined as $K_n = \frac{1}{n}(k(X_i, X_j))_{i,j=1,\ldots,n}$ and the *empirical covariance operator* $C_n : \mathcal{H}_k \to \mathcal{H}_k$ defined as $C_n = \frac{1}{n}\sum_{k=1}^n \Phi(X_k) \otimes \Phi(X_k)$. These are under some conditions finite sample approximations of operators $K : L_2(\mathcal{X}, P) \to L_2(\mathcal{X}, P)$ resp. $C : \mathcal{H}_k \to \mathcal{H}_k$ defined for the whole probability measure P.
We will study the properties of the operators K and C and the convergence of the empirical counterparts to the true operators under the following assumptions on the kernel.

- $k(x, y)$ is measurable,
- $k(x, y)$ is a positive definite kernel,
- $\int_{\mathcal{X}} k(x, x)dP(x) < \infty$.

Note that the second assumption implies $k \in L_2(\mathcal{X} \times \mathcal{X}, P \otimes P)$ by the Cauchy-Schwarz inequality. Also note that in our setting we have no assumptions on the separability of $\mathcal{H}$ or $L_2(\mathcal{X}, P)$.

Theorem 3.18 *Let $i : \mathcal{H} \to L_2(\mathcal{X}, P)$ be the canonical injection. Then under the stated assumptions i is continuous. Moreover i is a Hilbert-Schmidt operator with $\|i\|_{HS}^2 \leq \int_{\mathcal{X}} k(x, x)dP(x)$.*

Proof: Let i be the canonical injection $i : \mathcal{H} \to L_2(\mathcal{X}, P)$. Then for all $f \in \mathcal{H}$,

$$\|if\|_{L_2(\mathcal{X},P)}^2 = \int |f(x)|^2 dP(x) = \int \langle f, k(x, \cdot)\rangle_{\mathcal{H}}^2 \, dP(x) \leq \|f\|_{\mathcal{H}}^2 \int k(x, x)dP(x)\,.$$

Therefore i is a bounded operator.
Denote by $\{e_\alpha,\ \alpha \in A\}$ an orthonormal basis (possibly uncountable) of $L_2(\mathcal{X}, P)$. i is Hilbert-Schmidt if and only if $\sum_{\alpha\in A} \|i\, e_\alpha\|_{L_2(\mathcal{X},P)}^2 < \infty$. For all finite sets $F \subset A$ we have

$$\begin{aligned}\sum_{\alpha\in F} \|ie_\alpha\|_{L_2(\mathcal{X},P)}^2 &= \int_{\mathcal{X}} \sum_{\alpha\in F} |e_\alpha(x)|^2 dP(x) = \int_{\mathcal{X}} \sum_{\alpha\in F} |\langle e_\alpha, k(x, \cdot)\rangle_{\mathcal{H}}|^2 \, dP(x) \\ &\leq \int_{\mathcal{X}} \|k(x, \cdot)\|_{\mathcal{H}}^2 \, dP(x) = \int_{\mathcal{X}} k(x, x)dP(x)\end{aligned}$$

where we have used Bessel's inequality. Let now $S_{fin}(A) = \{P \subset A \,|\, P \text{ finite}\}$ be the directed set of finite subsets of A with the set inclusion as partial order. Since all summands are positive, the limit of the net of partial sums can be computed as follows

$$\sum_{\alpha\in A} \|i\, e_\alpha\|_{L_2(\mathcal{X},P)}^2 = \sup\{\sum_{\alpha\in F} \|i\, e_\alpha\|_{L_2(\mathcal{X},P)}^2\,,\ F \in S_{fin}(A)\} \leq \int_{\mathcal{X}} k(x, x)dP(x).$$

□

The next proposition connects the canonical injection i with the integral and the covariance operator:

Proposition 3.19 *The integral operator K*

$$K : L_2(\mathcal{X}, P) \to L_2(\mathcal{X}, P),\ (Kf)(x) = \int_{\mathcal{X}} k(x, y) f(y) dP(y)\,. \tag{3.3}$$

and the covariance operator C

$$C : \mathcal{H} \to \mathcal{H},\ \langle f, Cg\rangle = \int_{\mathcal{X}} f(x) g(x) dP(x)\,. \tag{3.4}$$

are both positive, self-adjoint, Hilbert-Schmidt, and trace-class. Moreover they can be decomposed as $K = ii^*$ and $C = i^*i$ and have the same spectrum which implies that $\operatorname{tr} K = \operatorname{tr} C$ and $\|C\|_{HS} = \|K\|_{HS} = \|k\|_{L_2(\mathcal{X}\times\mathcal{X}, P\otimes P)}$.

Proof: We showed in theorem 3.18 that i is continuous. Therefore the adjoint $i^* : L_2(\mathcal{X}, P) \to \mathcal{H}$ exists and is defined for $g \in L_2(\mathcal{X}, P)$ and $f \in \mathcal{H}$ as $\langle i^* g, f\rangle_{\mathcal{H}} = \langle g, if\rangle_{L_2(\mathcal{X},P)}$. In particular, choosing $f = k(x, \cdot) \in \mathcal{H}$, we see that

$$(i^* g)(x) = \langle k(x, \cdot), i^* g\rangle_{\mathcal{H}} = \langle ik(x, \cdot), g\rangle = \int_{\mathcal{X}} k(x, y) g(y) dP(y),$$

so that $K = ii^*$. As a consequence K is positive and self-adjoint. Moreover it is trace-class since

$$\operatorname{tr} K = \sum_{\alpha\in A} \langle e_\alpha, K\, e_\alpha\rangle_{L_2(\mathcal{X},P)} = \sum_{\alpha\in A} \|i^* e_\alpha\|_{\mathcal{H}} = \|i^*\|^2_{HS} \le \int_{\mathcal{X}} k(x, x) dP(x),$$

where we use the fact $\|i\|_{HS} = \|i^*\|_{HS}$.
Moreover, for $f, g \in \mathcal{H}$, $\langle f, i^* ig\rangle_{\mathcal{H}} = \langle if, ig\rangle_{L_2(\mathcal{X},P)} = \mathbb{E}\,[f(X)g(X)]$ so that C is positive, self-adjoint and $C = i^*i$. It follows easily that C is trace-class with

$$\operatorname{tr} C = \sum_{\alpha\in A} \langle e_\alpha, C\, e_\alpha\rangle_{\mathcal{H}} = \sum_{\alpha\in A} \|i\, e_\alpha\|^2_{L_2(\mathcal{X},P)} = \|i\|^2_{HS}\,.$$

Both C and K are trace-class and therefore compact which implies that they only have a discrete spectrum. Moreover they have the same spectrum and all non-zero eigenvalues have the same multiplicity. Let λ_n be an eigenvalue of K and denote by Λ_n the corresponding finite-dimensional eigenspace. Then

$$ii^*\Lambda_n = \lambda_n \Lambda_n \Rightarrow (i^*i) i^*\Lambda_n = \lambda_n i^*\Lambda_n \tag{3.5}$$

that is $i^*\Lambda_n$ is an eigenspace of C to the corresponding eigenvalue λ_n and the same argumentation holds in the other direction. Also $\dim \Lambda_n = \dim i^*(\Lambda_n)$ since it follows from (3.5) that $\Lambda_n \nsubseteq Ker(i^*)$ and $i^*(\Lambda_n) \nsubseteq Ker(i)$.
It is a classical result that $k \in L_2(\mathcal{X}\times\mathcal{X}, P\otimes P)$ implies that K is Hilbert-Schmidt and $\|K\|_{HS} = \|k\|_{L_2(\mathcal{X}\times\mathcal{X}, P\otimes P)}$, see [26] (note that this is true, even if $L_2(\mathcal{X}, P)$ is not separable). Since a compact self-adjoint operator is Hilbert-Schmidt if and only if $\sum_i \lambda_i^2 < \infty$ it follows directly from the equality of the spectra that C is Hilbert-Schmidt with $\|C\|_{HS} = \|K\|_{HS} = \|k\|_{L_2(\mathcal{X}\times\mathcal{X}, P\otimes P)}$. □

Corollary 3.20 *If $Ker(i) = 0$ then $\mathcal{H} = \overline{i^*(L_2(\mathcal{X}, P))}$ and $\mathcal{H}$ is separable.*

Proof: If $Ker(i) = 0$ then $\overline{Ran(i^*)} = Ker(i)^\perp = \mathcal{H}$. Since i^* is compact, $\overline{Ran(i^*)}$ is separable and therefore $\mathcal{H}$ is separable. □

In other words if the zero function is the only function in the RKHS $\mathcal{H}$ which is zero P-almost everywhere, then the image of the integral operator K is dense in the RKHS and the RKHS is automatically separable.

Corollary 3.21 *If $\mathcal{H}$ is separable then $\|i\|^2_{HS} = \operatorname{tr} C = \operatorname{tr} K = \int_{\mathcal{X}} k(x,x)dP(x)$.*

Proof: Let $\{e_n\}_{n=1}^{\infty}$ be a complete orthonormal basis of $\mathcal{H}$. Then

$$\begin{aligned}
\|i\|^2_{HS} &= \lim_{N\to\infty} \sum_{n=1}^{N} \|i\, e_n\|^2_{L_2(\mathcal{X},P)} = \lim_{N\to\infty} \sum_{n=1}^{N} \int_{\mathcal{X}} |e_n(x)|^2 dP(x) \\
&= \lim_{N\to\infty} \sum_{n=1}^{N} \int_{\mathcal{X}} |\, \langle k(x,\cdot), e_n\rangle_{\mathcal{H}}\,|^2 dP(x) = \int_{\mathcal{X}} \lim_{N\to\infty} \sum_{n=1}^{N} |\, \langle k(x,\cdot), e_n\rangle_{\mathcal{H}}\,|^2 dP(x) \\
&= \int_{\mathcal{X}} \|k(x,\cdot)\|^2_{\mathcal{H}}\, dP(x) = \int_{\mathcal{X}} k(x,x) dP(x) < \infty
\end{aligned}$$

where the fourth step follows from the monotone convergence theorem and the fifth step is Parseval's identity. □

The next corollary establishes a feature map in $L_2(\mathcal{X}, P)$.

Corollary 3.22 *If $k \in L_2(\mathcal{X}\times\mathcal{X}, P\otimes P)$, then there exists an orthonormal system (ϕ_n) in $L_2(P)$ such that*

$$k(x,y) = \sum_{n\in\mathbb{N}} \lambda_n \phi_n(x)\phi_n(y)\,, \tag{3.6}$$

where $\lambda_n \geq 0$ and the convergence of the sum occurs in $L_2(\mathcal{X}\times\mathcal{X}, P\otimes P)$. The associated feature map is thus

$$\Phi(x) = (\sqrt{\lambda_n}\phi_n(x))_{n\in\mathbb{N}}\,.$$

Proof: That is a classical result in functional analysis, see e.g. [73]. □

The remaining question is how the empirical counterparts K_n and C_n are related to the operators K and C.

Proposition 3.23 *Let K be the integral operator defined in (3.3) and X_i a set of i.i.d. random variables drawn from P. For all $f \in L_2(\mathcal{X}, P)$ we have:*

$$\begin{aligned}
\lim_{n\to\infty} \langle f, Kf\rangle_{L_2(\mathcal{X},P_n)} &= \lim_{n\to\infty} n^{-2} \sum_{i,j=1}^{n} f(X_i) f(X_j) k(X_i, X_j) \\
&= \int_{\mathcal{X}^2} f(x) f(y) k(x,y) dP(x) dP(y) = \langle f, Kf\rangle_{L_2(\mathcal{X},P)} \quad a.s.
\end{aligned}$$

Proof: The proof is essentially an application of a result in [36]. Given an i.i.d. set of random variables $X_i \in \mathcal{X}$ drawn from P and a measurable symmetric function $g(x,y) : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ it states that $\lim_{n\to\infty} n^{-2}\sum_{i,j=1}^{n} g(X_i, X_j) = Eg(X,Y)$ almost surely if $E|g(X,Y)| < \infty$ and $E\sqrt{|g(X,X)|} < \infty$. Let now $g(x,y) =$

$f(x)f(y)k(x,y)$ for some realization f of the equivalence class in $L_2(\mathcal{X}, P)$. The choice of the realization does not matter since the functions differ on a set of measure zero. Then the conditions require that $\int_{\mathcal{X}^2} f(x)f(y)k(x,y)dP(x)dP(y) < \infty$ and $\int_{\mathcal{X}} |f(x)|\sqrt{k(x,x)}dP(x) < \infty$. The second condition implies the first one and we have

$$\int_{\mathcal{X}} |f(x)|\sqrt{k(x,x)}dP(x) \leq \int_{\mathcal{X}} |f(x)|^2 dP(x) \int_X k(x,x)dP(x) < \infty,$$

since $\|f\|^2_{L_2(\mathcal{X},P)} \leq \|f\|^2_{\mathcal{H}} \int_{\mathcal{X}} k(x,x)dP(x)$. □

The next statement relates C_n and C:

Proposition 3.24

$$\langle f, C_n g\rangle_{\mathcal{H}_k} \xrightarrow{a.s.} \langle f, Cg\rangle_{\mathcal{H}_k}, \ \forall\, f, g \in \mathcal{H}_k.$$

Proof: The proof is a simple application of the strong law of large numbers. □

As a final remark we would like to note that if k is bounded then all the assumptions are fulfilled and the theorems of this section apply for any probability measure P.

3.5 Generalizations

Now that we have the general picture in mind, we investigate possible generalizations of the presented notions. We consider the generalization of kernel functions to operator-valued functions and of the RKHS to Hilbertian subspaces. We will show that they are both special cases of the general theory above. Finally we will consider the only real generalization the theory of reproducing kernel spaces with indefinite inner product.

3.5.1 Operator-Valued Kernels

Recently there was interest in the machine learning community to extend real-valued kernels to operator-valued kernels in order to learn vector-valued functions [66]. This concept is not new in the mathematics literature. It can at least traced back to the paper of [28].
Let $\mathcal{X}$ be a set and $\mathcal{G}$ a Hilbert space[11]. The goal is to generate a (generalized) RKHS whose functions are from $\mathcal{X}$ to $\mathcal{G}$ (instead of $\mathcal{X} \to \mathbb{R}$). We define a (generalized) notion of positive definite kernel:

Definition 3.25 *A function $k : \mathcal{X} \times \mathcal{X} \to L(\mathcal{G})$[12] such that $k(x,y) = k(y,x)^*$ is called a* **positive definite operator-valued kernel function** *if for all $n \geq 1$, $x_1, \ldots, x_n \in \mathcal{X}$, $c_1, \ldots, c_n \in \mathcal{G}$, $\sum_{i,j=1}^n \langle c_i\, k(x_i, x_j), c_j\rangle \geq 0$*

This seems to generalize the PD kernels we introduced before, and indeed, several papers deal with the notion of operator-valued kernels. However, a slight change of point of view allows to recast operator-valued kernels in the standard setting of real-valued ones, showing their great generality. We have the following result.

[11] the same theory can be developed for Banach spaces or locally convex spaces.
[12] set of bounded linear operators on $\mathcal{G}$

Proposition 3.26 *Let k be a PD operator-valued kernel $\mathcal{X} \times \mathcal{X} \to L(\mathcal{G})$. Define ℓ as the function on $(\mathcal{X} \times \mathcal{G})$ such that $\ell((x,f),(y,g)) = \langle f, k(x,y)g\rangle_{\mathcal{G}}$. The map $k \mapsto \ell$ thus defined, is a bijection between PD operator-valued kernels $\mathcal{X} \times \mathcal{X} \to L(\mathcal{G})$ and real-valued PD kernels $(\mathcal{X},\mathcal{G}) \times (\mathcal{X},\mathcal{G}) \to \mathbb{R}$ which are bilinear on $\mathcal{G} \times \mathcal{G}$[13]. If $\mathcal{G}$ is finite dimensional, dim $\mathcal{G} = d$, one can also define, (e_i) being an orthonormal basis of $\mathcal{G}$, $\ell((x,i),(y,j)) = \langle e_i, k(x,y)e_j\rangle$, such that $k \mapsto \ell$ is a bijection to real-valued PD kernels on $(\mathcal{X}, \{1,\ldots,d\})$.*

Proof: We prove the above proposition in the finite dimensional case (the general case has a similar proof). Let $\ell((x,i),(y,j)$ be a PD kernel on $(\mathcal{X}, \{1,\ldots,d\})$. Define a bilinear form on $\mathbb{R}^d$ by defining the matrix $k(x,y) : \mathbb{R}^d \to \mathbb{R}^d$ as

$$k_{ij}(x,y) = \langle e_i, k(x,y)e_j\rangle_{\mathbb{R}^d} = \ell((x,i),(y,j))$$

where e_i denotes a basis in $\mathbb{R}^d$. Conversely given the PD operator valued kernel $k(x,y)$, define by the above expression the kernel function $\ell((x,i),(y,j))$. Then we have with $v_m \in \{1,\ldots,d\}$

$$\begin{aligned}
\sum_{i,j=1}^{n}\sum_{m,n=1}^{d} \alpha_{im}\alpha_{jn}\ell((x_i,v_m),(x_j,v_n)) &= \sum_{i,j=1}^{n}\sum_{m,n=1}^{d} \alpha_{im}\alpha_{jn}k_{v_m v_n}(x_i,x_j) \\
&= \sum_{i,j=1}^{n}\left\langle \sum_{m=1}^{d}\alpha_{im}e_{v_m}, k(x_i,x_j)\sum_{n=1}^{d}\alpha_{jn}e_{v_n}\right\rangle \\
&= \sum_{i,j=1}^{n}\langle c_i, k(x_i,x_j)c_j\rangle
\end{aligned}$$

with $c_i = \sum_{m=1}^{d}\alpha_{im}e_{v_m}$. Now if ℓ is positive definite then consider the index set of size nd given by $z_{im} = (x_i, v_m)$ which gives the above expression and implies that $k(x,y)$ is a PD operator-valued kernel, since we can express any vector $c \in \mathbb{R}^d$ in the form $\sum_{m=1}^{d}\alpha_m e_{v_m}$. Conversely let $k(x,y)$ be a PD operator-valued kernel and take as vectors $c_i = \alpha_i e_{v_i}$, then $\ell((x,i),(y,j))$ is a PD kernel function since we can express all index sets in the form $z_i = (x_i, v_i)$. □

The meaning of the above proposition is that at the price of changing the index set, one can simply work with real-valued kernels, and the positive definiteness of these kernels implies the positivity of the corresponding operator valued kernels. Moreover one can use the properties of the real-valued kernels to derive the properties of the operator-valued one.

3.5.2 Hilbertian Subspaces

Instead of trying to generalize the PD kernels, one may, as in the work of Schwartz [83], generalize the notion of RKHS and kernel operator. The idea is to consider instead of Hilbert spaces of real-valued functions, that is a Hilbertian subspace of $\mathbb{R}^{\mathcal{X}}$, subspaces of quit general spaces equipped with the structure of a Hilbert space that may not even contain functions. The framework of Schwartz is formulated in the very general setting of locally convex topological vector spaces (l.c.s.), see [73, 76] for an introduction. Note that $\mathbb{R}^{\mathcal{X}}$ with the topology of pointwise convergence is

[13] i.e. $k((x,g_1),(y,g_2))$ is bilinear in g_1, g_2.

a complete l.c.s.. This topology is equivalent to the weak topology induced by the duality map $\langle \cdot, \cdot \rangle_{\mathbb{R}^{[\mathcal{X}]}, \mathbb{R}^{\mathcal{X}}}$ defined above. In the following E denotes a complete l.c.s.

Definition 3.27 *A linear subspace $\mathcal{H} \subset E$ is called a* **Hilbertian subspace** *if*

(i) it is provided with $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\mathcal{H}$ is a Hilbert space.

(ii) The injection of $\mathcal{H}$ into E is continuous; that is convergence in $\mathcal{H}$ implies convergence in E.

Definition 3.28 *A* **kernel operator** *K is a linear, symmetric map*[14] *from E'*[15] *into E. K is said to be* **positive** *if for all $e' \in E'$, $\langle e', Ke' \rangle_{E',E} \geq 0$.*

The following theorem gives the analogue of the bijection between positive definite kernels and RKHS.

Theorem 3.29 *[83] There is a one-to-one correspondence between the closed convex cone of Hilbertian subspaces $\mathcal{H}$ and the positive kernel operators K. To $\mathcal{H}$ corresponds the kernel operator $K = j \circ \theta \circ j'$, where $j : \mathcal{H} \to E$ is the natural injection, $j' : E' \to \mathcal{H}'$ its adjoint and $\theta : \mathcal{H}' \to \mathcal{H}$ the canonical isomorphism. Moreover given a positive kernel operator K, the Hilbert space is given by $\mathcal{H} = \overline{KE'}$ with the inner product on KE' defined as*

$$\langle Ke', Kf' \rangle_{\mathcal{H}} = \langle e', Kf' \rangle_{E',E} .$$

The inner product in $\mathcal{H}$ defined in the above way 'reproduces' the value of e' on any element of E contained in $\mathcal{H}$.

Example: [Hilbertian subspaces of $\mathbb{R}^{\mathcal{X}}$] We have defined in a previous section a positive symmetric kernel operator $K : \mathbb{R}^{[\mathcal{X}]} \to \mathbb{R}^{\mathcal{X}}$. Since $\mathbb{R}^{\mathcal{X}}$ is a complete l.c.s., K is also a positive kernel operator in the sense of Schwartz. Additionally by Theorem 3.10 the associated reproducing kernel Hilbert spaces are Hilbertian subspaces of $\mathbb{R}^{\mathcal{X}}$.

So one recovers the standard RKHS as a special case of Schwartz's theory, see [83]. The setting of Schwartz seems at first much to general for machine learning tasks. However as we will see soon it provides us with the right setting to deal with distribution valued kernels which is a generalization of the usual kernel function. One could ask at this point why it is a good idea to consider kernels on functions instead of kernels on points. One can argue that because of the limited precision of the measurement device measurements of real-valued physical quantities can never be made with arbitrary precision. This measurement error can be modelled by considering, instead of points, functions with compact support which are concentrated on the measured points. The width of the function then models the uncertainty in the measurement. This means we smear the points before we compare them with the kernel function. The following famous theorem characterizes the form of the kernel operator when one considers Hilbertian subspaces of distributions.

Theorem 3.30 (Schwartz kernel theorem) *The topological vector space of continuous linear maps $D(\mathbb{R}^n) \to D'(\mathbb{R}^n)$*[16]*, with the strong topology, is canonically isomorphic to the topological vector space $D'(\mathbb{R}^n \times \mathbb{R}^n)$.*

[14] Note that a linear, symmetric map is weakly continuous.

[15] E' denotes the topological dual space of E.

[16] $D'(\mathbb{R}^n)$ denotes the distributions on $\mathbb{R}^n$ and $D(\mathbb{R}^n)$ the space of smooth functions on $\mathbb{R}^n$ with compact support with the strict inductive limit topology.

This theorem guarantees that we have again a unique correspondence between the kernel operator and a generalized kernel function as in the case of usual positive definite kernels. Indeed, in the abstract framework of Hilbertian subspaces, it is not clear that a function of two variables is naturally associated to a subspace. However, thanks to this result, it is true in the case of Hilbertian subspaces of distributions: they are naturally associated to a (generalized) kernel function which is actually a distribution on $\mathbb{R}^n \times \mathbb{R}^n$. We give a simple yet illustrative example of this phenomenon.

Example: [$L_2(\mathbb{R}^n)$ as a Hilbertian subspace of $D'(\mathbb{R}^n)$] Let $K = \delta(x-y) \in D'(\mathbb{R}^n \times \mathbb{R}^n)$. Then we have for all $f \in D(\mathbb{R}^n)$, $(Kf)(x) = \int_{\mathbb{R}^n} \delta(x-y)f(y) = f(x)$ and the inner product on $KD(\mathbb{R}^n)$ is defined as:

$$\langle Kf, Kg \rangle := \langle Kf, g \rangle_{D'(\mathbb{R}^n), D(\mathbb{R}^n)} = \int_{\mathbb{R}^n} f(x)g(x)dx.$$

Since $D(\mathbb{R}^n)$ is dense in $L^2(\mathbb{R}^n)$ and the above inner product induces an isometry between $KD(\mathbb{R}^n)$ and $L^2(\mathbb{R}^n)$ restricted to $D(\mathbb{R}^n)$ we get the desired result that $L^2(\mathbb{R}^n)$ is isometrically isomorphic to the Hilbertian subspace $\overline{KD(\mathbb{R}^n)} \subset D'(\mathbb{R}^n)$.

Remark: The example on Hilbertian subspaces of $\mathbb{R}^{\mathcal{X}}$ suggests that the framework of Hilbertian subspaces is a generalization of the Aronszajn framework of RKHS. But one can always see the elements of the Hilbertian subspace $H \subset E$ as linear functions on the dual E' acting via $h(e') = \langle e', h \rangle_{E',E}$. So $\mathcal{H}$ can be considered as a Hilbertian subspace of $\mathbb{R}^{E'}$. Since E' must have a special structure, whereas the Aronszajn approach works for any set $\mathcal{X}$, from this point of view Hilbertian subspaces are actually less general. For example the framework of distributions can be seen as a RKHS on $\mathbb{R}^{\mathcal{X}}$. The problem of the Aronszajn approach is that the special properties of the underlying set $\mathcal{X}$ play no role and are 'forgotten'. It seems that from the structural point of view the framework of Schwartz is better, from the practical point of view the framework of Aronszajn is maybe easier to handle.

3.5.3 The General Indefinite Case

It is generally not easy to check if a given symmetric function is a positive definite kernel. In some cases like $k(x,y) = \tanh(\alpha \langle x, y \rangle + \beta)$ it is even known that the associated kernel matrix can have negative eigenvalues. Nevertheless it is sometimes used in support vector machines. Naturally the question arises if there still exists something like reproducing kernel spaces, such that we can interpret this non-positive definite kernel as an indefinite inner product in these space. The theory of reproducing kernel spaces with indefinite inner products was to our knowledge first explored by Schwartz [83] in the framework of Hermitian subspaces. A more explicit treatment following Aronszajn was done by Sorjonen [88].

Reproducing Kernel Pontryagin Spaces

Definition 3.31 *A symmetric kernel function $K(s,t) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to have κ **negative squares**, κ a nonnegative integer, if $\forall\, n \geq 1$, and all $x_1, \ldots, x_n \in \mathcal{X}$ the matrix $(k(x_i, x_j)_{i,j=1,\ldots,n})$ has at most κ negative eigenvalues and at least one such matrix has exactly κ negative eigenvalues.*

Now we define a generalization of Hilbert spaces.

Definition 3.32 *A **Krein space** is an inner product space $\mathcal{H}$, which can be written as the orthogonal sum $\mathcal{H} = \mathcal{H}_+ \oplus \mathcal{H}_-$ of a Hilbert space $\mathcal{H}_+$ and the antispace*[17] *$\mathcal{H}_-$ of a Hilbert space. If the antispace $\mathcal{H}_-$ is finite dimensional then $\mathcal{H}$ is called **Pontryagin space**.*

This decomposition is not unique, but the resulting spaces are all isomorphic. The dimensions of $\mathcal{H}_\pm$ are independent of the choice of the decomposition and are called positive and negative indices of $\mathcal{H}$.

Definition 3.33 *A **reproducing kernel Pontryagin space (RKPS)** $\mathcal{H}$ on $\mathcal{X}$ is a Pontryagin space of functions from $\mathcal{X}$ to $\mathbb{R}$ with a reproducing kernel $k(x, y)$ on $\mathcal{X} \times \mathcal{X}$ such that*

$$\begin{aligned} &\forall\, x \in \mathcal{X}, \quad k(x, \cdot) \in \mathcal{H}, \\ &\forall\, f \in \mathcal{H}, \quad \langle f(\cdot), k(x, \cdot)\rangle_{\mathcal{H}} = f(x). \end{aligned}$$

The RKPS are very similar in their structure as the following two theorems show.

Theorem 3.34 *[88] A Pontryagin space $\mathcal{H}$ of real-valued functions on Ω admits a reproducing kernel $K(s, t)$ if and only if all evaluation functionals are continuous. In this case $K(s, t)$ is unique, and it is a Hermitian kernel having κ negative squares, where k is the negative index of $\mathcal{H}$.*

Theorem 3.35 *[88] If $K(s, t)$ is a Hermitian kernel on $\mathcal{X} \times \mathcal{X}$ having κ negative squares, then there is a unique Pontryagin space $\mathcal{H}$ of functions on $\mathcal{X}$ with $\dim \mathcal{H}^- = \kappa$ having $K(s, t)$ as reproducing kernel.*

Reproducing Kernel Krein Spaces

The following theorem gives necessary and sufficient conditions for a symmetric function to be a reproducing kernel of a Krein space.

Theorem 3.36 *[83] If $k(x, y)$, $x, y \in \mathcal{X}$, is a symmetric function with values in $\mathbb{R}$, the following assertions are equivalent*

(i) k is the reproducing kernel of a Krein space $\mathcal{H}_k$ of functions on $\mathcal{X}$.

(ii) There exists an $\ell \in \mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$ such that $-\ell \preceq k \preceq \ell$.

(iii) $k = k_+ - k_-$ for some $k_+, k_- \in \mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$.

Unfortunately there exist counterexamples of symmetric functions which do not fulfill these conditions, but when the above conditions are satisfied, the reproducing kernel Krein space (RKKS) is characterized in the following way.

Proposition 3.37 *[83] If $k = k_+ - k_-$ with $k_+, k_- \in \mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$, then one can choose k_+ and k_- such that the associated RKHS of k_+ and k_-, $\mathcal{H}_+$ respectively $\mathcal{H}_-$, fulfill $\mathcal{H}_+ \bigcap \mathcal{H}_- = \{0\}$. In this case the RKKS associated to k consists of the functions $f = f_+ + f_-$, $f_+ \in \mathcal{H}_+$, $f_- \in \mathcal{H}_-$ with the indefinite inner product $[f, g] = \langle f_+, g_+\rangle_{\mathcal{H}_+} - \langle f_-, g_-\rangle_{\mathcal{H}_-}$.*

[17] An antispace of a Hilbert space is $(\mathcal{H}, \langle\cdot,\cdot\rangle)_{\mathcal{H}}$ is given by $(\mathcal{H}, -\langle\cdot,\cdot\rangle_{\mathcal{H}})$.

3.6 Conclusion

We have tried to extract from the mathematical literature on kernels the basic facts that are relevant to researchers in Machine Learning working with kernel methods. In particular, if one wants to develop generalizations of these concepts, it should be clear that there already exist several points of view for such generalizations and that, changing the point of view they can be cast in the same framework.
Finally, this chapter is far from being a complete overview on kernels. There exist many other notions which could be explored e.g. Gaussian measures, generalized stochastic processes, group representations in RKHS, spectral decompositions of kernels, regularization theory and various results of applications in approximation, interpolation etc.

3.7 Appendix

3.7.1 Structures Associated to a Gaussian Process

In this section we introduce extra objects that are naturally associated to a Gaussian process (hence to a PD kernel). We refer to [49] for additional details.
We denote by E a locally convex space.

Definition 3.38 *A Borel probability measure μ on E is a* **Gaussian measure** *if each $e' \in E'$, regarded as a random variable defined on the probability space (E, μ) is Gaussian.*

Definition 3.39 *A random variable X with values in E is a* **Gaussian vector** *if the real-valued random variable $\langle e', X\rangle_{E',E}$ is Gaussian for every $e' \in E'$, or equivalently, if the distribution of X is a Gaussian measure on E.*

Theorem 3.40 (Kolmogorov extension theorem) *Let $\Omega = \mathbb{R}^{\mathcal{X}}$, where $\mathcal{X}$ is an arbitrary index set, and let $\mathcal{F}$ be the product σ-field $\mathcal{B}^{\mathcal{X}}$ on Ω. Suppose that for every finite subset $\mathcal{Y} \subseteq \mathcal{X}$, we are given a (consistent) probability measure $P_{\mathcal{Y}}$ on $\mathbb{R}^{\mathcal{Y}}$; then there exists a unique probability measure on $\mathbb{R}^{\mathcal{X}}$ such that the projection onto $\mathbb{R}^{\mathcal{Y}}$ induces $P_{\mathcal{Y}}$ for every finite $\mathcal{Y}$.*

It follows from this theorem that all the objects introduced before are tightly related.

Proposition 3.41 *Every Gaussian process $(X_x)_{x\in\mathcal{X}}$ defines a unique Gaussian measure on $\mathbb{R}^{\mathcal{X}}$ and a unique random vector X with values in $\mathbb{R}^{\mathcal{X}}$.*

We now give the construction of a feature map via the Kolmogorov theorem [70]. Given a PD kernel k on $\mathcal{X}$ define for any finite subset $\mathcal{Y} = x_1, \ldots, x_n$ a probability measure which is centered Gaussian and has covariance matrix $(k(x_i, x_j))_{i,j}$. By Theorem 3.40 there exists a measure μ on $\mathbb{R}^{\mathcal{X}}$ and it is Gaussian. If we consider the Hilbert space $L_2(\mathbb{R}^{\mathcal{X}}, \mu)$ and define $X_x := f(x)$, $f \in \mathbb{R}^{\mathcal{X}}$ (where f has the distribution μ), then X_x is an element of $L_2(\mathbb{R}^{\mathcal{X}}, \mu)$ and $\mathbb{E}[X_x X_y] = \int f(x)f(y)d\mu(f) = k(x, y)$. Moreover one can check that the completion of the subspace of Gaussian random variables X_x in $L_2(\mathbb{R}^{\mathcal{X}}, \mu)$ still consists only of Gaussian random variables. Therefore it is called **Gaussian Hilbert space**. It is shown in Janson [49] that the Gaussian Hilbert space is isometric isomorphic to the RKHS associated to the PD kernel k.

Kapitel 4

Maximal Margin Classification in Metric Spaces

4.1 Introduction

If the only knowledge available to the statistician is that the data comes from a semi-metric space $(\mathcal{X}, d)$, where $\mathcal{X}$ is the input space and d is the corresponding semi-metric, it is reasonable to assume, for a classification task, that the class labels are somewhat related to the semi-metric. More precisely, since one has to make assumptions about the structure of the data (otherwise no generalization is possible), it is natural to assume that two points that are close (as measured by d) are likely to belong to the same class, while points that are far away may belong to different classes. Another way to express this assumed relationship between class membership and distances is to say that intra-class distances are on average smaller than inter-class distances.

Most classical classification algorithms rely, implicitly or explicitly, on such an assumption. On the other hand, it is not always possible to work directly in the space $\mathcal{X}$ where the data lies. In particular, some algorithms require a vector space structure (e.g. linear algorithms) or at least a feature representation (e.g. decision trees). So, if $\mathcal{X}$ does not have such a structure (e.g. if the elements of $\mathcal{X}$ are DNA sequences of variable length, or descriptions of the structure of proteins), it is typical to construct a new representation (usually as vectors) of the data. In this process, the distance between the data, that is the (semi)-metric, is usually altered. But with the above assumptions on the classification task this change means that information is lost or at least distorted.

It is thus desirable to avoid any distortion of the (semi)-metric in the process of constructing a new representation of the data. Or at least, the distortion should be consistent with the assumptions. For example a transformation which leaves the small distances unchanged and alters the large distances, is likely to preserve the relationship between distances and class membership. We later propose a precise formulation of this type of transformation.

Once the data is mapped into a vector space, there are several possible algorithms that can be used. However, there is one heuristic which has proven valuable both in terms of computational expense and in terms of generalization performance, it is the maximum margin heuristic. The idea of maximum margin algorithms is to look for a linear hyperplane as the decision function which separates the data with

maximum margin, i.e. such that the hyperplane is as far as possible from the data of the two classes. This is sometimes called the hard margin case. It assumes that the classes are well separated. In general one can always deal with the inseparable case by introducing slack variables, which corresponds to the soft margin case.
Our goal is to apply this heuristic to $(\mathcal{X}, d)$, the (semi)-metric input space directly. To do so, we proceed in two steps: we first embed $\mathcal{X}$ into a Banach space (i.e. a normed vector space which is complete with respect to its norm) and look for a maximum margin hyperplane in this space. The important part being that the embedding we apply is isometric, that is, all distances are preserved.
We explain how to construct such an embedding and show that the resulting algorithm can be approximated by the Linear Programming Machine proposed by Graepel *et al.* [37]. We also propose to use as a pre-processingßtep, a transformation of the metric which has the properties mentioned above (i.e. leaving the small distances unaltered and affecting the large ones) which may remove the unnecessary information contained in large distances and hence give a better result when combined with the above mentioned algorithm.
Embedding the data isometrically into a Banach space is convenient since it is possible for any metric space. But as we will show it has also the disadvantage that the obtained maximum margin algorithm cannot be directly implemented and has to be approximated. It may thus be desirable that the space into which the data is embedded has more structure. A natural choice is to use a Hilbert space (i.e. a Banach space where the norm is derived from an inner product). However, we recall a result of Schoenberg, see [80], which states that only a certain class of metric spaces can be isometrically embedded into a Hilbert space. Hence, we gain structure at the price of loosing generality. Moreover, we give a characterization of metric spaces that can be embedded into a Hilbert space with some distortion of the large distances. If the metric has the appropriate properties, we thus also derive an embedding into a Hilbert space and the corresponding maximal margin algorithm.
It turns out that the obtained algorithm is equivalent to the well-known Support Vector Machine (e.g. [81]). We thus obtain a new point of view on this algorithm which is based on an isometric embedding of the input space as a metric space, where the metric is induced by a kernel. However, the main distinction between our point of view and the more classical one, is that we show that the solution only depends on the metric induced by the kernel and not on the kernel itself. And given this metric, the effect of the algorithm is to perform maximal margin separation after an *isometric* embedding into a Hilbert space.

Finally we investigate the properties of the class of functions that are associated with these embeddings. In particular we want to measure their capacity. For that we use a (by now) standard measure of the size in learning theory, the Rademacher averages. These can be directly related to the generalization error of the algorithm. Our computations show that in the case of the Banach space embedding of an arbitrary metric space, the size of the obtained class of hypotheses is the same as the size of $(\mathcal{X}, d)$ itself as a metric space, where the size is measured by the covering numbers. For the second embedding into a Hilbert space, we get results similar to the previously known ones for SVM, but we express them in terms of the induced (semi)-metric. Finally, in the case where $\mathcal{X}$ can be embedded isometrically both in a Banach and a Hilbert space, we compare the capacities of both obtained hypotheses classes and show that the SVM algorithm corresponds to a more parsimoniousßpace

of functions.

The paper is organized as follows. Section 4.2 introduces the general approach of embedding into a Banach space and performing maximum margin classification in this space. In particular, several possible embeddings with their effects on the metric are discussed. In Section 4.3 this approach is applied to an arbitrary metric space and we give the resulting general algorithm. Then, section 4.4 deals with the special case of metric spaces that can be isometrically embedded into a Hilbert space. These metrics are characterized and we derive, with our general approach, an algorithm which turns out to be equivalent to the SVM algorithm. Finally in section 4.5, we compute Rademacher averages corresponding to the previously mentioned algorithms and compare them.
Part of this chapter has been published in [43, 42].

4.2 The general approach

We are working in the following setting. We are given a set $\mathcal{X}$, together with a (semi)-metric defined on it, which makes it a (semi-)metric space $(\mathcal{X}, d)$. Recall that a semi-metric is a non-negative symmetric function, $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which satisfies the triangle inequality and $d(x, x) = 0$ for all $x \in \mathcal{X}$ (it is a metric if $d(x, y) = 0$ implies $x = y$).
Remark: In the following we will consider only metric spaces. But all the results remain true for semi-metric spaces. The reason why we restrict ourselves to metric spaces is on the one hand simplicity but on the other hand the in general undesired implications of a semi-metric, see Appendix 4.7.1 for this issue.

Our basic assumption is that this metric is consistent with the classification problem to be solved in the sense that when two points are close, they are likely to belong to the same class. Of course, there are many algorithms that can take into account such an assumption to build a classifier (e.g. nearest neighbors classifiers). Moreover if one has more structure than the pure metric space e.g. when $\mathcal{X}$ is a differentiable manifold, then this knowledge should be used in the classifier. In the sense that one should build functions which satisfy stronger smoothness requirements. One could argue that then the approach presented here is too general since at first sight we only use the metric structure of the input space. However as we will show later the functions used by the general maximal margin algorithm are always Lipschitz functions which can be regarded as the lowest level of smoothness. Moreover if the metric has stronger smoothness properties e.g. in the case of a Riemannian manifold then these smoothness properties are also transferred to the associated function space used by the maximal margin classifier. This will become obvious from the form of embedding we use. In that sense the maximal margin algorithm adapts to the smoothness of $\mathcal{X}$.

One of the cornerstones of the algorithm we use is the large margin heuristic. Thus we work with hyperplanes in a linear space. Since $\mathcal{X}$ need not be a linear space, we have to *transform* it into one, which can be done by *embedding* it into a linear space (with a norm defined on it). Since the metric information is the only information available to us to perform classification and we assume that the local structure is correlated to the class affiliations, we should not distort it too much in the embedding process. Or in other words the minimal requirement for our embedding is that it preserves neighborhoods, so it should at least be a homeomorphism of $(\mathcal{X}, d)$ onto

a subset of a linear space.
The following diagram summarizes this procedure:

$$(\mathcal{X}, d) \xrightarrow{embedding} (\mathcal{B}, \|\cdot\|) \rightarrow \text{maximal margin classification}$$

4.2.1 First step: embedding into a normed space

Maximal margin hyperplane classification requires that we work in a linear normed space. We thus have to map $\mathcal{X}$ to a subset of a normed space B (chosen to be complete, hence a Banach space).

Formally, we define a *feature map* $\Phi : \mathcal{X} \rightarrow \mathcal{B}$, $x \rightarrow \Phi(x)$, and denote by $d_{\mathcal{B}}$ the induced metric on $\mathcal{X}$.

$$d_{\mathcal{B}}(x, y) = \|\Phi(x) - \Phi(y)\|_{\mathcal{B}} \ .$$

We require that d and $d_{\mathcal{B}}$ are not too different since we want to preserve the metric information, which we assume to be relevant for classification. In other words we want that the map Φ seen as the identity map id between the metric spaces $(\mathcal{X}, d)$ and $(\mathcal{X}, d_{\mathcal{B}})$ to have one of the properties in the following list. We give the embeddings in the order of increasing requirements and each embedding is a special case of the previous one.

1. Φ is an embedding if and only if Φ is a homeomorphism, that is

$$\begin{aligned}&\forall\, x, y \in \mathcal{X}, \forall\, \epsilon > 0, \exists\, \delta_1, \delta_2 \text{ such that:}\\ &d(x, y) < \delta_1 \Rightarrow d_{\mathcal{B}}(x, y) < \epsilon, \quad d_{\mathcal{B}}(x, y) < \delta_2 \Rightarrow d(x, y) < \epsilon.\end{aligned}$$

2. Φ is a uniform embedding if and only if id : $(\mathcal{X}, d) \rightarrow (\mathcal{X}, d_{\mathcal{B}})$ is a uniform homeomorphism, that is

$$\begin{aligned}&\forall\, \epsilon > 0, \exists\, \delta_1, \delta_2 \text{ such that } \forall\, x, y \in \mathcal{X} :\\ &d(x, y) < \delta_1 \Rightarrow d_{\mathcal{B}}(x, y) < \epsilon, \quad d_{\mathcal{B}}(x, y) < \delta_2 \Rightarrow d(x, y) < \epsilon.\end{aligned}$$

3. Φ is a Bi-Lipschitz embedding, that is

$$\exists \lambda > 0,\ \forall\, x, y \in \mathcal{X},\ \frac{1}{\lambda}\, d(x, y) \leq d_{\mathcal{B}}(x, y) \leq \lambda\, d(x, y)\,.$$

4. Φ is an isometric embedding,

$$\forall\, x, y \in \mathcal{X},\ d_{\mathcal{B}}(x, y) = \|\Phi(x) - \Phi(y)\| = d(x, y)\,.$$

In this paper we will consider two cases.

- In the first case we assume that the metric $d(x, y)$ is meaningful and helpful for the classification task on all scales. That means we should preserve the metric in the embedding process, that is Φ should be an isometric embedding.
- In the second case we assume that the metric $d(x, y)$ is only locally meaningful. What do we mean by that? In the construction of a metric on a set $\mathcal{X}$ for a real-world problem one has some intuition about what it means for two elements $x, y \in \mathcal{X}$ to be 'close' and can encode this information in the metric $d(x, y)$. However larger distances are sometimes not very meaningful or even completely

arbitrary. Consider for example the edit-distance for sequences. It is fairly clear, what it means to have an edit-distance of one or two, namely the word sequence is roughly the same. However for two completely different sequences the distance will be large without any meaning and will possibly have a great influence on the construction of the classifier with the danger of fitting an irrelevant feature. Therefore in cases where we trust our metric only locally it makes no difference if we change the global structure as long as we preserve the local structure. Additionally this change of the global structure should fulfill two requirements. First it should be uniform over $\mathcal{X}$, since without further information we have no reason to change it differently in some regions. Second it should eliminate the influence of high distance values.
In mathematical terms:

Definition 4.1 *The local distortion of a map $\phi : (\mathcal{X}, d) \to (\mathcal{X}, d_{\mathcal{B}})$ is given by*

$$\mu(x) = D_+(x)/D_-(x)$$

where the functions $D_+(x)$ and $D_-(x)$ are defined as

$$D_+(x) = \limsup_{y \to x} \frac{d_{\mathcal{B}}(x,y)}{d(x,y)}, \quad D_-(x) = \liminf_{y \to x} \frac{d_{\mathcal{B}}(x,y)}{d(x,y)}.$$

Definition 4.2 *A uniform local isometry is a uniform homeomorphism $\phi : (\mathcal{X}, d) \to (\mathcal{X}, d_{\mathcal{B}})$ with local distortion $\mu(x) \equiv 1$.*

A uniform local isometry preserves the local structure up to a global rescaling, which does not matter for the maximal margin classification. Finally our embedding should be a uniform local isometry such that the transformed metric is bounded, i.e. $\sup_{x,y \in \mathcal{X}} d_{\mathcal{B}}(x,y)$ exists.

It is interesting to note here that for all embeddings $\Phi : (\mathcal{X}, d) \to (\mathcal{B}, \|\cdot\|)$ one can adopt two points of view:

- Direct embedding: $\Phi : (\mathcal{X}, d) \to (\mathcal{B}, \|\cdot\|_{\mathcal{B}})$
- Indirect embedding: identity $\mathrm{id} : (\mathcal{X}, d) \to (\mathcal{X}, d_{\mathcal{B}})$ and isometric embedding $\phi : (\mathcal{X}, d_{\mathcal{B}}) \to (\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ with $\Phi = \phi \circ \mathrm{id}$

The above two points of view are completely equivalent (i.e. any embedding Φ can be written as $\phi \circ \mathrm{id}$ where id is the identity and ϕ an isometric embedding and conversely) but the second point of view emphasizes the importance of isometric embeddings. Namely any embedding can be decomposed into a transformation of the initial metric followed by an isometric embedding. This equivalence allows us to treat isometric and uniform locally isometric embeddings in the same framework.

The first question is how to construct such a uniform local isometry. One general way to do this are the so called metric transforms introduced by Blumenthal. (We use here and in the following $\mathbb{R}_+ = \{x \in \mathbb{R} \,|\, x \geq 0\}$.)

Definition 4.3 *Let $(\mathcal{X}, d)$ be a metric space and let $F : \mathbb{R}_+ \to \mathbb{R}_+$ be a function with $F(0) = 0$. Then $(\mathcal{X}, F(d))$ is called a metric transform of $(\mathcal{X}, d)$.*

The following lemma gives sufficient conditions for a metric transform $F(d)$ to be a metric.

Lemma 4.4 *Let $F : \mathbb{R}_+ \to \mathbb{R}_+$ be a monotone increasing concave function, such that $F(0) = 0$ and $F(x) > 0$ for all $x > 0$. If d is a metric on $\mathcal{X}$, then $F(d)$ is also a metric on $\mathcal{X}$.*

The proof of this lemma can be found e.g. in [29]. We denote the functions which fulfill the assumptions of the above lemma as true metric transforms. Note that the map id : $(\mathcal{X}, d) \to (\mathcal{X}, F(d))$ is a uniform homeomorphism for every true metric transform. The next lemma characterizes all true metric transforms which are in addition uniform local isometries.

Lemma 4.5 *Let F be a true metric transform. If $\lim_{t\to 0} \frac{F(t)}{t}$ exists and is positive, then the identity* id : $(\mathcal{X}, d) \to (\mathcal{X}, F(d))$ *is a uniform local isometry. Moreover the resulting metric space $(\mathcal{X}, F(d))$ is bounded if F is bounded.*

Proof: With the assumptions the functions $D_+(x)$ and $D_-(x)$ defined in Definition 4.1 exist and $\forall x \in \mathcal{X}$, $D_+(x) = D_-(x) > 0$, so that $\mu \equiv 1$. □

In order to illustrate this lemma, we give two examples of metric transforms F which result in uniform local isometries, where $(\mathcal{X}, F(d))$ is bounded.

$$F(t) = \frac{t}{1+t}, \qquad F(t) = 1 - \exp(-\lambda t),\ \forall\, \lambda > 0 \tag{4.1}$$

Furthermore an important question is whether there exists for any given metric space $(\mathcal{X}, d)$ a Banach space $\mathcal{B}$ and a map Φ which embeds $(\mathcal{X}, d)$ isometrically into $\mathcal{B}$. In the following we will answer this positively, namely any metric space $(\mathcal{X}, d)$ can be embedded isometrically via the *Kuratowski* embedding into $(C_b(\mathcal{X}), \|\cdot\|_\infty)$, where $C_b(\mathcal{X})$ denotes the continuous, bounded functions on $\mathcal{X}$. However in the later analysis of the maximal margin algorithm it turns out that an embedding into a Hilbert space provides a simpler structure of the space of solutions. Therefore we will consider after the general case of an isometric embedding into a Banach space the special case of an isometric embedding into a Hilbert space.
Moreover all isometric embeddings we consider have the following minimal property:

Definition 4.6 (Total isometric embedding) *Given a metric space $(\mathcal{X}, d)$ and an isometric embedding $\Phi : \mathcal{X} \to \mathcal{B}$ where $\mathcal{B}$ is a Banach space, we say that Φ is a* total isometric embedding *if $\Phi(\mathcal{X})$ is total, that is $\mathcal{B}$ is the norm-closure of* $\operatorname{span}\{\Phi(x) | x \in \mathcal{X}\}$.

This definition is in a sense trivial, since if we have an isometric embedding Φ into a Banach space $\mathcal{C}$, then the norm closure of $\operatorname{span}\Phi(\mathcal{X})$ is again a Banach space $\mathcal{B}$ with the same norm. But this 'minimal' isometric embedding allows then to associate to the dual space $\mathcal{B}'$ (the space of continuous linear functionals on $\mathcal{B}$ endowed with the norm $\|w'\| = \sup_{b\in\mathcal{B},\,\|b\|\leq 1} |w'(b)|$) [1] an isometrically isomorphic Banach space of functions on $\mathcal{X}$ as we will see now.

Proposition 4.7 *Let $\Phi : \mathcal{X} \to B$ be a total isometric embedding. Then there exists a Banach space $\mathcal{F}_{\mathcal{B}'}$ of real-valued Lipschitz functions on $\mathcal{X}$ and a map $\Gamma : \mathcal{B}' \to \mathcal{F}_{\mathcal{B}'}$ such that Γ is an isometric isomorphism. The map Γ is given by*

$$\Gamma(w')(\cdot) = \langle w', \Phi(\cdot)\rangle_{\mathcal{B}',\mathcal{B}}$$

[1] Given an element b of a Banach space B and an element w' of its dual B', we write $w'(b) = \langle w', b\rangle_{B',B}$. This should not be confused with the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ in a Hilbert space.

and we define $\|\Gamma(w')\|_{\mathcal{F}_{\mathcal{B}'}} = \|w'\|_{\mathcal{B}'}$. *The Lipschitz constant of* $\Gamma(w')$ *is upper bounded by* $\|w'\|_{\mathcal{B}'}$.

We need for the proof of the proposition and in the rest of the article the following notions and a theorem relating them.

Definition 4.8 *Let* M, N *be subspaces of* $\mathcal{B}$ *resp.* $\mathcal{B}'$. *Then the annihilators* $M^\perp$ *and* ${}^\perp N$ *are defined as*

$$M^\perp = \{w' \in \mathcal{B}' : \langle w', m\rangle = 0, \forall\, m \in M\},$$
$${}^\perp N = \{b \in \mathcal{B} : \langle n, b\rangle = 0, \forall\, n \in N\}.$$

Theorem 4.9 *[75]*

- ${}^\perp(M^\perp)$ *is the norm closure of* M *in* $\mathcal{B}$.
- $({}^\perp N)^\perp$ *is the weak*-closure of* N *in* $\mathcal{B}'$

Now we can prove Proposition 4.7.
Proof: The only thing we have to prove is that Γ is injective. Let $f, g \in \mathcal{F}_{\mathcal{B}'}$, then

$$f \equiv g \;\Leftrightarrow\; \langle w_f - w_g, \Phi(x)\rangle_{\mathcal{B}',\mathcal{B}} = 0,\ \forall\, x \in \mathcal{X}.$$

Since Φ is a total isometric embedding, $\mathcal{B} = \overline{\operatorname{span}\{\Phi(\mathcal{X})\}} = {}^\perp(\operatorname{span}\{\Phi(\mathcal{X})\}^\perp)$. In particular, we have $\operatorname{span}\{\Phi(\mathcal{X})\}^\perp = \{0\}$. Therefore $w_f - w_g = 0$, that is, Γ is injective. The Lipschitz constant of $\Gamma(w')$ can be computed as follows. For all $x, y \in \mathcal{X}$,

$$\begin{aligned}|\Gamma(w')(x) - \Gamma(w')(y)| &= |\,\langle w', \Phi(x) - \Phi(y)\rangle_{\mathcal{B}',B}\,| \le \|w'\|_{\mathcal{B}'}\,\|\Phi(x) - \Phi(y)\|_{\mathcal{B}} \\ &= \|w'\|_{\mathcal{B}'}\, d(x, y).\end{aligned}$$

□

The fact that one always obtains Lipschitz functions has been pointed out in [103, 104] where it is shown that a large class of isometric embeddings can be obtained via an embedding into the predual of Lipschitz functions.

4.2.2 Second step: maximal margin classification

Maximal margin and its dual problem

What does maximal margin classification mean? The classifier is a hyperplane in $\mathcal{B}$, which can be identified with an element in the dual of $\mathcal{B}'$ plus an offset, such that the distance, the margin, to the two classes is maximized. This problem is equivalent to the problem of determining the distance between the convex hulls of the two classes of our training data. This duality was proven in the generality of an arbitrary Banach space by Zhou et al. [109]. We define the convex hull of a finite set $T \subset \mathcal{B}$ as

$$co(T) = \left\{ \sum_{i\in I} \alpha_i x_i \,|\, \sum_{i\in I} \alpha_i = 1,\ x_i \in T,\ \alpha_i \ge 0,\ |I| < \infty \right\}.$$

Theorem 4.10 *[109] Let T_1 and T_2 be two finite sets in a Banach space $\mathcal{B}$. Then if $co(T_1) \cap co(T_2) = \emptyset$*

$$\begin{aligned} d(co(T_1), co(T_2)) &= \inf_{y \in co(T_1), z \in co(T_2)} \|y - z\| \\ &= \sup_{w' \in \mathcal{B}'} \frac{\inf_{y \in T_1, z \in T_2} \langle w', y - z \rangle_{\mathcal{B}', B}}{\|w'\|}. \end{aligned} \tag{4.2}$$

The condition $co(T_1) \cap co(T_2) = \emptyset$ is equivalent to the condition of separability.

Corollary 4.11 *The maximal margin problem is translation invariant in the Banach space $\mathcal{B}$.*

Proof: This is a trival statement, since we are only interested in distances. □

Later we will use the above dual formulation in order to derive properties of the solution $w' \in \mathcal{B}'$.

Maximal margin formulations

In this section we derive from the dual problem the usual maximal margin formulation. We consider an input sample $x_1, \ldots, x_n \in \mathcal{X}$ with labels $y_1, \ldots, y_n \in \{-1, 1\}$. These samples can be embedded via Φ into a Banach space $\mathcal{B}$. We denote by Φ_x the embedded point $\Phi(x)$ and by T_1 the set $\{\Phi_{x_i} : y_i = +1\}$ of positive examples and by $T_2 = \{\Phi_{x_i} : y_i = -1\}$ the set of negative examples.

First we rewrite the second line of (4.2) by using the definition of the infimum:

$$\begin{aligned} &\sup_{x' \in \mathcal{B}',\, c,d \in \mathbb{R}} \quad \frac{c - d}{\|x'\|} \\ &\text{subject to:} \quad \langle x', y \rangle_{\mathcal{B}', B} \geq c, \quad \forall\, y \in T_1, \quad \langle x', z \rangle_{\mathcal{B}', B} \leq d, \quad \forall\, z \in T_2. \end{aligned}$$

Now subtract $-\frac{c+d}{2}$ from both inequalities, and define the following new quantities: $b = \frac{c+d}{d-c}$, $w' = \frac{2}{c-d} x'$, $T = T_1 \cup T_2$. Then taking the inverse we arrive at the standard hard margin formulation:

$$\begin{aligned} &\min_{w' \in \mathcal{B}', b} \|w'\| \\ &\text{subject to:} \quad y_i(\langle w', \Phi_{x_i} \rangle_{\mathcal{B}', B} + b) \geq 1, \quad \forall i = 1, \ldots, n. \end{aligned} \tag{4.3}$$

Another equivalent formulation where we use the space of functions $\mathcal{F}_{\mathcal{B}'}$ which we defined in Proposition 4.7 takes more the point of view of regularization.

$$\min_{f_{w'} \in \mathcal{F}_{\mathcal{B}'}, b} \|f_{w'}\|_{\mathcal{F}_{\mathcal{B}'}} + \sum_{i=1}^{n} \ell\left(y_i(f_{w'}(x_i) + b)\right) \tag{4.4}$$

where the loss function ℓ is given by $\ell(x) = 0$, $\forall\, x \geq 1$, $\ell(x) = \infty$, $\forall\, x < 1$.

In principle we have two points of view on the hard margin problem. One is based on the geometric interpretation (4.2), (4.3) of finding a separating hyperplane with maximal distance to the two classes. The other is based on (4.4) and regards the problem as the search for a function which classifies correctly and has minimal norm, where we assume that the norm is some measure of smoothness. In this paper we will switch between these two viewpoints depending on which is better suited to illustrate a certain property.

Form of the solution

Let us now come back to the initial formulation (4.2). Our goal is to obtain a characterization of the solutions $w' \in \mathcal{B}'$. We consider the following subspace $A = \operatorname{span}\{\Phi_{x_1} - \Phi_{x_2} \,|\, x_1 \in T_1, x_2 \in T_2\} \subset \operatorname{span}\{\Phi_{x_i} \,|\, x_i \in T\}$ which can be equivalently written as

$$A = \left\{ \sum_{i=1}^{n} \alpha_i \Phi_{x_i} : \sum_{i=1}^{n} \alpha_i = 0 \right\}.$$

The following lemma characterizes the space of solution $w' \in \mathcal{B}'$.

Lemma 4.12 *The quotient space*[2] $\mathcal{B}'/A^\perp$*, endowed with the quotient norm, is a Banach space. It is isometrically isomorphic to the dual* A' *of* A *and has dimension* $n-1$*. Moreover the problem of maximal margin separation in* $\mathcal{B}'$*,* (4.3)*, is equivalent to the following problem in* $\mathcal{B}'/A^\perp$*:*

$$\min_{w' \in \mathcal{B}'/A^\perp, b} \|w'\|_{\mathcal{B}'/A^\perp} \tag{4.5}$$
$$\text{subject to:} \quad y_i \Big(\langle w', \Phi_{x_i} \rangle_{\mathcal{B}',\mathcal{B}} + b \Big) \geq 1, \quad \forall i = 1, \ldots, n.$$

Proof: A is finite dimensional hence closed in $\mathcal{B}$. It is thus a Banach space with the induced norm. It is well known (see e.g. [75]) that then $\mathcal{B}'/A^\perp$ with the quotient norm $\|b'\|_{\mathcal{B}'/A^\perp} = \inf\{\|b' - a'\| : a' \in A^\perp\}$ is a Banach space isometric isomorphic to A', the dual of A. Since A is a normed space of finite dimension $n-1$, its dual has the same dimension.
Since addition of elements of $A^\perp$ does not change the numerator of (4.2), but will change the norm in the denominator, the problem can be equivalently formulated in the quotient space $\mathcal{B}'/A^\perp$.
In the constraint of (4.5), w' is an arbitrary representative of its equivalence class $w' + A^\perp$. This is well defined, since if u' is another representative of the equivalence class we have $m' = u' - w' \in A^\perp$ and $\forall m' \in A^\perp, \phi_{x_i}, \phi_{x_j} \in T$,

$$\left\langle m', \phi_{x_i} - \phi_{x_j} \right\rangle_{\mathcal{B}',\mathcal{B}} = 0.$$

That is, m' is constant on the data. Therefore if w' satisfies the constraint with constant b, u' will satisfy the constraint with the constant $c = b - \langle m, \phi_{x_i} \rangle_{\mathcal{B}',\mathcal{B}}$. □

Remarkably, this lemma tells us that the solution of the maximum margin problem is effectively in a finite dimensional subspace of $\mathcal{B}'$ which is determined by the data. However, it gives no explicit description how this subspace depends on the data, which makes it hard to be effectively used in general.

Moreover, in order to solve the initial problem using the above lemma, one has to first solve the finite dimensional problem in $\mathcal{B}'/A^\perp$ and then to solve the minimum norm interpolation problem in $\mathcal{B}'$. Indeed, if a is a solution in $\mathcal{B}'/A^\perp$, one has to find an element b' in the equivalence class a. For this one has to solve

$$\inf_{b' \in \mathcal{B}' : b'|_A = a|_A} \|b'\|_{\mathcal{B}'},$$

which corresponds to minimizing the norm provided the values on a finite dimensional subspace are known.

[2] A closed subspace C of a Banach space $\mathcal{B}$ defines a linear equivalence relation $\sim$ by $u \sim v$ if $u - v \in C$. The quotient space $\mathcal{B}/C$ is the vector space of these equivalence classes.

We give an interpretation of this lemma from the point of view of functions which we developed in the previous section. The closed subspace $A^\perp$ of $\mathcal{B}'$ defines a closed subspace of functions $\mathcal{F}_{A^\perp}$ of $\mathcal{F}_{\mathcal{B}'}$ on $(\mathcal{X}, d)$ which are constant on all data points, namely $\forall\, w' \in A^\perp$, $x_1 \in T_1, x_2 \in T_2$

$$f_{w'}(x_1) - f_{w'}(x_2) = \langle w', \Phi(x_1) - \Phi(x_2)\rangle_{\mathcal{B}',B} = 0$$

The proposition then states that the solution is only defined up to a constant function on the data or in other words we are looking for a solution f in $\mathcal{F}_{\mathcal{B}'}/\mathcal{F}_{A^\perp}$ with the usual quotient norm $\|f\|_{\mathcal{F}_{\mathcal{B}'}/\mathcal{F}_{A^\perp}} = \inf_{g\in\mathcal{F}_{A^\perp}} \|f - g\|$. In particular, if there are constant functions (constant functions are constant on the data) in our function class $\mathcal{F}_{\mathcal{B}'}$, they will not be penalized in the norm. This reflects the fact that constant functions are useless for classification and should therefore not be considered in the norm of our solution space. Since we use the threshold 0 for classification we have to compensate for the constant functions on the data with the bias term b in the final solution.

$$f_{w'}(x) = \operatorname{sgn}(\langle w', \Phi(x)\rangle_{\mathcal{B}',B} + b).$$

Later we will consider also isometric embeddings into a Hilbert space $\mathcal{H}$. There we have $(A^\perp)^\perp = A$ and we can actually decompose $\mathcal{H}$ into $\mathcal{H} = A^\perp \oplus A$. Then the solution of the maximal margin problem is an element of A, which is itself a Hilbert space and consists of all functions $f \in \mathcal{H}$, orthogonal to the functions which are constant on the data. This is a stronger statement than the usual representer theorem, which says that the solution lies in the space spanned by the data.

4.3 Metric based maximal margin classifier in a Banach space

In this section we treat the general case, where we embed isometrically a given metric space $(\mathcal{X}, d)$ into a Banach space $\mathcal{B}$ followed by a maximal margin classification in $\mathcal{B}$. In general there exist for each metric space several Banach spaces, into which it can be embedded isometrically. In this section we use the very simple Kuratowski embedding. After the definition of the Kuratowski embedding Φ and the corresponding Banach space $\mathcal{B}$ we finally formulate the algorithm of maximal margin classification in $\mathcal{B}$. Unfortunately the full problem cannot be solved exactly. We provide a reasonable approximation to the full problem, which is exact if one considers the training set and a possible test point as a finite metric space.
The following diagram illustrates the employed procedure

$$(\mathcal{X}, d) \xrightarrow{isometric} (D, \|.\|_\infty) \subset (C_b(\mathcal{X}), \|.\|_\infty) \to \text{maximal margin separation}$$

where D is a Banach space of (continuous and bounded) functions defined on $\mathcal{X}$ (see definitions below).

4.3.1 Isometric embedding into a Banach space

Let $(\mathcal{X}, d)$ be a metric space and denote by $C_b(\mathcal{X})$ the Banach space of continuous and bounded functions on $\mathcal{X}$ endowed with the supremum norm. If $\mathcal{X}$ is compact

the topological dual of $C_b(\mathcal{X})$ is the space of regular signed Borel measures $\mathcal{M}(\mathcal{X})$ with the measure norm $\|\mu\| = \int_{\mathcal{X}} d\mu_+ - \int_{\mathcal{X}} d\mu_-$ (where μ_+ and μ_- are respectively the positive and negative parts of μ).
Consider an arbitrary $x_0 \in \mathcal{X}$ and define the following map

$$\Phi : \mathcal{X} \;\to\; \mathbb{R}^{\mathcal{X}}, \quad x \;\mapsto\; \Phi_x := d(x,\cdot) - d(x_0,\cdot)$$

Let $D = \overline{\operatorname{span}\{\Phi_x : x \in \mathcal{X}\}}$, where the closure is taken in $(C_b(\mathcal{X}), \|\cdot\|_\infty)$.
We will show that Φ defines an isometric embedding of the metric space $\mathcal{X}$ into D.

Lemma 4.13 *Φ is a total isometric embedding from $(\mathcal{X}, d)$ into the Banach space $(D, \|\cdot\|_\infty) \subset (C_b(\mathcal{X}), \|\cdot\|_\infty)$.*

Proof: We have $\|\Phi_x\|_\infty \le d(x, x_0) < \infty$ and $|\Phi_x(y) - \Phi_x(y')| \le |d(x,y) - d(x,y')| + |d(x_0,y) - d(x_0,y')| \le 2d(y,y')$, so that $\Phi_x \in C_b(\mathcal{X})$. In addition $\|\Phi_x - \Phi_y\|_\infty = \|d(x,\cdot) - d(y,\cdot)\|_\infty = d(x,y)$ and the supremum is attained at x and y. Hence, Φ is an isometry from $(\mathcal{X}, d)$ into $(D, \|\cdot\|_\infty)$ which is a closed subspace of $C_b(\mathcal{X})$. Therefore $(D, \|\cdot\|)$ is also Banach space and Φ a total isometric embedding, since by definition $D = \overline{\operatorname{span}\,\Phi(\mathcal{X})}$. □

Note that, as an isometry, Φ is continuous, and x_0 is mapped to the origin of D. The choice of this origin Φ_{x_0} has no influence on the classifier since the maximal margin problem is translation invariant.

4.3.2 The algorithm

The maximal margin formulation (4.3) can be directly stated as:

$$\begin{aligned} &\min_{w' \in D', b \in \mathbb{R}} \|w'\| \\ &\text{subject to:} \quad y_j \left(\langle w', \Phi_{x_j} \rangle_{D',D} + b \right) \ge 1, \quad \forall\, j = 1, \dots, n. \end{aligned} \tag{4.6}$$

Note that since we have no explicit description of the dual space D' we cannot solve this directly. If $\mathcal{X}$ is compact it is well-known that the dual of $C_b(\mathcal{X})$ is isometrically isomorphic to the Banach space of regular signed Borel measures $\mathcal{M}(\mathcal{X})$ on $\mathcal{X}$ with the measure norm. Thus we can state the problem explicitly. Note that even though we work in a bigger space than D', we will get the same solution lying in A' isometrically isomorphic to $\mathcal{M}(X)/A^{\perp}$[3]since we are minimizing the norm:

$$\begin{aligned} &\min_{w' \in \mathcal{M}(\mathcal{X}), b \in \mathbb{R}} \|\mu\|_{\mathcal{M}(\mathcal{X})} \\ &\text{subject to:} \quad y_j \left(\int_{\mathcal{X}} (d(x_j, x) - d(x, x_0)) d\mu(x) + b \right) \ge 1, \quad \forall\, j = 1, \dots, n. \end{aligned}$$

This problem also cannot be solved directly, since we have no parametrization of $\mathcal{M}(\mathcal{X})$.
Let us now consider again the general problem (4.6). Since we have neither a description of the dual $A' \simeq D'/A^{\perp}$ nor of D', we develop a reasonable approximation in the bigger space $C_b(X)'$. We introduce the space E defined as the span of evaluation functionals:

$$E := \operatorname{span}\{\delta_x : x \in \mathcal{X}\}\,.$$

First we have the following lemma:

[3]Let $A^{\perp}_{\mathcal{M}(X)}, A^{\perp}_{D'}$ denote the annihilator of A in $\mathcal{M}(X)$ resp. D'. Then we have $A' \simeq \mathcal{M}(X)/A^{\perp}_{\mathcal{M}(X)} \simeq D'/A^{\perp}_{D'}$.

Lemma 4.14 *The space E defined above is weak*-dense in the dual of $C_b(\mathcal{X})$ and the norm is given by $\|\sum_{i=1}^n \alpha_i \delta_{x_i}\|_{C_b(\mathcal{X})'} = \sum_{i=1}^n |\alpha_i|$.*

Proof: The evaluation functionals are in the dual of $C_b(\mathcal{X})$ since

$$|\delta_x(f) - \delta_x(g)| = |f(x) - g(x)| \leq \|f - g\|_\infty$$

Consider now the span of evaluation functionals $\text{span}\{\delta_x : x \in \mathcal{X}\}$. The norm induced by $C_b(\mathcal{X})$ is given as

$$\left\| \sum_{i=1}^n \alpha_i \delta_{x_i} \right\|_{C_b(\mathcal{X})'} = \sup_{f \in C_b(\mathcal{X})'} \frac{|\langle \sum_{i=1}^n \alpha_i \delta_{x_i}, f \rangle|}{\|f\|_\infty} = \sup_{f \in C_b(\mathcal{X})'} \frac{|\sum_{i=1}^n \alpha_i f(x_i)|}{\|f\|_\infty} = \sum_{i=1}^n |\alpha_i|$$

Further on we have that ${}^\perp\{\delta_x : x \in \mathcal{X}\} = 0$ since $\langle \delta_x, f \rangle = 0, \forall\, x \in \mathcal{X} \Leftrightarrow f \equiv 0$. That implies

$$({}^\perp\{\delta_x : x \in \mathcal{X}\})^\perp = C_b(\mathcal{X})'$$

Therefore by Theorem 4.9 the weak*-closure of $\text{span}\{\delta_x : x \in \mathcal{X}\}$ is $C_b(\mathcal{X})'$. □

Let us explain shortly what this result means. The weak*-topology is the topology of pointwise convergence on $C_b(\mathcal{X})$. Therefore the weak*-denseness of E in $C_b(\mathcal{X})'$ can be equivalently formulated as follows: $\forall\, \mu \in C_b(\mathcal{X})'$, $\exists\, \{e_\alpha\}_{\alpha \in I} \in E$ such that $e_\alpha \to \mu$ in the weak*-topology, that is

$$\forall\, f \in C_b(\mathcal{X}), \quad \langle e_\alpha, f \rangle_{C_b(\mathcal{X})', C_b(\mathcal{X})} \longrightarrow \langle \mu, f \rangle_{C_b(\mathcal{X})', C_b(\mathcal{X})}\,.$$

In other words one can approximate in the above sense any element of $C_b(\mathcal{X})'$ arbitrarily well with elements from E. On the other hand weak*-dense does not imply norm-dense.

Our first step is that we formulate the problem in $C_b(\mathcal{X})'$ which seems at first to be an approximation. But according to the same argument as before we have $A' \simeq C_b(X)'/A^\perp \simeq D'/A^\perp$. Since we are minimizing the norm under the given constraints this implies that the solution will lie in $C_b(\mathcal{X})'/A^\perp$ which is isometrically isomorphic to A'. Then as an first approximation we restrict $C_b(\mathcal{X})'$ to E. Since the span of evaluation functionals is not norm dense in $C_b(\mathcal{X})'$, this implies that even in the limit of an infinite number of evaluation functionals we might not get the optimal solution.
This approximation can be formulated as the following optimization problem:

$$\begin{aligned} \inf_{e \in E, b} \|e\| = \inf_{m \in \mathbb{N},\, z_1, \ldots, z_m \in \mathcal{X}^m,\, b} \quad & \sum_{i=1}^m |\beta_i| \\ \text{s.t. } y_j \left(\sum_{i=1}^m \beta_i \left\langle \delta_{z_i}, \Phi_{x_j} \right\rangle + b \right) &= y_j \left(\sum_{i=1}^m \beta_i (d(x_j, z_i) - d(x_0, z_i)) + b \right) \geq 1 \\ \forall j = 1, \ldots, n. & \end{aligned}$$

Unfortunately it is not possible to prove that the solution can be expressed in terms of the data points only (which would be a form of a representer theorem for this algorithm). We could actually construct explicit counterexamples. Note however that in [103] a representer theorem was derived for a similar but different setting. Namely they showed that if one considers all Lipschitz functions together with the

Lipschitz constant as a norm, the solution lies in the *vector lattice* spanned by the data. However it is also shown there that this setting is not equivalent to the setting presented here. Moreover as we will show later the capacity of all Lipschitz functions measured by Rademacher averages is higher than of our approach.
In order to make the problem computationally tractable, we have to restrict the problem to a finite dimensional subspace of E. A simple way to do this is to consider only the subspace of E generated by a finite subset $Z \in \mathcal{X}, |Z| = m$, which includes the training set $T \subset Z$. We are free to choose the point x_0 in the embedding, so we choose it as $x_0 = z_1, z_1 \in Z$. Since the problem stated in Theorem 4.10 is translation invariant, this choice has no influence on the solution. This leads to the following optimization problem:

$$\begin{aligned} &\min_{\beta_i, b} \sum_{i=1}^{m} |\beta_i| \\ &\text{subject to:} \quad y_j \left(\sum_{i=1}^{m} \beta_i (d(x_j, z_i) - d(z_1, z_i)) + b \right) \geq 1, \quad \forall x_j \in T. \end{aligned}$$

A convenient choice for Z is $Z = T$. In a transduction setting one can use for Z the union of labelled and unlabelled data.
As the second term in the constraint, $\sum_{i=1}^{m} \beta_i d(z_1, z_i)$, does not depend on j, we can integrate it in a new constant c and solve the equivalent problem:

$$\begin{aligned} &\min_{\beta_i, c} \sum_{i=1}^{m} |\beta_i| \\ &\text{subject to:} \quad y_j \left(\sum_{i=1}^{m} \beta_i \, d(x_j, z_i) + c \right) \geq 1, \quad \forall x_j \in T. \end{aligned} \tag{4.7}$$

The corresponding decision function is given by

$$f(x) = \operatorname{sgn} \left(\sum_{i=1}^{m} \beta_i \, d(x, z_i) + c \right) .$$

The above optimization problem can be transformed into a linear programming problem, and is easily solvable with standard methods. Note that if we take $Z = T$ we recover the algorithm proposed by Graepel et al. [37]. We also note that it is easily possible to obtain a soft-margin version of this algorithm. In this case there still exists the equivalent problem of finding the distance between the reduced convex hulls [12, 109]. This algorithm was compared to other distance based classifiers by Pekalska et al. in [71] and showed good performance.

The approximation with a finite subset Z, $|Z| = m$, such that $T \subset Z$ can also be seen from another point of view. Namely consider the finite metric space (Z, d). Since the isometric embedding Φ is possible for any metric space, we can use it also in this special case and the Banach space of continuous, bounded functions $(C_b(Z), \|\cdot\|_\infty)$ is actually equal to $l_\infty^m = (\mathbb{R}^m, \|\cdot\|_\infty)$. We note that in the case of finite dimension m the dual of l_∞^m is given by l_1^m. Formulating the maximal margin problem in the Banach space l_∞^m leads then exactly to the optimization problem (4.7). Therefore the approximation to the maximal margin problem for $(\mathcal{X}, d)$ using a finite subset of

evaluation functionals indexed by Z is equivalent to the maximal margin problem for the finite metric space (Z, d) without any approximation. Moreover one can embed $m+1$ points isometrically into l_∞^m with the embedding Φ (z_1 is mapped to the origin of l_∞^m). Thus the resulting classifier is not only defined on Z but by embedding Z plus a possible test point $x \in \mathcal{X}$ isometrically into l_∞^m we can classify all points $x \in \mathcal{X}$ respecting all the distance relationships of x to Z.

4.4 Metric based maximal margin classifier in a Hilbert space

In the previous section we constructed a maximal margin classifier in the Banach space $D \subset (C_b(\mathcal{X}), \|\cdot\|_\infty)$ which works for any metric space $(\mathcal{X}, d)$, since any metric space can be embedded isometrically into $(C_b(\mathcal{X}), \|\cdot\|_\infty)$. The problem of the resulting maximal margin classifier is that the space of solutions $D'/A^\perp$ is not easily accessible. However in a Hilbert space the dual space $\mathcal{H}'$ is isometrically isomorphic to $\mathcal{H}$. Therefore we have $\mathcal{H}/A^\perp = (A^\perp)^\perp = A$, that is given n data points we have an explicit description of the at most $(n-1)$-dimensional space of solutions.
Regarding these properties of the space of solutions in $\mathcal{H}$ it seems desirable to rather embed isometrically into a Hilbert space than into a Banach space. It turns out that isometric embeddings into Hilbert spaces are only possible for a subclass of metric spaces. Following the general framework we first treat isometric and uniform locally isometric embeddings. Then the resulting maximal margin classifier is determined. Finally we show the equivalence to the SVM and provide an alternative point of view on kernels regarding SVMs.

4.4.1 Isometric embedding into a Hilbert space

We have seen in the previous part that all metric spaces can be embedded isometrically into a Banach space. Is this true also for isometric embeddings into Hilbert spaces? The answer was given by Schoenberg in 1938 in terms of the following class of functions, by now well-known as positive definite resp. conditionally positive definite kernels.

Definition 4.15 *A real valued function k on $\mathcal{X} \times \mathcal{X}$ is positive definite (resp. conditionally positive definite) if and only if k is symmetric and*

$$\sum_{i,j}^{n} c_i c_j k(x_i, x_j) \geq 0, \tag{4.8}$$

for all $n \in \mathbb{N}$, $x_i \in \mathcal{X}, i = 1, ..., n$, and for all $c_i \in \mathbb{R}, i = 1, ..., n$, (resp. for all $c_i \in \mathbb{R}, i = 1, ..., n$, with $\sum_i^n c_i = 0$).

The metric spaces which can be isometrically embedded into a Hilbert space can be characterized as follows:

Theorem 4.16 (Schoenberg [80]) *A metric space $(\mathcal{X}, d)$ can be embedded isometrically into a Hilbert space if and only if $-d^2(x, y)$ is conditionally positive definite.*

Based on this characterization, one can introduce the following definition.

Definition 4.17 *A metric d defined on a space $\mathcal{X}$ is called a* Hilbertian metric *if $(\mathcal{X}, d)$ can be isometrically embedded into a Hilbert space, or equivalently if $-d^2$ is conditionally positive definite.*

We notice that isometric embeddings into a Hilbert space are only possible for a restricted subclass of metric spaces. So we achieve the advantage of having a small and easily accessible space of solutions by losing the ability to handle the whole class of metric spaces in this framework.
Let us now construct explicitly the corresponding isometric embedding.

Proposition 4.18 *Let $d(x, y)$ be a Hilbertian metric. Then for every point $x_0 \in \mathcal{X}$ there exists a reproducing kernel Hilbert space $\mathcal{H}_k$ and a map $\Psi : \mathcal{X} \to \mathcal{H}_k$ given by*

$$x \to \Psi_x(\cdot) = \frac{1}{2}\left(-d^2(x, \cdot) + d^2(x, x_0) + d^2(\cdot, x_0)\right)$$

such that

- $\{\Psi_x | x \in X\}$ *is total in* $\mathcal{H}_k$
- $\|\Psi_x - \Psi_y\|_{\mathcal{H}_k} = d(x, y)$.
- $\Psi_{x_0} = 0$

We need the following two lemmata to prove this proposition.

Lemma 4.19 *[14] Let $\mathcal{X}$ be a nonempty set, $x_0 \in \mathcal{X}$, and let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric function. Let $\tilde{k}(x, y)$ be given by*

$$\tilde{k}(x, y) = k(x, y) - k(x, x_0) - k(x_0, y) + k(x_0, x_0).$$

Then $\tilde{k}$ is positive definite if and only if k is conditionally positive definite.

Lemma 4.20 *[81] Given a positive definite kernel $k(x, y) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ there exists a unique reproducing kernel Hilbert space (RKHS) of functions on $\mathcal{X}$, where $\mathcal{H} = \overline{\text{span}\{k(x, \cdot) | x \in \mathcal{X}\}}$.*

Proof: [Proposition 4.18] Define the symmetric kernel function $k(x, y) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$k(x, y) = \frac{1}{2}\left(-d^2(x, y) + d^2(x, x_0) + d^2(y, x_0)\right).$$

Using Lemma 4.19, $k(x, y)$ is a positive definite kernel. Moreover by Lemma 4.20 there exists a unique reproducing kernel Hilbert space $\mathcal{H}_k$ associated to $k(x, y)$ such that $k(x, y) = \langle \Psi_x, \Psi_y \rangle_{\mathcal{H}_k}$ and $\{k(x, \cdot) | x \in \mathcal{X}\} = \{\Psi_x | x \in \mathcal{X}\}$ is total in $\mathcal{H}_k$. Moreover we have

$$\|\Psi_x - \Psi_y\|^2 = k(x, x) + k(y, y) - 2k(x, y) = d^2(x, y)$$

and $\Psi_{x_0}(\cdot) = \frac{1}{2}(-d^2(x_0, \cdot) + d^2(x_0, \cdot)) = 0$. □

4.4.2 Uniform locally isometric embedding into a Hilbert space

In the previous section we constructed an isometric embedding into a Hilbert space. If one trusts the metric $d(x, y)$ only locally we argued in section 4.2.1 that one should use a uniform locally isometric embedding.
The following proposition gives necessary and sufficient conditions for a uniform embedding of a metric space into a Hilbert space:

Proposition 4.21 *[1] A metric space $(\mathcal{X}, d)$ can be uniformly embedded into a Hilbert space if and only if there exists a positive definite kernel $k(x, y)$ on $\mathcal{X}$ such that*

- *For every $x \in \mathcal{X}$, $k(x, x) = 1$*
- *k is uniformly continuous*
- *For every $\varepsilon > 0$, $\inf\{1 - k(x, y) : d(x, y) \geq \varepsilon\} > 0$*
- *$\lim_{\varepsilon \to 0} \sup\{1 - k(x, y) : d(x, y) \leq \varepsilon\} = 0$*

The following corollary extends the previous proposition to uniform local isometries.

Corollary 4.22 *Let $(\mathcal{X}, d)$ be a metric space and k a positive definite kernel which fulfills the conditions of Proposition 4.21. If the limits*

$$\limsup_{y \to x} \frac{1 - k(x, y)}{d^2(x, y)}, \qquad \liminf_{y \to x} \frac{1 - k(x, y)}{d^2(x, y)}$$

exist and are non-zero then $\phi_x : x \to k(x, \cdot)$ is a uniform local isometry of $(\mathcal{X}, d)$ onto a subset of the RKHS associated to k.

Proof: Simply calculate the metric induced by the positive definite kernel k, $d_k^2(x, y) = 2 - 2k(x, y)$ and use the definition of the functions D_+ and D_- in Definition 4.1. The explicit embedding ϕ follows from Lemma 4.20. □

In principle the above proposition and the corollary are not very satisfying since they provide no explicit construction of a positive definite kernel which fulfills the conditions for a given metric.
In the case where the given metric is a Hilbertian metric we can use a result of Schoenberg. It characterizes the metric transforms F of a given Hilbertian metric d, such that $F(d)$ is also a Hilbertian metric. This implies that the the identity map $\mathrm{id} : (\mathcal{X}, d) \to (\mathcal{X}, F(d))$ is a uniform homeomorphism. Moreover using Lemma 4.20 we get a uniform embedding into a Hilbert space.

Theorem 4.23 *[79] Let $F : \mathbb{R}_+ \to \mathbb{R}_+$ be a function such that $F(0) = 0$ and all derivatives of F exist on $\mathbb{R}_+ \backslash \{0\}$. Then the following assertions are equivalent:*

- *$F(d)$ is a Hilbertian metric, if d is a Hilbertian metric.*
- $$F(t) = \left(\int_0^\infty \frac{1 - e^{-t^2 u}}{u} \, d\gamma(u) \right)^{1/2},$$

  where $\gamma(u)$ is monotone increasing for $u \geq 0$ and satisfying $\int_1^\infty \frac{d\gamma(u)}{u} < \infty$[4].

[4] The integrals are Lebesgue-Stieltjes integrals.

- $(-1)^{n-1}\frac{d^n}{dt^n}F^2(\sqrt{t}) \geq 0$ *for all* $t > 0$ *and* $n \geq 1$.

Moreover F *is bounded if and only if*

$$\lim_{\epsilon \to 0} \gamma(\epsilon) = \gamma(0), \text{ and } \lim_{\epsilon \to 0} \int_{\epsilon}^{1} \frac{d\gamma(u)}{u} \text{ exists.}$$

For a uniform, local isometric embedding one has to fulfill in addition the requirements of Lemma 4.5. Combining Theorem 4.23 and Lemma 4.5 we get a complete description of all metric transforms for a given Hilbertian metric which induce a uniform local isometry and where the transformed metric is Hilbertian. The examples given in (4.1) fulfill both the conditions of Theorem 4.23 and of Lemma 4.5. Therefore they provide two examples of metric transforms which induce uniform local isometries and produce Hilbertian metrics if one starts with a Hilbertian metric. The drawback of the Theorem 4.23 is that we have to start with a Hilbertian metric. A more general theorem which characterizes metric transforms of an arbitrary metric space such that the transformed metric is Hilbertian seems not to be available in the literature.

4.4.3 The maximal margin algorithm

In the general considerations we defined the subspace $A \subset \text{span}\{\Psi_x | x \in T\}$ by

$$A := \left\{ \sum_{i=1}^{n} \alpha_i \Psi_{x_i} : \sum_{i=1}^{n} \alpha_i = 0 \right\}.$$

Since in a Hilbert space the dual is isometrically isomorphic to the Hilbert space itself we get the following form of the space of solutions:

Lemma 4.24 *The space of solutions* $\mathcal{H}/A^{\perp}$ *is equal to* A.

Proof: We have the simple equalities $\mathcal{H}/A^{\perp} = (A^{\perp})^{\perp} = A$. □

Following Zhou [109] note that if in (4.2) the infimum on the left is achieved by $y_0 \in co(T_1)$ and $z_0 \in co(T_2)$ then w' is aligned with $y_0 - z_0$, that is

$$\langle y_0 - z_0, w' \rangle_{\mathcal{H}} = \|y_0 - z_0\|_{\mathcal{H}} \|w'\|_{\mathcal{H}}$$

In a Hilbert space it follows from the Cauchy-Schwarz inequality that in this case $w' = y_0 - z_0$. Therefore in a Hilbert space the problem of maximal margin separation is not only equivalent to the problem of finding the distance of the convex hulls but it has also the same solution. Therefore we can equivalently formulate the problem of maximal margin separation as finding the distance of the convex hulls of the isometrically embedded training data in $\mathcal{H}_k$.
The optimization problem corresponding to the maximum margin hyperplane can be written as

$$\min_{\alpha} \left\| \sum_{i:y_i=+1} \alpha_i \Psi_{x_i} - \sum_{i:y_i=-1} \alpha_i \Psi_{x_i} \right\|_{\mathcal{H}_k}^2$$
$$\text{subject to: } \sum_{i:y_i=+1} \alpha_i = \sum_{i:y_i=-1} \alpha_i = 1, \quad \alpha_i \geq 0,$$

The distance $\left\|\sum_{i:y_i=+1}\alpha_i\Psi_{x_i}-\sum_{i:y_i=-1}\alpha_i\Psi_{x_i}\right\|_{\mathcal{H}_k}$ can be calculated explicitly with the expression of the inner product $\langle\Psi_x,\Psi_y\rangle_{\mathcal{H}_k}=k(x,y)$ from the proof of Proposition 4.18:

$$\begin{aligned}\left\|\sum_i y_i\alpha_i\Phi_{x_i}\right\|^2_{\mathcal{H}_k} &= \sum_{i,j=1}^n y_iy_j\alpha_i\alpha_jk(x_i,x_j)\\ &= \frac{1}{2}\sum_{i,j=1}^n y_iy_j\alpha_i\alpha_j(-d^2(x_i,x_j)+d^2(x_i,x_0)+d^2(x_0,x_j))\\ &= -\frac{1}{2}\sum_{i,j=1}^n y_iy_j\alpha_i\alpha_jd^2(x_i,x_j)\end{aligned}$$

where the other terms vanish because of the constraint $\sum_{i=1}^n y_i\alpha_i=0$. So the final optimization problem becomes

$$\begin{aligned}&\min_\alpha -\frac{1}{2}\sum_{i,j=1}^n y_iy_j\alpha_i\alpha_jd^2(x_i,x_j)\\ &\text{subject to: } \sum_i y_i\alpha_i=0,\quad \sum_i\alpha_i=2,\quad \alpha_i\geq 0,\end{aligned}$$

and with $w=\sum_{i=1}y_i\alpha_i\Phi_{x_i}$ the final classifier has the form

$$\begin{aligned}f(x)&=\langle w,\Phi_x\rangle_{\mathcal{H}_k}+b=\sum_{i=1}^n y_i\alpha_ik(x_i,x)+b\\ &=-\frac{1}{2}\sum_{i=1}^n\alpha_iy_i(d^2(x_i,x)-d^2(x_i,x_0))+b=-\frac{1}{2}\sum_{i=1}^n\alpha_iy_id^2(x_i,x)+c\end{aligned}$$

The constant c is determined in such a way that the hyperplane lies exactly half way between the two closest points of the convex hulls. Following this consideration the point $m=\frac{1}{2}\sum_{i=1}^n\alpha_i\Phi_{x_i}$ lies on the hyperplane. Then c can be calculated by:

$$c=-\langle w,m\rangle_{\mathcal{H}_k}=\frac{1}{2}\sum_{i,j=1}^n y_i\alpha_i\alpha_j(d^2(x_i,x_j)-d^2(x_i,x_0))$$

4.4.4 Equivalence to the Support Vector Machine

The standard point of view on SVM is that we have an input space $\mathcal{X}$ which describes the data. This input space $\mathcal{X}$ is then embedded via Φ into a Hilbert space $\mathcal{H}$ with a positive definite kernel[5] and then maximal margin separation is done. The following diagram summarizes this procedure:

$$\mathcal{X}\xrightarrow{\text{kernel }k}\mathcal{H}_k\longrightarrow\text{maximal margin separation}\tag{4.9}$$

[5]Originally the SVM was only formulated with positive definite kernels. Later it was shown in [82] that due to the translation invariance of the maximal margin problem in feature space one can use the class of conditionally positive definite kernels. In this case the kernel $k(x,y)$ is not equal to an inner product $\langle\Phi_x,\Phi_y\rangle$ in a Hilbert space, but it defines an inner product on a subspace which includes A.

where the kernel k is positive definite.
We show now that this is equivalent to the point of view in this paper:

$$(\mathcal{X}, d) \xrightarrow{isometric} \mathcal{H}_k \longrightarrow \text{maximal margin separation}$$

where d is a Hilbertian metric.
The next proposition is the key to this equivalence. It is a characterization of the class of all conditionally positive definite kernels in terms of the class of Hilbertian metrics. It can be found in Berg et al. (see Proposition 3.2 of [14]). We have rewritten it in order to stress the relevant result.

Proposition 4.25 *All conditionally positive definite kernels $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are generated by a Hilbertian metric $d(x, y)$ in the sense that there exists a function $g : \mathcal{X} \to \mathbb{R}$ such that*

$$k(x, y) = -\frac{1}{2} d^2(x, y) + g(x) + g(y)\,, \tag{4.10}$$

and any kernel of this form induces the Hilbertian metric d via

$$d^2(x, y) = k(x, x) + k(y, y) - 2k(x, y). \tag{4.11}$$

This proposition establishes a many-to-one correspondence between the set of conditionally positive definite kernels and Hilbertian metrics. This is rather obvious since already any change of the origin in the RKHS corresponds to a new kernel function on $\mathcal{X}$ but the induced metric (4.11) is invariant. Moreover the following theorem shows that only the Hilbertian metric d matters for classification with the SVM.

Theorem 4.26 *The SVM is equivalent to the metric based maximal margin classifier in a Hilbert space. The solution of the SVM does not depend on the specific isometric embedding Φ, nor on the corresponding choice of the kernel in a given family determined by a Hilbertian metric, see* (4.10)*. The optimization problem and the solution can be completely expressed in terms of the (semi)-metric d of the input space,*

$$\min_{\alpha} \left\| \sum_i y_i \alpha_i \Phi_{x_i} \right\|^2_{\mathcal{H}_k} = -\frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j d^2(x_i, x_j)$$
$$\textit{subject to} : \sum_i y_i \alpha_i = 0, \quad \sum_i \alpha_i = 2, \quad \alpha_i \geq 0.$$

The solution can be written as

$$f(x) = -\frac{1}{2} \sum_i y_i \alpha_i d^2(x_i, x) + c.$$

Proof: By Proposition 4.25 all conditionally positive definite kernels are generated by a Hilbertian metric $d(x, y)$. Using (4.10) one can show now that for each kernel associated to a Hilbertian metric the corresponding optimization problem for maximal margin separation and the corresponding solution are equivalent to the metric maximal margin classification problem in a Hilbert space for the associated Hilbertian metric.

The expression of the optimization problem of the SVM in terms of the (semi)-metric follows from (4.10);

$$\begin{aligned}\left\|\sum_i y_i\alpha_i\Phi_{x_i}\right\|^2_{\mathcal{H}_k} &= \sum_{i,j} y_iy_j\alpha_i\alpha_j k(x_i,x_j)\\ &= \sum_{i,j} y_iy_j\alpha_i\alpha_j[-\frac{1}{2}d^2(x_i,x_j)+g(x_i)+g(x_j)]\\ &= -\frac{1}{2}\sum_{i,j} y_iy_j\alpha_i\alpha_j d^2(x_i,x_j),\end{aligned}$$

where the terms with g vanish due to the constraint $\sum_i y_i\alpha_i = 0$.
The solution expressed in terms of a CPD kernel k can also be expressed in terms of the (semi)-metric by using (4.10):

$$\begin{aligned}f(x) &= \sum_i y_i\alpha_i k(x_i,x)+b = \sum_i y_i\alpha_i[-\frac{1}{2}d(x_i,x)^2+g(x_i)+g(x)]\\ &= -\frac{1}{2}\sum_i y_i\alpha_i d^2(x_i,x)+c,\end{aligned}$$

where again $\sum_i y_i\alpha_i g(x)$ vanishes and $c = b+\sum_i y_i\alpha_i g(x_i)$, but c can also be directly calculated with the average value of $b = y_j+\frac{1}{2}\sum_i y_i\alpha_i d^2(x_i,x_j)$, where j runs over all indices with $\alpha_j > 0$. Since neither the specific isometric embedding Φ nor a corresponding kernel k enter the optimization problem or the solution, the SVM only depends on the (semi)-metric. □

The kernel is sometimes seen as a similarity measure. The last theorem, however, shows that this property of the kernel does not matter for support vector classifiers. On the contrary the (semi)-metric as a dissimilarity measure of the input space only matters for the maximal margin problem. Nevertheless it seems to be easier to construct a conditionally positive definite kernel than a Hilbertian metric, but one should have in mind that only the induced metric has an influence on the solution, and therefore compare two different kernels through their induced metrics. This should also be considered if one uses eigenvalues of the kernel matrix. They depend on the underlying Hilbertian metric and as well on the function $g(x)$ in (4.10) whereas the solution of the SVM only depends on the Hilbertian metric. In other words properties which are not uniform over the class of kernels induced by a semi-metric are not relevant for the solution of the SVM.
One could use the ambiguity in the kernel to chose from the whole class of kernels which induce the same (semi)-metric (4.10) the one which is computationally the cheapest, because the solution does not change as is obvious from the last theorem. Furthermore note that Lemma 4.24 provides a slight refinement of the usual representer theorem of the SVM which states that the solution lies in an at most n dimensional space spanned by the data (see e.g. [81]). This refinement seems to be a marginal effect for large training sets. However the crucial point here is that the constraint on the subspace implies that the SVM is actually equivalent to the metric based maximal margin classifier in a Hilbert space.
As a final note we would like to add that the whole argumentation on the isometric embedding of the (semi)-metric space into a Hilbert space also applies to the

soft-margin-formulation of the SVM. The reformulation in terms of reduced convex hulls is a little bit tricky, and we refer to [19, 12, 109] for this issue.

4.5 Measuring the capacity via Rademacher averages

In this section we compute the Rademacher averages corresponding to the function classes induced by our embeddings. The Rademacher average is a measure of capacity of a function class with respect to classification, and can be used to derive upper bounds on the error of misclassification.

4.5.1 General case

Given a sample of input points $x_1, \ldots, x_n$, we define the empirical Rademacher average $\widehat{R}_n$ of the function class $\mathcal{F}$ as

$$\widehat{R}_n(\mathcal{F}) := E_\sigma \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i)\,, \tag{4.12}$$

where σ are Rademacher variables, that are independent uniform random variables with values $\{-1, +1\}$, and E_σ denotes the expectation conditional to the sample (i.e. with respect to the σ_i only). We repeat here Theorem 7 from [6] to show how the expectation of the Rademacher average can be used to bound the error of misclassification, followed by a Lemma provided in [7] which shows that the empirical Rademacher average is concentrated.

Theorem 4.27 *Let $\mathcal{F}$ be a set of real-valued functions on $\mathcal{X}$ with $\sup\{|f(x)| : f \in \mathcal{F}\} < \infty$ for all $x \in \mathcal{X}$. Suppose that $\Gamma : \mathbb{R} \to [0, 1]$ satisfies $\Gamma(\alpha) \geq \mathbb{1}_{\alpha \leq 0}$ and is Lipschitz with constant L. Then with probability at least $1 - \delta$, every function in $\mathcal{F}$ satisfies*

$$P(Y f(X) \leq 0) \leq \frac{1}{n} \sum_{i}^{n} \Gamma(y_i f(x_i)) + L\, \mathbb{E}\, \widehat{R}_n(\mathcal{F}) + \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}.$$

The next Lemma shows how $\mathbb{E}\, \widehat{R}_n(\mathcal{F})$ can be upper bounded by $\widehat{R}_n(\mathcal{F})$.

Lemma 4.28 *Fix $x > 0$ and let $\mathcal{F}$ be a class of functions with range in $[a, b]$. Then with probability at least $1 - e^{-x}$,*

$$\mathbb{E}\, \widehat{R}_n(\mathcal{F}) \leq \inf_{\alpha \in (0,1)} \left(\frac{1}{1-\alpha} \widehat{R}_n(\mathcal{F}) + \frac{(b-a)x}{4n\alpha(1-\alpha)} \right)$$

Also, with probability $1 - e^{-x}$,

$$\widehat{R}_n(\mathcal{F}) \leq \inf_{\alpha > 0} \left((1+\alpha) \mathbb{E}\, \widehat{R}_n(\mathcal{F}) + \frac{(b-a)x}{2n} \Big(\frac{1}{2\alpha} + \frac{1}{3} \Big) \right)$$

The function classes we are interested in are hyperplanes with a given margin. Now hyperplanes correspond to elements of the dual of the Banach space into which the data is embedded and the margin corresponds to the norm in that space. Therefore

we have to consider the Rademacher averages of balls in the dual space.
For a function $f_{w'}$ in $\mathcal{F}_{\mathcal{B}'}$, $f_{w'}(x) = \langle w', \Phi_x \rangle_{\mathcal{B}',\mathcal{B}}$ with $\|f_{w'}\|_{\mathcal{F}_{\mathcal{B}'}} = \|w'\|_{\mathcal{B}'}$ so that

$$E_\sigma \sup_{\|w'\|_{\mathcal{B}'} \leq B} \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i) = \frac{B}{n} E_\sigma \left\| \sum_{i=1}^n \sigma_i \Phi_{x_i} \right\|_{\mathcal{B}} .$$

Notice that even if the embedding Φ is isometric, the above quantity depends on how the $\Phi(x_i)$ are located in the embedded linear space. So, a priori, the above quantity depends on the embedding and not only on the geometry of the input space.
More precisely, we consider the following two classes. For a given positive definite kernel k, let $\tilde{k}$ be defined as $\tilde{k}(x,y) = k(x,y) - k(x,x_0) - k(x_0,y) + k(x_0,x_0)$[6] and $\mathcal{H}$ be the associated RKHS for $\tilde{k}$. We define $\mathcal{F}_1 = \{g \in \mathcal{H}, \|g\| \leq B\}$. Also, with the notations of the previous section, we define $\mathcal{F}_2 = \{e \in D, \|e\| \leq B\}$.

Theorem 4.29 *With the above notation, we have*

$$\widehat{R}_n(\mathcal{F}_1) \leq \frac{B}{n} \sqrt{\sum_{i=1}^n d(x_i, x_0)^2} .$$

where $d(x_i, x_0) = \|k(x_i, \cdot) - k(x_0, \cdot)\|_{\mathcal{H}}$ is the distance induced by the kernel on $\mathcal{X}$. Also, there exists a universal constant C such that

$$\widehat{R}_n(\mathcal{F}_2) \leq \frac{CB}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\frac{\varepsilon}{2}, \mathcal{X}, d)}\, d\varepsilon.$$

Proof: We first compute the Rademacher average of $\mathcal{F}_2$:

$$\widehat{R}_n(\mathcal{F}_2) = \frac{B}{n} E_\sigma \left\| \sum_{i=1}^n \sigma_i \Phi_{x_i} \right\|_\infty = \frac{B}{n} E_\sigma \sup_{x \in \mathcal{X}} \left| \sum_{i=1}^n \sigma_i \Phi_{x_i}(x) \right| \tag{4.13}$$

We will use Dudley's upper bound on the empirical Rademacher average [30] which states that there exists an absolute constant C for which the following holds: for any integer n, any sample $\{x_i\}_{i=1}^n$ and every class $\mathcal{F}_2$,

$$\widehat{R}_n(\mathcal{F}_2) \leq \frac{C}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{F}_2, \ell_2^n)}\, d\varepsilon, \tag{4.14}$$

where $N(\varepsilon, \mathcal{F}_2, \ell_2^n)$ are the covering numbers of the function class $\mathcal{F}_2$ with respect to the ℓ_2 distance on the data, i.e. $\|f - g\|_{\ell_2^n}^2 := \frac{1}{n} \sum_{i=1}^n (f(x_i) - g(x_i))^2$.
In order to apply this result of Dudley, we notice that the elements of $\mathcal{X}$ can be considered as functions defined on $\mathcal{X}$. Indeed, for each $y \in \mathcal{X}$, one can define the function $f_y : x \mapsto \Phi_x(y)$. We denote by $\mathcal{G}$ the class of all such functions, i.e. $\mathcal{G} = \{f_y : y \in \mathcal{X}\}$. Then using (4.13), we get

$$\widehat{R}_n(\mathcal{F}_2) = B\, E_\sigma \sup_{x \in \mathcal{X}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \Phi_{x_i}(x) \right| = B\, \widehat{R}_n(\mathcal{G}). \tag{4.15}$$

[6] where $k(x_0, \cdot)$ corresponds to the origin in $\mathcal{H}$ and is introduced to make the comparison with the space D easier

We now upper bound the empirical L_2-norm of $\mathcal{G}$:

$$\begin{aligned}\|f_{y_1} - f_{y_2}\|_{\ell_2^n} &\leq \max_{x_i \in T} |\Phi_{x_i}(y_1) - \Phi_{x_i}(y_2)| \\ &= \max_{x_i \in T} |d(x_i, y_1) - d(x_i, y_2) + d(x_0, y_2) - d(x_0, y_1)| \\ &\leq 2d(y_1, y_2). \end{aligned} \tag{4.16}$$

Combining (4.14) and (4.16) we arrive at

$$\widehat{R}_n(\mathcal{G}) \leq \frac{C}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\frac{\varepsilon}{2}, \mathcal{X}, d)}\, d\varepsilon$$

This gives the first result. Similarly, we have

$$\widehat{R}_n(\mathcal{F}_1) = \frac{B}{n} E_\sigma \left\| \sum_{i=1}^n \sigma_i (k(x_i, .) - k(x_0, .)) \right\|_{\mathcal{H}} \leq \frac{B}{n} \sqrt{\sum_{i=1}^n d(x_i, x_0)^2}\,,$$

where the second step follows from Jensen's inequality (applied to the concave function $\sqrt{\cdot}$). □

If we can assume that the data is inside a subset of $\mathcal{X}$ with finite diameter R, then this simplifies to

$$\widehat{R}_n(\mathcal{F}_2) \leq \frac{CB}{\sqrt{n}} \int_0^R \sqrt{\log N(\frac{\varepsilon}{2}, \mathcal{X}, d)}\, d\varepsilon.$$

The above theorem gives an upper bound on the Rademacher average directly in terms of the covering numbers of the metric space $(\mathcal{X}, d)$.

In particular, this shows that the Rademacher average corresponding to the Kuratowski embedding are much smaller than those corresponding to the Lipschitz embedding of [103]. Indeed, for a bounded subset of the metric space $\mathbb{R}^d$, the covering numbers behave like ϵ^{-d} so that the Rademacher average in our case is of order $\sqrt{d/n}$ while in the Lipschitz case it is of order $(1/n)^{1/d}$. Note that in order to establish these results one needs to use a modified version of the metric entropy bound of Dudley, see [104][Chapter 3, Theorem 17], since the integral in Theorem 4.29 diverges in these cases.

Notice that a trivial bound on $\widehat{R}_n(\mathcal{F}_2)$ can be found from (4.13) and

$$\left| \sum_{i=1}^n \sigma_i (d(x_i, x) - d(x_0, x)) \right| \leq \sum_{i=1}^n d(x_i, x_0),$$

which gives the upper bound

$$\widehat{R}_n(\mathcal{F}_2) \leq \frac{B}{n} \sum_{i=1}^n d(x_i, x_0)\,,$$

which is also an upper bound on $\widehat{R}_n(\mathcal{F}_1)$. However, this upper bound is loose since if all the data is at approximately the same distance from x_0 (e.g. on a sphere), then this quantity does not decrease with n. This is undesirable as it would mean that the bound on the error does not decrease when the sample size is increased.

4.5.2 Comparing the approaches

More interesting than upper bounds on the Rademacher averages of the individual algorithms is to compare them directly in the cases where both algorithms can be applied (i.e. when $-d^2$ is conditionally positive definite). In this case, one can choose to embed isometrically the input space either into a Hilbert space or into a Banach space. The question is then how different balls of same radius in the dual spaces are.

Theorem 4.30 *If d is a Hilbertian metric, then*

$$\widehat{R}_n(\mathcal{F}_1) \leq \frac{B}{n}\sqrt{\sum_{i=1}^{n} d(x_i, x_0)^2} \leq \sqrt{2}\widehat{R}_n(\mathcal{F}_2)\,.$$

Proof: We have

$$\begin{aligned}\widehat{R}_n(\mathcal{F}_2) &= \frac{B}{n} E_\sigma \sup_{x\in\mathcal{X}} \left|\sum_{i=1}^{n} \sigma_i \Phi_{x_i}(x)\right| \geq \frac{B}{n} E_\sigma \left|\sum_{i=1}^{n} \sigma_i \Phi_{x_i}(x_0)\right| \\ &\geq \frac{B}{\sqrt{2}n}\sqrt{E_\sigma \sum_{i,j=1}^{n} \sigma_i\sigma_j \Phi_{x_j}(x_0)\Phi_{x_i}(x_0)} \\ &= \frac{1}{\sqrt{2}}\frac{B}{n}\sqrt{\sum_{i=1}^{n} d(x_i, x_0)^2} \geq \frac{1}{\sqrt{2}}\widehat{R}_n(\mathcal{F}_1)\end{aligned}$$

The second step follows from the Khintchine-Kahane inequality. The constant $1/\sqrt{2}$ is optimal, see e.g. [59]. □

This result can be seen as an indication that the SVM is as good as the general algorithm for arbitrary metric spaces in terms of complexity of the unit ball. However, this does not directly allow to compare the generalization abilities of both algorithms. Indeed, the obtained margin in each case could be quite different.

4.6 Conclusion and perspectives

In this article we have built a general framework for the generation of maximal margin algorithms for metric spaces. We considered two general cases. In the first one we trust the metric globally, in the second one we believe only in the local structure of the metric which seems to be often the case for metrics defined on real-world data. In the first case we embed directly isometrically into a Banach space, in the second one we first perform a uniform transformation of the metric such that the local structure is preserved and then embed isometrically the transformed space into a Banach space.

For each metric space we presented a Banach space into which it can be embedded isometrically. It turned out that the optimization problem of the maximal margin algorithm in this Banach space cannot be solved exactly. We provided an approximation which is exact if one considers the training data plus one test point as a finite metric space. One special approximation is the LP-machine for distances of [37].

Since the space of classifiers has a considerably nicer structure if one embeds in a Hilbert space, we considered in the second part isometric embeddings into a Hilbert space. These are no longer possible for all metric spaces, but are restricted to the subclass of Hilbertian metrics. We showed that the resulting algorithm is equivalent to the SVM classifier, but since the relationship between kernels and Hilbertian metrics is many-to-one, the metric based point of view provides a better insight into the structural properties of the SVM.
For the class of Hilbertian metrics we can compare the two isometric embeddings. They both preserve the metric structure, that is, all available information on the data. Therefore the question arises which norm on the linear extension provides the better results in the sense of generalization error. We provided a first answer to this question by comparing the Rademacher averages of both embeddings. It turned out that the Rademacher average of the SVM are upper bounded by a constant times the Rademacher average of the metric based classifier in the Banach space. This result suggests that the SVM has a better generalization performance. But further work has to be done in that direction.

4.7 Appendix

4.7.1 Semi-metric spaces compared to metric spaces for classification

In this article all results were stated for metric spaces. As the following observations show they can be formulated equivalently for semi-metric spaces. In fact there is a connection between both of them which we want to clarify in this appendix.

Theorem 4.31 *Let $(\mathcal{X}, d)$ be a (semi)-metric space and $\sim$ be the equivalence relation defined by $x \sim y \Leftrightarrow d(x,y) = 0$. Then $(\mathcal{X}/\sim, d)$ is a metric space, and if $-d^2(x,y)$ is a conditionally positive definite kernel and k a positive definite kernel on $\mathcal{X}$ which induces d on $\mathcal{X}$, then $-d^2$ is also a conditionally positive definite kernel and k a positive definite kernel on $(\mathcal{X}/\sim, d)$.*

Proof: The property $d(x,y) = 0$ defines an equivalence relation on $\mathcal{X}$, $x \sim y \Longleftrightarrow d(x,y) = 0$. Symmetry follows from the symmetry of d, and transitivity $x \sim y, y \sim z \Rightarrow x \sim z$ follows from the triangle inequality $d(x,z) \leq d(x,y) + d(y,z) = 0$. Then $d(x,y)$ is a metric on the quotient space $\mathcal{X}/\sim$ because all points with zero distance are identified, so

$$d(x,y) = 0 \iff x = y,$$

and obviously symmetry and the triangle inequality are not affected by this operation. d is well-defined because if $x \sim z$ then $|d(x,.) - d(z,.)| \leq d(x,z) = 0$.
The fact that $-d^2$ is conditionally positive definite on $\mathcal{X}/\sim$ follows from the fact that all possible representations of equivalence classes are points in $\mathcal{X}$ and $-d^2$ is conditionally positive definite on $\mathcal{X}$. It is also well defined because if $x \sim z$ then

$$|d^2(x,.) - d^2(z,.)| \leq d(x,z)|(d(x,.) + d(z.,)| = 0.$$

The argumentation that k is also positive definite on $\mathcal{X}/\sim$ is the same as above. It is well defined because if $x \sim x'$ then $\|\Phi_x - \Phi_{x'}\| = 0$, so that actually $k(x,\cdot) = k(x',\cdot)$ (since for all $y \in \mathcal{X}$, $|k(x,y) - k(x',y)| \leq \|\Phi_x - \Phi_{x'}\| \, \|\Phi_y\|$). □

The equivalence relation defined in Theorem 4.31 can be seen as defining a kind of

global invariance on $\mathcal{X}$. For example in the SVM setting when we have the kernel $k(x, y) = \langle x, y \rangle^2$, the equivalence relation identifies all points which are the same up to a reflection. This can be understood as one realization of an action of the discrete group $D = \{-e, +e\}$ on $\mathbb{R}^n$, so this kernel can be understood as a kernel on $\mathbb{R}^n/D$. Assume now that there are no invariances in the data and two different points $x \neq y$ with different labels are such that $d(x, y) = 0$. Then they cannot be separated by any hyperplane. This means that using semi-metrics implicitly assumes invariances in the data, which may not hold.

Kapitel 5

Hilbertian Metrics and Positive Definite Kernels on Probability Measures

5.1 Introduction

Kernel methods have shown in the last years that they are one of the best and generally applicable tools in machine learning. Their great advantage is that positive definite (pd) kernels can be defined on every set. Therefore they can be applied to data of any type. Nevertheless in order to get good results the kernel should be adapted as well as possible to the underlying structure of the input space. This has led in the last years to the definition of kernels on graphs, trees, and manifolds. Kernels on probability measures also belong to this category, but they are already one level higher since they are not defined on the structures directly but on probability measures on these structures[1]. In recent time they have become quite popular due to the following possible applications:

- Direct application on probability measures e.g. histogram data of text [57] and colors [21].
- Given a statistical model for the data, one can first fit the model to the data and then use the kernel to compare two fits, see [57, 51], thereby linking parametric and non-parametric models.
- Given a bounded probability space $\mathcal{X}$, one can use the kernel to compare arbitrary sets in that space, e.g by putting the uniform measure on each set.

In this paper we consider Hilbertian metrics and pd kernels on $\mathcal{M}_+^1(\mathcal{X})^2$. Previous approaches to pd kernels on probability measures were often only defined on a subset of $\mathcal{M}_+^1(\mathcal{X})$ namely a parametric model of probability measures. In the work of Lafferty and Lebanon on diffusion kernels on statistical manifolds[3] in [57] they propose to use the heat kernel on a statistical manifold as a positive definite kernel. Though mathematically beautiful the approach is hard to apply in practice since

[1] However note that one can always use the kernel K on probability measures to define a kernel k on $\mathcal{X}$ by: $k(x, y) = K(\delta_x, \delta_y)$.

[2] $\mathcal{M}_+^1(\mathcal{X})$ denotes the set of positive measures μ on $\mathcal{X}$ with $\mu(\mathcal{X}) = 1$, i.e. the set of probability measures.

[3] A statistical manifold is a parametric set of probability measures which under certain conditions can be seen as a manifold. It becomes a Riemannian manifold by using the Fisher information matrix as Riemannian metric.

the heat kernel can in general not be computed. The authors then suggest to use certain approximations of the heat kernel which however are not guaranteed to be positive definite. In [50], Jebara, Kondor, and Howard introduce a family of so called probability product kernels. One special case is the so called Bhattacharyya kernel which is defined on all probability measures. However the other members of this family do not fulfill this. Then the authors concentrate on computing these kernels for various interesting models of probability measures. Our goal is tackle the problem from a more general perspective and to build Hilbertian metrics and positive definite kernels on the set of all probability measures. The results in this chapter have been partially published in [46, 45].
In the first section we will summarize the close connection between Hilbertian metrics and pd kernels which was already discused in the last chapter.
We will consider two types of kernels on probability measures. The first one is general covariant and defined on the whole set $\mathcal{M}_+^1(\mathcal{X})$. That means that arbitrary smooth coordinate transformations of the underlying probability space will have no influence on the kernel. Such kernels can be applied if only the probability measures themselves are of interest, but not the space they are defined on. We introduce and extend a two parameter family of covariant pd kernels which encompasses all previously used kernels of this type. Despite the great success of these general covariant kernels in text and image classification, they have some shortcomings. For example for some applications we might have a similarity measure resp. a pd kernel on the probability space which we would like to use for the kernel on probability measures. In the second part we investigate types of kernels on probability measures which incorporate such a similarity measure. We provide an alternative descriptions of these kernels which on one hand makes it easier to understand which properties of the measures are effectively used, and on the other hand gives in some cases an efficient way of computing these kernels. Finally we apply these kernels on two text (Reuters and WebKB) and two image classification tasks (Corel14 and USPS).

5.2 Hilbertian Metrics versus Positive Definite Kernels

It is a well-known fact, see Chapter 3, that a pd kernel $k(x, y)$ corresponds to an inner product $\langle \phi_x, \phi_y \rangle_{\mathcal{H}}$ in some feature space $\mathcal{H}$. The class of conditionally positive definite (cpd) kernels is less well known. Nevertheless this class is of great interest since Schölkopf showed in [82] that all translation invariant kernel methods can also use the bigger class of cpd kernels. Therefore we give a short summary of this type of kernels and their connection to Hilbertian metrics[4].

Definition 5.1 *A real valued function k on $\mathcal{X} \times \mathcal{X}$ is pd (resp. cpd) if and only if k is symmetric and $\sum_{i,j}^n c_i c_j k(x_i, x_j) \geq 0$, for all $n \in \mathbb{N}$, $x_i \in \mathcal{X}, i = 1, ..., n$, and for all $c_i \in \mathbb{R}, i = 1, ..., n$, (resp. for all $c_i \in \mathbb{R}, i = 1, ..., n$, with $\sum_i^n c_i = 0$).*

Note that every pd kernel is also cpd. The close connection between the two classes is shown by the following lemma:

Lemma 5.2 *[14] Let k be a kernel defined as $k(x, y) = \hat{k}(x, y) - \hat{k}(x, x_0) - \hat{k}(x_0, y) + \hat{k}(x_0, x_0)$, where $x_0 \in \mathcal{X}$. Then k is pd if and only if $\hat{k}$ is cpd.*

[4] A semi-metric $d(x, y)$ fulfills the conditions of a metric except that $d(x, y) = 0$ does not imply $x = y$. It is called Hilbertian if one can embed the (semi)-metric space $(\mathcal{X}, d)$ isometrically into a Hilbert space. A (semi)-metric d is Hilbertian if and only if $-d^2(x, y)$ is cpd. That is a classical result of Schoenberg, see [80].

Similar to pd kernels one can also characterize cpd kernels. Namely one can write all cpd kernels in the form: $k(x,y) = -\frac{1}{2}\left\|\phi_x - \phi_y\right\|^2_{\mathcal{H}} + f(x) + f(y)$. The cpd kernels corresponding to Hilbertian (semi)-metrics are characterized by $f(x) = 0$ for all $x \in \mathcal{X}$, whereas if k is pd it follows that $f(x) = \frac{1}{2}k(x,x) \geq 0$. We refer to Chapter 4 for further details. We also would like to point out that for SVM's the class of Hilbertian (semi)-metrics is in a sense more important than the class of pd kernels. Namely as we have shown in the previous chapter the solution and optimization problem of the SVM only depends on the Hilbertian (semi)-metric which is implicitly defined by each pd kernel. Moreover a whole family of pd kernels induces the same semi-metric. In order to avoid confusion we will in general speak of Hilbertian metrics since, using Lemma 5.2, one can always define a corresponding pd kernel. Nevertheless for the convenience of the reader we will often explicitly state the corresponding pd kernels.

5.3 γ-homogeneous Hilbertian Metrics and Positive Definite Kernels on $\mathbb{R}_+$

The class of Hilbertian metrics on probability measures we consider in this chapter are based on a pointwise comparison of the densities $p(x)$ using a Hilbertian metric on $\mathbb{R}_+$. Therefore Hilbertian metrics on $\mathbb{R}_+$ are the basic ingredient of our approach. In principle we could use any Hilbertian metric on $\mathbb{R}_+$, but, as we will explain later, we require the metric on probability measures to have a certain property. This in turn requires that the Hilbertian metric on $\mathbb{R}_+$ is γ-homogeneous[5]. The class of γ-homogeneous Hilbertian metrics on $\mathbb{R}_+$ was recently characterized by Fuglede:

Theorem 5.3 (Fuglede [34]) *A continuous, symmetric function $d : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ with $d(x,y) = 0 \iff x = y$ is a γ-homogeneous Hilbertian metric d on $\mathbb{R}_+$ if and only if there exists a (necessarily unique) non-zero bounded measure $\rho \geq 0$ on $\mathbb{R}_+$ such that d^2 can be written as*

$$d^2(x,y) = \int_{\mathbb{R}_+} \left|x^{(\gamma+i\lambda)} - y^{(\gamma+i\lambda)}\right|^2 d\rho(\lambda). \tag{5.1}$$

Using Lemma 5.2, we define the corresponding class of pd kernels on $\mathbb{R}_+$ by choosing $x_0 = 0$. We will see later that this corresponds to choosing the zero-measure as origin of the RKHS.

Corollary 5.4 *A continuous, symmetric function $k : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ with $k(x,x) = 0 \iff x = 0$ is a 2γ-homogeneous pd kernel k on $\mathbb{R}_+$ if and only if there exists a (necessarily unique) non-zero bounded symmetric measure $\kappa \geq 0$ on $\mathbb{R}$ such that k is given as*

$$k(x,y) = \int_{\mathbb{R}} x^{(\gamma+i\lambda)} y^{(\gamma-i\lambda)}\, d\kappa(\lambda). \tag{5.2}$$

Proof: If k has the form given in (5.2), then it is obviously 2γ-homogeneous and since $k(x,x) = x^{2\gamma}\kappa(\mathbb{R})$ we have $k(x,x) = 0 \iff x = 0$. The other direction follows by first noting that $k(0,0) = \langle \phi_0, \phi_0 \rangle = 0$ and then applying theorem 5.3, where κ is the symmetrized version of ρ around the origin, together with lemma 5.2 and $k(x,y) = \langle \phi_x, \phi_y \rangle = \frac{1}{2}\left(-d^2(x,y) + d^2(x,0) + d^2(y,0)\right)$. □

[5] A symmetric function k on $\mathbb{R}_+ \times \mathbb{R}_+$ is γ-homogeneous if $k(c\,x, c\,y) = c^\gamma k(x,y)$ for all $c \in \mathbb{R}_+$

At first glance Theorem 5.3, though mathematically beautiful, seems not to be very helpful from the viewpoint of applications. But as we will show in the section on structural pd kernels on $\mathcal{M}_+^1(\mathcal{X})$, this result allows us to compute this class of kernels very efficiently.
Recently Topsøe and Fuglede proposed an interesting two-parameter family of Hilbertian metrics on $\mathbb{R}_+$ [98, 34]. We extend now the parameter range of this family. This allows us in the next section to recover all previously used Hilbertian metrics on $\mathcal{M}_+^1(\mathcal{X})$ from this family.

Theorem 5.5 *The function $d : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ defined as:*

$$d^2_{\alpha|\beta}(x,y) = \frac{|\alpha\beta|}{\alpha-\beta}\left[\left(\frac{x^\alpha+y^\alpha}{2}\right)^{\frac{1}{\alpha}} - \left(\frac{x^\beta+y^\beta}{2}\right)^{\frac{1}{\beta}}\right] \tag{5.3}$$

is a $1/2$-homogeneous Hilbertian metric on $\mathbb{R}_+$, if $\alpha \in [1,\infty]$, $\beta \in [\frac{1}{2},\alpha]$ or $\beta \in [-\infty,-1]$. Moreover the pointwise limit for $\alpha \to \beta$ is given as:

$$\lim_{\alpha\to\beta} d^2_{\alpha|\beta}(x,y) = \beta^2 \frac{\partial}{\partial\beta}\left(\frac{x^\beta+y^\beta}{2}\right)^{\frac{1}{\beta}} =$$
$$\left(\frac{x^\beta+y^\beta}{2}\right)^{\frac{1}{\beta}}\left[\frac{x^\beta}{x^\beta+y^\beta}\log\left(\frac{2x^\beta}{x^\beta+y^\beta}\right) + \frac{y^\beta}{x^\beta+y^\beta}\log\left(\frac{2y^\beta}{x^\beta+y^\beta}\right)\right].$$

Note that $d^2_{\alpha|\beta} = d^2_{\beta|\alpha}$. Before we give the proof let us give the following equivalent description for negative values of β. We introduce the parameter $\rho = -\beta$ and get for $1 \le \rho \le \infty$,

$$d^2_{\alpha|\rho}(x,y) = \frac{\alpha\rho}{\alpha+\rho}\left[\left(\frac{x^\alpha+y^\alpha}{2}\right)^{\frac{1}{\alpha}} - xy\left(\frac{2}{x^\rho+y^\rho}\right)^{\frac{1}{\rho}}\right].$$

We need the following lemmas in the proof:

Lemma 5.6 *[14, 2.10] If $k : \mathcal{X}\times\mathcal{X}$ is cpd and $k(x,x) \le 0$, $\forall x \in \mathcal{X}$, then $-(-k)^\gamma$ is also cpd for $0 < \gamma \le 1$.*

Lemma 5.7 *If $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is cpd and $k(x,y) < 0$, $\forall\, x,y \in \mathcal{X}$, then $-1/k$ is pd.*

Proof: It follows from Theorem 2.3 in [14] that if $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}_-$ is cpd, then $1/(t-k)$ is pd for all $t > 0$. The pointwise limit of a sequence of cpd resp. pd kernels is cpd resp. pd if the limit exists, see e.g. [81]. Therefore $\lim_{t\to 0} 1/(t-k) = -1/k$ is positive definite if k is strictly negative. □

We can now prove Theorem 5.5:

Proof: The proof for the symmetry, the limit $\alpha \to \beta$ and the parameter range $1 \le \alpha \le \infty$, $1/2 \le \beta \le \alpha$ can be found in [34]. We prove that $-d^2_{\alpha|\beta}$ is cpd for $1 \le \alpha \le \infty$, $-\infty \le \beta \le -1$. First note that $k(x,y) = -(f(x)+f(y))$ is cpd on $\mathbb{R}_+$, for any function $f : \mathbb{R}_+ \to \mathbb{R}_+$ and satisfies $k(x,y) \le 0$, $\forall\, x,y \in \mathcal{X}$. Therefore by Lemma 5.6, $-(x^\alpha+y^\alpha)^{1/\alpha}$ is cpd for $1 \le \alpha < \infty$. The pointwise limit $\lim_{\alpha\to\infty} -(x^\alpha+y^\alpha)^{1/\alpha} = -\max\{x,y\}$ exists, therefore we can include the limit $\alpha = \infty$. Next we consider $k(x,y) = -(x+y)^{1/\beta}$ for $1 \le \beta \le \infty$ which is

cpd as we have shown and strictly negative if we restrict k to $\mathbb{R}_+^* \times \mathbb{R}_+^*$. Then all conditions for lemma 5.7 are fulfilled, so that $k(x,y) = (x+y)^{-1/\beta}$ is pd. But then also $k(x,y) = (x^{-\beta} + y^{-\beta})^{-1/\beta}$ is pd. Moreover k can be continuously extended to 0 by $k(x,y) = 0$ for $x = 0$ or $y = 0$. Multiplying both parts with the positive factor $\frac{|\alpha\beta|}{\alpha-\beta}$ and adding them gives the result. By construction $-d^2_{\alpha|\beta}$ is then cpd and $d_{\alpha|\beta}$ a semi-metric since

$$d^2_{\alpha|\beta}(x,x) = 0, \quad \forall\, x \in \mathbb{R}_+.$$

Now $d_{\alpha|\beta}(x,y) = 0$ implies $x = y$ since

$$\sqrt{\frac{|\alpha\beta|}{2^{\frac{1}{\alpha}}(\alpha-\beta)}}\sqrt{x} = d(x,0) \le d(x,y) + d(y,0) = \sqrt{\frac{|\alpha\beta|}{2^{\frac{1}{\alpha}}(\alpha-\beta)}}\sqrt{y}$$

so that $\sqrt{x} \le \sqrt{y}$. Vice versa we get $\sqrt{y} \le \sqrt{x}$ which implies $\sqrt{x} = \sqrt{y}$ so that $d_{\alpha|\beta}$ is a metric. □

It is easy then to define with the help of Lemma 5.2 a corresponding family of positive definite kernels $k_{\alpha|\beta}$. We fix the family by requiring that $k_{\alpha|\beta}(0,0) = 0$ which corresponds to the choice $x_0 = 0$ in Lemma 5.2 so that the origin in $\mathbb{R}_+$ is mapped to the origin in the corresponding RKHS.

Corollary 5.8 *The function $k : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ defined as:*

$$k_{\alpha|\beta}(x,y) = \frac{\alpha\beta}{\alpha-\beta}\left[\left(2^{-\frac{1}{\alpha}} - 2^{-\frac{1}{\beta}}\right)(x+y) - \left(\frac{x^\alpha + y^\alpha}{2}\right)^{1/\alpha} + \left(\frac{x^\beta + y^\beta}{2}\right)^{1/\beta}\right]$$

is a 1-homogeneous positive definite kernel on $\mathbb{R}_+$, if $\alpha \in [1,\infty]$, $\beta \in [\frac{1}{2}, \alpha]$. For $\alpha = \beta$ one gets

$$\begin{aligned} k_{\beta|\beta}(x,y) =& \log(2)\frac{x+y}{2^{\frac{1}{\beta}}} \\ &- \left(\frac{x^\beta + y^\beta}{2}\right)^{\frac{1}{\beta}}\left[\frac{x^\beta}{x^\beta+y^\beta}\log\left(\frac{2x^\beta}{x^\beta+y^\beta}\right) + \frac{y^\beta}{x^\beta+y^\beta}\log\left(\frac{2y^\beta}{x^\beta+y^\beta}\right)\right]. \end{aligned}$$

For negative values of β we introduce $\rho = -\beta$. Then for $\alpha \in [1,\infty]$, $\rho \in [1,\infty]$

$$k_{\alpha|-\rho}(x,y) = \frac{\alpha\rho}{\alpha+\rho}\left[\frac{x+y}{2^{\frac{1}{\alpha}}} - \left(\frac{x^\alpha + y^\alpha}{2}\right)^{\frac{1}{\alpha}} + xy\left(\frac{2}{x^\rho + y^\rho}\right)^{\frac{1}{\rho}}\right]$$

is a 1-homogeneous positive definite kernel on $\mathbb{R}_+$.

5.4 Covariant Hilbertian Metrics on $\mathcal{M}_+^1(\mathcal{X})$

In this section we define Hilbertian metrics on $\mathcal{M}_+^1(\mathcal{X})$ by comparing the densities pointwise with a Hilbertian metric on $\mathbb{R}_+$ and integrating these distances over $\mathcal{X}$. Since densities can only be defined with respect to a dominating measure[6] our definition will at first depend on the choice of the dominating measure. The final goal is

[6] A measure μ dominates a measure ν if $\mu(E) > 0$ whenever $\nu(E) > 0$ for all measurable sets $E \subset \mathcal{X}$. In $\mathbb{R}^n$ the dominating measure μ is usually the Lebesgue measure.

however to become independent of the dominating measure so that we get a metric on the set of all probability measures. The dependence on the dominating measure would restrict the applicability of our approach since there exists no universal dominating measure which dominates all probability measures. For example if we had $\mathcal{X} = \mathbb{R}^n$ and choose the dominating measure μ to be the Lebesgue measure, then we could not deal with Dirac measures δ_x since they are not dominated by the Lebesgue measure.
Therefore we construct the Hilbertian metric such that it is independent of the dominating measure. This justifies the term 'covariant' since independence from the dominating measure also yields invariance from arbitrary one-to-one coordinate transformations. In turn this also implies that all structural properties of the probability space will be ignored so that the metric on $\mathcal{M}_+^1(\mathcal{X})$ only depends on the probability measures and not on properties of $\mathcal{X}$. As an example take the color histograms of images. Covariance here means that the choice of the underlying color space say RGB, HSV, or CIE Lab does not influence our metric since these color spaces are all related by one-to-one coordinate transformations. Note however that in practice the results will usually slightly differ due to different discretizations of the color space.
In order to simplify the notation we define $p(x)$ to be the Radon-Nikodym derivative $(dP/d\mu)(x)$ [7] of P with respect to the dominating measure μ.

Proposition 5.9 *Let P and Q be two probability measures on $\mathcal{X}$, μ an arbitrary dominating measure[8] of P and Q, and $d_{\mathbb{R}_+}$ a 1/2-homogeneous Hilbertian metric on $\mathbb{R}_+$. Then $D_{\mathcal{M}_+^1(\mathcal{X})}$ defined as*

$$D^2_{\mathcal{M}_+^1(\mathcal{X})}(P,Q) := \int_{\mathcal{X}} d^2_{\mathbb{R}_+}(p(x),q(x))d\mu(x)\,, \tag{5.4}$$

is a Hilbertian metric on $\mathcal{M}_+^1(\mathcal{X})$. $D_{\mathcal{M}_+^1(\mathcal{X})}$ is independent of the dominating measure μ.

Proof: First we show by using the 1/2-homogeneity of $d_{\mathbb{R}_+}$ that $d_{\mathcal{M}_+^1(\mathcal{X})}$ is independent of the dominating measure μ. We have

$$\int_{\mathcal{X}} d^2_{\mathbb{R}_+}(\frac{dP}{d\mu},\frac{dQ}{d\mu})d\mu = \int_{\mathcal{X}} d^2_{\mathbb{R}_+}(\frac{dP}{d\nu}\frac{d\nu}{d\mu},\frac{dQ}{d\nu}\frac{d\nu}{d\mu})\frac{d\mu}{d\nu}d\nu = \int_{\mathcal{X}} d^2_{\mathbb{R}_+}(\frac{dP}{d\nu},\frac{dQ}{d\nu})d\nu$$

where we use that $d^2_{\mathbb{R}_+}$ is 1-homogeneous. It is easy to show that $-d^2_{\mathcal{M}_+^1(\mathcal{X})}$ is conditionally positive definite, simply take for every $n \in \mathbb{N}$, $P_1,\ldots,P_n$ the dominating measure $\frac{\sum_{i=1}^n P_i}{n}$ and use that $-d^2_{\mathbb{R}_+}$ is conditionally positive definite. □

Remark: By plugging into Equation 5.4 arbitrary Hilbertian metrics on $\mathbb{R}_+$, one can also define Hilbertian metrics on a subset of probability measures. Namely the subset of probability measures which are dominated by μ. The problem with such a definition is that the domain of the metric as well as the metric itself depend on the choice of the dominating measure. There might exist cases where from a practical point of view this is even desired. However from a theoretical point of view the metric should only measure differences in the assignment of probability mass on $\mathcal{X}$.

[7] In case of $\mathcal{X} = \mathbb{R}^n$ and when μ is the Lebesgue measure we can think of $p(x)$ as the normal density function.
[8] Such a dominating measure always exists take e.g. $M = (P+Q)/2$

Therefore it should not depend on the dominating measure or roughly equivalent the choice of coordinates since this introduces an implicit dependence on the description of the space $\mathcal{X}$.
Using the same principle one can define positive definite kernels on probability measures.

Proposition 5.10 *Let P and Q be two probability measures on $\mathcal{X}$, μ an arbitrary dominating measure of P and Q, and k a 1-homogeneous continuous positive definite kernel on $\mathbb{R}_+$. Then K defined as*

$$K(P,Q) := \int_{\mathcal{X}} k\big(p(x), q(x)\big) d\mu(x)\,, \tag{5.5}$$

is a positive definite kernel on $\mathcal{M}_+^1(\mathcal{X})$ and K is independent of the dominating measure μ.

The beautiful Theorem 5.3 of Fuglede allows us now to characterize all Hilbertian metrics on probability measures of the type introduced in Equation 5.4.

Theorem 5.11 *All continuous, covariant Hilbertian metrics d on $\mathcal{M}_+^1(\mathcal{X})$ of the form* (5.4) *are given up to constants as:*

$$d^2(P,Q) \quad = \frac{1}{\rho(\mathbb{R}_+)} \int_{\mathcal{X}} \int_{\mathbb{R}_+} \left| x^{(\frac{1}{2}+i\lambda)} - y^{(\frac{1}{2}+i\lambda)} \right|^2 d\rho(\lambda) d\mu(x)\,, \tag{5.6}$$

where ρ is non-zero bounded measure on $\mathbb{R}_+$.

As a corollary one can derive a similar result for covariant kernels on $\mathcal{M}_+^1(\mathcal{X})$.

Corollary 5.12 *All continuous, covariant positive definite kernels on $\mathcal{M}_+^1(\mathcal{X})$ defined in* (5.5) *with $K(P,P) = 0 \Longleftrightarrow P = 0$ are up to constants of the form:*

$$K(P,Q) = \frac{1}{\rho(\mathbb{R}_+)} \int_{\mathcal{X}} \int_{\mathbb{R}_+} Re(p(x)^{(\frac{1}{2}+i\lambda)} \overline{q(x)^{(\frac{1}{2}+i\lambda)}}) d\rho(\lambda) d\mu(x). \tag{5.7}$$

where ρ is non-zero bounded measure on $\mathbb{R}_+$.

We would like to note that Proposition 5.9, Theorem 5.11 and Corollary 5.12 can be easily extended to all bounded positive measures $\mathcal{M}_+^b(\mathcal{X})$ on $\mathcal{X}$. It can be seen from Corollary 5.12 that

$$K(P,P) = \frac{1}{\rho(\mathbb{R}_+)} \int_{\mathcal{X}} \int_{\mathbb{R}_+} p(x)\, d\rho(\lambda) d\mu(x) = P(\mathcal{X})$$

so that measures P with equal mass of $\mathcal{X}$ are mapped to the sphere in the corresponding RKHS with radius $r = P(\mathcal{X})$.
The principle introduced in Proposition 5.9 can also be applied to the two-parameter family $d_{\alpha|\beta}$ of 1/2-homogeneous Hilbertian metrics on $\mathbb{R}_+$ in order to get the corresponding family of covariant Hilbertian metrics $D_{\alpha|\beta}$ on $\mathcal{M}_+^1(\mathcal{X})$.
As special cases we get the following well-known measures on probability distribu-

tions. Note that these metrics are not normalized as in Theorem 5.11.

$$
\begin{aligned}
D^2_{1|-1}(P,Q) =& \frac{1}{4}\int_{\mathcal{X}} \frac{(p(x)-q(x))^2}{p(x)+q(x)} d\mu(x),\\
D^2_{\frac{1}{2}|1}(P,Q) =& \frac{1}{4}\int_{\mathcal{X}} (\sqrt{p(x)}-\sqrt{q(x)})^2 d\mu(x),\\
D^2_{1|1}(P,Q) =& \frac{1}{2}\int_{\mathcal{X}} p(x)\log\left[\frac{2p(x)}{p(x)+q(x)}\right] + q(x)\log\left[\frac{2q(x)}{p(x)+q(x)}\right] d\mu(x),\\
D^2_{\infty|1}(P,Q) =& \frac{1}{2}\int_{\mathcal{X}} |p(x)-q(x)| d\mu(x).
\end{aligned} \tag{5.8}
$$

$D^2_{1|-1}$ is the symmetric χ^2-measure, $D_{\frac{1}{2}|1}$ the Hellinger distance, $D^2_{1|1}$ the Jensen-Shannon divergence and $D^2_{\infty|1}$ the total variation. The symmetric χ^2-metric was for some time wrongly assumed to be pd and is new in this family due to our extension of $d^2_{\alpha|\beta}$ to negative values of β. The Hellinger metric is well known in the statistics community and was for example used in [51, 50] respectively the induced positive definite Bhattacharyya kernel. The total variation was implicitly used in SVM's through a pd counterpart which we will give below. Finally the Jensen-Shannon divergence is very interesting since it is a symmetric and smoothed variant of the Kullback-Leibler divergence. The Jensen-Shannon (JS) divergence, see [61] for some basic properties, attracted recently some interest since it was proven by Endres and Schindelin [31] that the JS-divergence is a square of a metric and shortly afterwards Fuglede and Topsøe showed in [33] that this metric is Hilbertian. Instead of the work in [67] where they have a heuristic approach to get from the Kullback-Leibler divergence to a pd matrix, the Jensen-Shannon divergence is a theoretically sound alternative. As a final remark we would like to note that the four special cases of metrics on $\mathcal{M}^1_+(\mathcal{X})$ can all be written as f-divergences[9], see [97, 98]. Up to $D^2_{\infty|1}(P,Q)$ they are all smooth and fulfill $f(1)=0$. Then it is a standard result in information geometry [3] that all smooth f-divergences with $f(1)=0$ induce up to constants the same Riemannian metric on a statistical manifold. That implies that locally all these metrics are equivalent for a given parametric model.
For completeness we also give the corresponding pd kernels on $\mathcal{M}^1_+(\mathcal{X})$ where we take in Lemma 5.2 the zero measure as x_0 in $\mathcal{M}^1_+(\mathcal{X})$. This choice seems strange for probability measures. But as noted before the whole framework presented in this chapter can easily be extended to all positive, bounded measures $\mathcal{M}^b_+(\mathcal{X})$ on $\mathcal{X}$. For

[9] Let f be a convex function. Then for $P,Q \in \mathcal{M}^1_+(\mathcal{X})$ the f-divergence $D(P,Q)$ is defined as

$$D(P,Q) = \int_{\mathcal{X}} f\left(\frac{dP}{dQ}\right) dQ$$

the set $\mathcal{M}_+^b(\mathcal{X})$ the zero measure is a natural choice of the origin.

$$\begin{aligned}
K_{1|-1}(P,Q) &= \int_{\mathcal{X}} \frac{p(x)q(x)}{p(x)+q(x)} d\mu(x), \\
K_{\frac{1}{2}|1}(P,Q) &= \frac{1}{2} \int_{\mathcal{X}} \sqrt{p(x)q(x)} d\mu(x), \\
K_{1|1}(P,Q) &= -\frac{1}{2} \int_{\mathcal{X}} p(x) \log\left(\frac{p(x)}{p(x)+q(x)}\right) + q(x) \log\left(\frac{q(x)}{p(x)+q(x)}\right) d\mu(x), \\
K_{\infty|1}(P,Q) &= \int_{\mathcal{X}} \min\{p(x), q(x)\} d\mu(x), \qquad (5.9)
\end{aligned}$$

where μ is an arbitrary dominating measure of P and Q. For the derivation of the last kernel $K_{\infty|1}(P,Q)$ note that $|x-y| = \max\{x,y\} - \min\{x,y\}$.
The astonishing fact is that we find the four (partially) previously used Hilbertian metrics resp. pd kernels on $\mathcal{M}_+^1(\mathcal{X})$ as special cases of a two-parameter family of Hilbertian metrics resp. pd kernels on $\mathcal{M}_+^1(\mathcal{X})$. Due to the symmetry of $d^2_{\alpha|\beta}$ (which implies symmetry of $D^2_{\alpha|\beta}$) we can even see all of them as special cases of the family restricted to $\alpha = 1$. This on the one hand shows the close relation of these metrics among each other and on the other hand gives us the opportunity to do model selection in this one-parameter family of Hilbertian metrics. Yielding an elegant way to handle both the known similarity measures and intermediate ones in the same framework.

5.5 Structural Positive Definite Kernels

The covariant Hilbertian metrics proposed in the last section have the advantage that they only compare the probability measures, thereby ignoring all structural properties of the space $\mathcal{X}$ where the measures are defined on. On the other hand there exist cases where we have a reasonable similarity measure on the space $\mathcal{X}$, which we would like to be incorporated into the metric. We will consider in this section two ways of doing this.

5.5.1 Structural Kernel I

To incorporate structural information about the probability space $\mathcal{X}$ is helpful, when we compare probability measures with disjoint support. For the covariant metrics disjoint measures have always maximal distance, irrespectively how closeör fartheir support is. Obviously if our training set consists only of disjoint measures, learning is not possible with covariant metrics.
If one has a metric or similarity measures on $\mathcal{X}$, one can use it to measure the distance resp. similarity between disjoint measures. One prominent example of a metric on probability measure using such information is the Kantorovich metric also known as Wasserstein distance, see [101].

Definition 5.13 (Kantorovich metric) *Let $(\mathcal{X}, d_{\mathcal{X}})$ be a complete, separable and bounded metric space. Then the Kantorovich metric d_K on $\mathcal{M}_+^1(\mathcal{X})$ is defined as:*

$$d_K(P,Q) = \inf_{\nu} \left\{ \int_{\mathcal{X}\times\mathcal{X}} d(x,y)\, d\nu(x,y) \Big| \nu \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{X}), \pi_1(\nu) = P, \pi_2(\nu) = Q \right\}$$

where π_i denotes the marginal with respect to i-th coordinate.

There exist various generalizations by replacing the metric with more general cost functions. When $\mathcal{X}$ is finite, the Kantorovich metric is also known as Earthmover distance.
In the same spirit we propose here a positive definite kernel which incorporates a given similarity measure, namely a pd kernel, on the probability space. The only disadvantage is that this kernel is not invariant with respect to the dominating measure. That means we can only define it for the subset $\mathcal{M}_+^1(\mathcal{X}, \mu) \subset \mathcal{M}_+^1(\mathcal{X})$ of measures dominated by μ. On the other hand in some cases one has anyway a preferred measure, or one is only interested in a kernel on a parametric model which is dominated by a certain measure. Such a preferred measure is then a natural choice for the dominating measure so that practically it does not seem to be a major restriction. For our experiments it does not make any difference since we anyway use only probabilities over finite spaces, so that the uniform measure dominates all other measures and therefore $\mathcal{M}_+^1(\mathcal{X}, \mu) \equiv \mathcal{M}_+^1(\mathcal{X})$.

Theorem 5.14 (Structural Kernel I) *Let k be a bounded PD kernel on $\mathcal{X}$ and $\hat{k}$ a bounded PD kernel on $\mathbb{R}_+$. Then*

$$K_I(P, Q) = \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y)\, \hat{k}(p(x), q(y))\, d\mu(x)\, d\mu(y) \tag{5.10}$$

is a pd kernel on $\mathcal{M}_+^1(\mathcal{X}, \mu) \times \mathcal{M}_+^1(\mathcal{X}, \mu)$.

Proof: Note first that the product $k(x, y)\hat{k}(r, s)$ $(x, y \in \mathcal{X}, r, s \in \mathbb{R}_+)$ is a positive definite kernel on $\mathcal{X} \times \mathbb{R}_+$. The corresponding RKHS $\mathcal{H}$ is the tensor product of the RKHS $\mathcal{H}_k$ and $\mathcal{H}_{\hat{k}}$, that is $\mathcal{H} = \mathcal{H}_k \otimes \mathcal{H}_{\hat{k}}$. We denote the corresponding feature map by $(x, r) \to \phi_x \otimes \psi_r$. Now let us define a linear map $L_q : \mathcal{H} \to \mathbb{R}$ by

$$\begin{aligned} L_q : \phi_x \otimes \psi_r \longrightarrow \quad & \int_{\mathcal{X}} k(x, y)\hat{k}(r, q(y))d\mu(y) = \int_{\mathcal{X}} \langle \phi_x, \phi_y \rangle_{\mathcal{H}_k} \left\langle \psi_r, \psi_{q(y)} \right\rangle_{\mathcal{H}_{\hat{k}}} d\mu(y) \\ & \leq \|\phi_x \otimes \psi_r\|_{\mathcal{H}} \int_{\mathcal{X}} \left\|\phi_y \otimes \psi_{q(y)}\right\|_{\mathcal{H}} d\mu(y) \end{aligned}$$

Therefore by the assumption L_q is continuous. By the Riesz lemma, there exists a vector u_q such that $\forall v \in \mathcal{H}$, $\langle u_q, v \rangle_{\mathcal{H}} = L_q(v)$. It is obvious from

$$\begin{aligned} \langle u_p, u_q \rangle_{\mathcal{H}} &= \int_{\mathcal{X}} \left\langle u_p, \phi_y \otimes \psi_{q(y)} \right\rangle_{\mathcal{H}} d\mu(y) = \int_{\mathcal{X}^2} \left\langle \phi_x \otimes \psi_{p(x)}, \phi_y \otimes \psi_{q(y)} \right\rangle_{\mathcal{H}} d\mu(y) d\mu(x) \\ &= \int_{\mathcal{X}^2} k(x, y)\, \hat{k}(p(x), q(y))\, d\mu(x)\, d\mu(y) \end{aligned}$$

that K is positive definite. □

Note that this kernel can easily be extended to all bounded, signed measures. This structural kernel generalizes previous work done by Suquet, in [91, 92, 93], where the special case with $\hat{k}(p(x), q(y)) = p(x)q(y)$ has been considered. The advantage of this choice for $\hat{k}$ is that $K_I(P, Q)$ becomes independent of the dominating measure. In fact it is easy to see that among the family of structural kernels $K_I(P, Q)$ of the form (5.10) this choice of $\hat{k}$ yields the only structural kernel $K(P, Q)$ which is

independent of the dominating measure. Indeed for independence bilinearity of $\hat{k}$ is required which yields $\hat{k}(x,y) = xy\,\hat{k}(1,1)$.
The structural kernel has the disadvantage that the computational cost increases dramatically compared to the covariant one, since one has to integrate twice over $\mathcal{X}$. An implementation seems therefore only to be possible for either very localized probability measures or a sharply concentrated similarity kernel $\hat{k}$ e.g. a compactly supported radial basis function on $\mathbb{R}^n$.
The following equivalent representation of this kernel will provide a better understanding and at the same time will show a way to reduce the computational cost considerably. The same construction has been done by Suquet in [93] with $\hat{k}(p(x), q(y)) = p(x)q(y)$.

Proposition 5.15 *Let k and $\hat{k}$ be two bounded kernels on $\mathcal{X}\times\mathcal{X}$ resp. $\mathbb{R}_+\times\mathbb{R}_+$ which can be written as*

$$k(x,y) = \int_T \Gamma(x,t)\overline{\Gamma(y,t)}d\omega(t),$$
$$\hat{k}(p(x),q(y)) = \int_S \Psi(p(x),\lambda)\overline{\Psi(q(y),\lambda)}d\kappa(\lambda).$$

where ω resp. κ are σ-finite measures. Then the structural kernel $K_I(P,Q)$ can be equivalently written as the inner product in $L_2(T,\omega)\otimes L_2(S,\kappa)$:

$$K_I(P,Q) = \int_T\int_S \phi_P(t,\lambda)\overline{\phi_Q(t,\lambda)}\,d\kappa(\lambda)d\omega(t)$$

for some sets T, S with the feature map:

$$\phi : \mathcal{M}_+^1(\mathcal{X},\mu) \to L_2(T,\omega)\otimes L_2(S,\kappa),$$
$$P \to \phi_P(t,\lambda) = \int_{\mathcal{X}} \Gamma(x,t)\Psi(p(x),\lambda)d\mu(x).$$

Proof: First note that one can write every pd kernel in the form:

$$k(x,y) = \langle \Gamma(x,\cdot), \Gamma(y,\cdot)\rangle_{L_2(T,\omega)} = \int_T \Gamma(x,t)\overline{\Gamma(y,t)}d\omega(t),$$

where $\Gamma(x,\cdot)\in L_2(T,\mu)$ for each $x\in\mathcal{X}$. In general the space T is very big since one can show that such a representation always exists in $L_2(\mathbb{R}^{\mathcal{X}},\mu)$, see e.g. [49]. For the product of two positive definite kernels we have such a representation on the set $T\times S$. The rest of the argument follows by applying Fubini-Tonelli's theorem several times. First note that

$$k(x,x)\hat{k}(s,s) = \int_{T\times S} |\Gamma(x,t)\Psi(s,\lambda)|^2 d(\omega\times\kappa)(t,\lambda) \le C < \infty$$

where $C = \sup_{x\in\mathcal{X}, s\in\mathbb{R}_+} k(x,x)\hat{k}(s,s)$. Now introduce $\Phi(x,s,\lambda,t) = \Gamma(x,t)\Psi(s,\lambda)$

$$\int_{\mathcal{X}\times\mathcal{X}}\int_{T\times S} |\Phi(x,r,\lambda,t)\overline{\Phi(y,s,\lambda,t)}|\,d(\omega\times\kappa)(t,\lambda)\,d(\mu\times\mu)(x,y)$$
$$\le \int_{\mathcal{X}\times\mathcal{X}} \Big(\int_{T\times S} |\Phi(x,r,\lambda,t)|^2 d(\omega\times\kappa)(t,\lambda) \int_{T\times S} |\Phi(y,s,\lambda,t)|^2 d(\omega\times\kappa)(t,\lambda)\Big) d(\mu\times\mu)(x,y)$$
$$< \infty$$

This allows us to use Fubini-Tonelli on $(\mathcal{X}\times\mathcal{X})\times(T\times S)$ so that we can interchange the integration order. Moreover it implies that $\omega\times\kappa$-almost everywhere.

$$\int_{\mathcal{X}\times\mathcal{X}} |\Phi(x,r,\lambda,t)\overline{\Phi(y,s,\lambda,t)}|\, d(\mu\times\mu)(x,y) < \infty$$

which implies using Fubini-Tonelli on $\mathcal{X}\times\mathcal{X}$

$$\int_{\mathcal{X}} |\Phi(x,r,\lambda,t)|\, d\mu(x) < \infty$$

□

This representation has several advantages. First, the functions $\Gamma(x,t)$ give us a better idea what properties of the measure P are used in the structural kernel. Second, in the case where $S\times T$ is of the same or smaller size than $\mathcal{X}$ we can decrease the computation cost, since we now have to do only an integration over $T\times S$ instead of an integration over $\mathcal{X}\times\mathcal{X}$. Finally, this representation is a good starting point if one wants to approximate the structural kernel. Since any discretization of T, S, or $\mathcal{X}$ or integration over smaller subsets, will nevertheless give a pd kernel in the end. We illustrate this result with a simple example. We take $\mathcal{X}=\mathbb{R}^n$ and $k(x,y)=k(x-y)$ to be a translation invariant kernel, furthermore we take $\hat{k}(p(x),q(y))=p(x)q(y)$. The characterization of translation invariant kernels is a classical result due to Bochner:

Theorem 5.16 *A continuous function $k(x,y)=k(x-y)$ is pd on $\mathbb{R}^n$ if and only if $k(x-y)=\int_{\mathbb{R}^n} e^{i\langle t,x-y\rangle}d\omega(t)$ where ω is a finite non-negative measure on $\mathbb{R}^n$.*

Obviously we have in this case $T=\mathbb{R}^n$. Then the above proposition tells us that we are effectively computing the following feature vector for each P, $\phi_P(t)=\int_{\mathbb{R}^n} e^{i\langle x,t\rangle}p(x)d\mu(x)=\mathbb{E}_P\, e^{i\langle x,t\rangle}$. Finally the structural kernel can in this case be equivalently written as $K_I(P,Q)=\int_{\mathbb{R}^n}\mathbb{E}_P\, e^{i\langle x,t\rangle}\mathbb{E}_Q\, e^{i\langle x,t\rangle}d\omega(t)$. That means the kernel is in this case nothing else than the inner product between the characteristic functions of the measures in $L_2(\mathbb{R}^n,\omega)$[10]. Moreover the computational cost has decreased since we only have to integrate over $T=\mathbb{R}^n$ instead of $\mathbb{R}^n\times\mathbb{R}^n$. Therefore in this case the kernel computation has the same computational complexity as in the case of the covariant kernels. The calculation of the features, here the characteristic functions, can be done as a preprocessing step for each measure.

5.5.2 Structural Kernel II

The second structural kernel we propose has almost the opposite properties compared to the first one. It is invariant with respect to the dominating measure and therefore defined on the set of all probability measures $\mathcal{M}_+^1(\mathcal{X})$. On the other hand it can also incorporate a similarity function on $\mathcal{X}$, but the distance between disjoint measures will not correspond to their 'closeness' in $\mathcal{X}$.

Theorem 5.17 (Structural Kernel II) *Let $s:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be a bounded non-negative function, $\hat{k}$ a one-homogeneous pd kernel on $\mathbb{R}_+$, and μ a dominating measure of P and Q. Then*

$$K_{II}(P,Q)=\int_{\mathcal{X}\times\mathcal{X}} s(x,y)\,\hat{k}(p(x),q(x))\,\hat{k}(p(y),q(y))\,d\mu(x)\,d\mu(y), \tag{5.11}$$

[10] Note that ω is not the Lebesgue measure.

is a pd kernel on $\mathcal{M}_+^1(\mathcal{X})$. K_{II} is independent of the dominating measure. Moreover $K_{II}(P,Q) \geq 0$, $\forall\, P,Q \in \mathcal{M}_+^1(\mathcal{X})$ if $s(x,y)$ is a bounded positive definite kernel.

Proof: We first prove that K_{II} is positive definite on $\mathcal{M}_+^1(\mathcal{X})$. Note that

$$\sum_{i,j=1}^{n} c_i c_j K_{II}(P_i,P_j) = \int_{\mathcal{X}^2} s(x,y) \sum_{i,j=1}^{n} c_i c_j \hat{k}(p_i(x),p_j(x))\hat{k}(p_i(y),p_j(y))d\mu(x)d\mu(y)$$

The second term is a non-negative function in x and y since $\hat{k}^2$ is positive definite on $(\mathbb{R}_+ \times \mathbb{R}_+) \times (\mathbb{R}_+ \times \mathbb{R}_+)$. Now since $s(x,y)$ is a non-negative function, the integration over $\mathcal{X} \times \mathcal{X}$ is positive. The independence of $K_{II}(P,Q)$ of the dominating measure follows directly from the one-homogeneity of $\hat{k}(x,y)$. Define now $f(x) = \hat{k}(p(x),q(x))$. Then $f \in L_1(\mathcal{X},\mu)$ since

$$\begin{aligned}\int_{\mathcal{X}} |f(x)|d\mu(x) &\leq \int_{\mathcal{X}} \sqrt{\hat{k}(p(x),p(x))\hat{k}(q(x),q(x))}d\mu(x) \\ &= \kappa(\mathbb{R})^2 \int_{\mathcal{X}} \sqrt{p(x)q(x)}d\mu(x) \leq \kappa(\mathbb{R})^2,\end{aligned}$$

where we have used the representation of one-homogeneous kernels. A bounded positive definite kernel $s(x,y)$ defines a positive definite integral operator

$$I : L_1(\mathcal{X},\mu) \to L_\infty(\mathcal{X},\mu), \quad (Ig)(x) = \int_{\mathcal{X}} s(x,y)g(y)d\mu(y).$$

With the definition of $f(x)$ as above, K_{II} is positive since

$$K_{II}(P,Q) = \int_{\mathcal{X}} \int_{\mathcal{X}} s(x,y)f(x)f(y)\, d\mu(x)\, \mu(y) \geq 0.$$

□

Even if the kernel looks quite similar to the first one, it cannot be decomposed as the first one since $s(x,y)$ need not be a positive definite kernel. We just give the equivalent representation without proof:

Proposition 5.18 *If $s(x,y)$ is a positive definite kernel on $\mathcal{X}$ then $K_{II}(P,Q)$ can be equivalently written as:*

$$K_{II}(P,Q) = \int_T \left| \int_{\mathcal{X}} \Gamma(x,t)\hat{k}(p(x),q(x))d\mu(x) \right|^2 d\omega(t),$$

where $s(x,y) = \int_T \Gamma(x,t)\overline{\Gamma(x,t)}d\omega(t)$.

We illustrate this representation with a simple example. Let $s(x,y)$ be a translation-invariant kernel on $\mathbb{R}^n$. Then we can again use Bochner's theorem for the representation of $s(x,y)$. Also assume that P and Q are dominated by the Lebesgue measure. Then note that $f(t) = \int_{\mathcal{X}} \Gamma(x,t)\hat{k}(p(x),q(x))d\mu(x)$ is the Fourier transform of $\hat{k}(p(x),q(x))$ so that in this case the kernel $K_{II}(P,Q)$ is nothing else than the integrated power spectrum of the function $\hat{k}(p(x),q(x))$ with respect to ω.

5.6 Experiments

We compared the performance of the proposed metrics/kernels in four classification tasks. All used data sets consist of inherently positive data resp. counts of terms, counts of pixels of a given color, intensity at a given pixel. Also we will never encounter an infinite number of counts in practice, so that the assumption that the data consists of bounded, positive measures seems reasonable. Moreover we normalize always so that we get probability measures. For text data this is one of the standard representations, also for the Corel data this is quite natural, since all images have the same size and therefore the same number of pixels. This in turn implies that all images have the same mass in color space. For the USPS dataset it might seem at first a little bit odd to see digits as probability measures. Still the results we get are comparable to that of standard kernels without normalization, see [81]. Nevertheless we don't get state-of-the-art results for USPS since we don't implement invariance of the digits with respect to translations and small rotations.
Details of the datasets and used similarity measures:

- *Reuters* text data set. The documents are represented as term histograms. Following [57] we used the five most frequent classes *earn, acq, moneyFx, grain* and *crude*. Documents which belong to more than one of theses classes are excluded. This results in a data set with 8085 examples of dimension 18635.

- *WebKB* web pages data set. The documents are also represented as term histograms. The four most frequent classes *student, faculty, course* and *project* are used. 4198 documents remain each of dimension 24212, see [57]. For both structural kernels we took for both text data sets the correlation matrix in the bag of documents representation as a pd kernel on the space of terms.

- *Corel* image data base. We chose the categories Corel14 from the Corel image database as in [21]. The Corel14 has 14 classes each with 100 examples. As reported in [21] the classes are very noisy, especially the bear and polar bear classes. We performed a uniform quantization of each image in the RGB color space, using 16 bins per color, yielding 4096 dimensional histograms. For both structural kernels we used as a similarity measure on the RGB color space the compactly supported positive definite RBF kernel, $k(x,y) = (1 - \|x-y\| / d_{max})_+^2$, with $d_{max} = 0.15$, see [105].

- *USPS* data set. 7291 training and 2007 test samples. For the first structural kernel we used again the compactly supported RBF kernel with $d_{max} = 2.2$ where we take the euclidean distance on the pixel space such that the smallest distance between two pixels is 1. For the second structural kernel we used as the similarity function $s(x,y) = 1_{\|x-y\| \leq 2.2}$.

All data sets were split into a training (80%) and a test (20%) set. The multiclass problem was solved by one-vs-all with SVM's. For all experiments we used the one-parameter family $d^2_{\alpha|1}$ of Hilbertian metrics resp. their positive definite kernel counterparts $k_{\alpha|1}$ as basic metrics resp. kernels on $\mathbb{R}_+$ in order to build the covariant Hilbertian metrics and both structural kernels. In the table they are denoted as *dir*. Then a second run was done by plugging the metric $D_{\alpha|1}(P,Q)$ on $\mathcal{M}_+^1(\mathcal{X})$ induced

by the covariant resp. structural kernels into a Gaussian[11]:

$$K_{\alpha|1,\lambda}(P,Q) = e^{-D^2_{\alpha,1}(P,Q)/\lambda} \tag{5.12}$$

They are denoted in the table as *exp*. As a comparison we show the results if one takes the linear kernel on $\mathbb{R}_+$, $k(x,y) = xy$, as a basis kernel. Note that this kernel is 2-homogeneous compared to the 1-homogeneous kernels $k_{\alpha|1}$. Therefore the linear kernel will not yield a covariant kernel. As mentioned earlier the first structural kernel becomes independent of the dominating measure with this choice of $\hat{k}$. Also in this case we plugged the resulting metric on $\mathcal{M}^1_+(\mathcal{X})$ into a Gaussian for a second series of experiments. In the simplest case this gives the Gaussian kernel $k(x,y) = \exp(-\|x-y\|^2/\lambda)$.
For the penalty constant we chose from $C = \{10^k, k = -1, 0, 1, 2, 3, 4\}$ and for α from $\alpha = \{1/2, \pm 1, \pm 2, \pm 4, \pm 16, \infty\}$ ($\alpha = -\infty$ coincides with $\alpha = \infty$). For the Gaussian (5.12) we chose additionally from $\lambda = 0.2 * \sigma * \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13\}$, where $\sigma = \frac{1}{n}\sum_{m=1}^{n} K(P_m, P_m)$. In order to find the best parameters for C, α resp. C, α, λ we performed 10-folds cross validation. For the best parameters among α, C resp. α, C, λ we evaluated the test error. Since the Hilbertian metrics of (5.8) were not yet compared or even used in kernel methods, we also give the test errors for the kernels corresponding to $\alpha = -1, 1/2, 1, \infty$. The results are shown in table 5.1.

5.6.1 Interpretation

- The test error for the best α among the family $k_{\alpha|1}$ selected by cross-validation gives for all three types of kernels and their Gaussian transform always optimal or close to optimal results.
- For the text classification the covariant kernels were always better than the structured ones. We think that by using a better similarity measure on terms the structural kernels should improve. For the two image classification tasks the test errors of the best structural kernel is roughly 10% better than the best covariant one.
- The linear resp. Gaussian kernel were for the first three data-sets always worse than the corresponding covariant ones. This remains valid even if one only compares the direct covariant ones with the Gaussian kernel (so that one has in both cases only a one-parameter family of kernels). For the USPS dataset the results are comparable. Future experiments have to show whether this remains true if one considers unnormalized data.

5.7 Conclusion

We extended a family of Hilbertian metrics proposed by Topsøe so that now all previously used measures on probabilities are now included in this family. Moreover we studied with our structural kernels two ways of incorporating similarity information in the space $\mathcal{X}$ into the kernel on probability measures. We gave an equivalent representation for our first structural kernel on $\mathcal{M}^1_+(\mathcal{X})$ which on the one hand provides a better understanding how it captures structure of the probability measures and on the other hand gives in some cases a more efficient way to compute it.

[11] It is well-known that this transform yields a positive definite kernel iff D is a Hilbertian metric, see e.g. [14].

Tabelle 5.1: The table shows the test errors for the covariant and the two structural kernels resp. of their Gaussian transform for each data set. The first column shows the test error and the α-value of the kernel with the best cross-validation error over the family $D^2_{\alpha|1}$ denoted as *dir* resp. of the Gaussian transform denoted as *exp*. The next four columns provide the results for the special cases $\alpha = -1, 1/2, 1, \infty$ in $D^2_{\alpha|1}$ resp. $K_{\alpha|1,\lambda}$. The last column $\langle \cdot, \cdot \rangle$ gives the test error if one takes the linear kernel as basis kernel resp. of the Gaussian transform.

			Best α		$\alpha = -1$	$\alpha = \frac{1}{2}$	$\alpha = 1$	$\alpha = \infty$	$\langle \cdot, \cdot \rangle$
Reuters	cov	dir	**1.36**	-1	**1.36**	1.42	**1.36**	1.79	1.98
	cov	exp	*1.54*	1/2	1.73	1.54	1.79	1.91	1.73
	str	dir	1.85	1	*1.60*	1.91	1.85	1.67	2.16
	str	exp	*1.54*	1	1.60	*1.54*	*1.54*	1.60	2.10
	str2	dir	*1.54*	1	1.85	1.67	*1.54*	2.35	2.41
	str2	exp	*1.67*	1/2	2.04	1.67	1.91	2.53	2.65
WebKB	cov	dir	4.88	16	4.76	4.88	*4.52*	4.64	7.49
	cov	exp	4.76	1	4.76	**4.40**	4.76	4.99	7.25
	str	dir	*4.88*	∞	5.47	5.95	5.23	*4.88*	6.30
	str	exp	*5.11*	∞	5.35	5.23	*5.11*	*5.11*	6.42
	str2	dir	*4.88*	1/2	5.59	*4.88*	5.59	6.30	9.39
	str2	exp	*5.59*	1/2	6.18	5.59	5.95	7.13	9.04
Corel14	cov	dir	12.86	-1	12.86	20.71	15.71	*12.50*	30.00
	cov	exp	12.50	1	*11.43*	14.29	12.50	11.79	34.64
	str	dir	15.71	-1	15.71	23.21	16.43	*12.14*	29.64
	str	exp	**10.36**	1	10.71	12.50	**10.36**	11.07	20.36
	str2	dir	20.00	16	*18.57*	21.43	19.29	20.00	36.79
	str2	exp	*17.14*	1/2	18.57	*17.14*	19.29	18.93	35.71
USPS	cov	dir	*7.82*	-2	8.07	7.92	8.17	7.87	9.02
	cov	exp	*4.53*	-16	4.58	4.58	*4.53*	5.28	*4.53*
	str	dir	*7.52*	-1	*7.52*	8.87	7.77	7.87	9.07
	str	exp	4.04	1/2	3.99	4.04	**3.94**	4.78	4.09
	str2	dir	5.48	2	5.18	5.28	5.33	6.03	*5.03*
	str2	exp	4.29	1/2	*4.09*	4.29	4.24	5.03	4.88

Further we proposed a second structural kernel which is independent of the dominating measure, therefore yielding a structural kernel on all probability measures. Finally we could show that doing model selection in $d^2_{\alpha|1}$ resp. $k_{\alpha|1}$ gives almost optimal results for covariant and structural kernels. Also the covariant kernels and their Gaussian transform are almost always superior to the linear resp. the Gaussian kernel which suggests that the considered family of kernels is a serious alternative whenever one has data which is generically positive. It remains an open problem if one can improve the structural kernels for text classification by using a better similarity function/kernel.

Notation

General:

$\mathbb{R}_+$	Set of positive real numbers including zero
$\mathbb{R}_+^*$	Set of strictly positive real numbers
k	Kernel function (not necessarily positive definite)
$(\mathcal{X}, d)$	metric space, set $\mathcal{X}$ with metric d
B'	dual space to Banach space B

Chapter II:

V	Set of vertices
E	Set of edges
$\langle \cdot, \cdot \rangle_V$	Inner product on the functions on the vertices
$\langle \cdot, \cdot \rangle_E$	Inner product on the functions on the edges
χ	Weight function in $\langle \cdot, \cdot \rangle_V$
ϕ	Weight function in $\langle \cdot, \cdot \rangle_E$
γ	Weight function in the difference operator d
Δ_{norm}	Normalized graph Laplacian
Δ_{unnorm}	Unnormalized graph Laplacian
M	Submanifold
∂M	Boundary of the submanifold M
g, g_{ij}	Metric tensor and its coordinate representation
$\mathrm{II}, \mathrm{II}^k{}_{ij}$	Second fundamental form and its coordinate representation
R_{ijkl}	Riemannian curvature tensor and its coordinate representation
R	Scalar curvature
$\nabla_U V$	Covariant derivative of V in the direction of U
$D_t V$	Covariant derivative of V along a curve
$\overline{\nabla}, \overline{\mathrm{II}}$	Cov. derivative and second fundamental form of the boundary ∂M
$dV(x)$	natural volume element of M
$\exp_p$	Exponential map at p
$\mathrm{inj}(p)$	Injectivity radius at p
R_0	Largest ball in M for which one has global bounds on the volume
i	Isometric embedding of M into $\mathbb{R}^d$
ρ	Radius of curvature
κ	measures the global self-'nearness' of M with respect to $\mathbb{R}^d$
$\delta(x)$	more local version of κ
C_1, C_2, R_k, r_k	Constants associated to the kernel function k
$P, p(x)$	Probability measure and its density with respect to $dV(x)$
$C^k(M)$	Space of k-times continuously differentiable functions on M

Chapter III:

$\mathbb{R}^{[\mathcal{X}]}$	Vector Space of finite linear comb. of evaluation functionals
$\mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$	Set of positive definite kernels on $\mathcal{X}$
$L_+(\mathbb{R}^{\mathcal{X}})$	Set of positive, symmetric kernel operators
$\mathrm{Hilb}(\mathbb{R}^{\mathcal{X}})$	Set of Hilbertian subspaces of $\mathbb{R}^{\mathcal{X}}$

Chapter IV:

$M^\perp$	Subspace of the dual space which annihilates the subspace M
${}^\perp N$	Subspace which annihilates the subspace N of the dual space
$C_b(\mathcal{X})$	Banach space of cont., bounded funct. on $\mathcal{X}$ with the $\|\cdot\|_\infty$-norm
$\widehat{R}_n(\mathcal{F})$	empirical Rademacher average of the function class $\mathcal{F}$
$N(\varepsilon, \mathcal{F}, d)$	Covering numbers of the set $\mathcal{F}$ at scale ϵ with respect to d

Chapter V:

$\mathcal{M}_+^1(\mathcal{X})$	Set of probability measures on $\mathcal{X}$
$\mathcal{M}_+^b(\mathcal{X})$	Set of bounded positive measures on $\mathcal{X}$

Literaturverzeichnis

[1] I. Aharoni, B. Maurey, and B. S. Mityagin. Uniform embeddings of metric spaces and of Banach spaces into Hilbert spaces. *Israel J. Math.*, 52:251–265, 1985.

[2] R. Alexander and S. Alexander. Geodesics in Riemannian manifolds with boundary. *Indiana Univ. Math. J.*, 30:481–488, 1981.

[3] S. Amari and H. Nagaoka. *Information Geometry*. AMS, Providence, RI, 2000.

[4] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[5] M. Atteia. *Hilbertian Kernels and Spline Functions*. North Holland, Amsterdam, 1992.

[6] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2002.

[7] P. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Ann. Stat.*, 33:1497–1537, 2005.

[8] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comp.*, 15(6):1373–1396, 2003.

[9] M. Belkin and P. Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine Learning*, 56:209–239, 2004.

[10] M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. In P. Auer and R. Meir, editors, *Proc. of the 18th Conf. on Learning Theory (COLT)*, Berlin, 2005. Springer.

[11] M. Belkin. *Problems of Learning on Manifolds*. PhD thesis, University of Chicago, 2003. http://www.people.cs.uchicago.edu /˜misha /thesis.pdf.

[12] K. P. Bennett and E. J. Bredensteiner. Duality and geometry in SVM classifiers. In *Proc. of the 17th Int. Conf. on Machine Learning (ICML)*, pages 57–64, 2000.

[13] P. H. Bérard. *Spectral Geometry: Direct and Inverse Problems*. Springer, Berlin, 1986.

[14] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer, New York, 1984.

[15] M. Bernstein, V. de Silva, J. C. Langford, and J.B. Tenenbaum. Graph approximations to geodesics on embedded manifolds. Technical report, Stanford University, 2001.

[16] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In O. Bousquet, U. von Luxburg, and G. Rätsch, editors, *Advanced Lectures on Machine Learning*, pages 169–207, Berlin, 2004. Springer.

[17] O. Bousquet, S. Boucheron, and G. Lugosi. Theory of classification: A survey of recent advances. *ESAIM: Probability and Statistics*. in press.

[18] O. Bousquet, O. Chapelle, and M. Hein. Measure based regularization. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Adv. in Neur. Inf. Proc. Syst. (NIPS)*, volume 16. MIT Press, 2004.

[19] C. J. C. Burges and D. J. Crisp. Uniqueness of the SVM solution. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Adv. in Neur. Inf. Proc. Syst. (NIPS)*, volume 12, 1999.

[20] S. Canu and A. Elisseeff. Regularization, kernels and sigmoid net. unpublished, 1999.

[21] O. Chapelle, P. Haffner, and V. Vapnik. SVMs for histogram-based image classification. *IEEE Trans. on Neural Netw.*, 10:1055–1064, 1999.

[22] I. Chavel. *Riemannian geometry - a modern introduction*. Cambridge University Press, Cambridge, 1992.

[23] Y. C. Choquet-Bruhat, C. DeWitt-Morette, and M. Dillard-Bleick. *Analysis, Manifolds and Physics*. North-Holland Publishing Co., Amsterdam, second edition, 1982.

[24] F. Chung. *Spectral Graph Theory*. AMS, Providence, RI, 1997.

[25] S. Coifman and S. Lafon. Diffusion maps. Preprint, Jan. 2005, to appear in Appl. and Comp. Harm. Anal., 2005.

[26] J. B. Conway. *A course in Functional Analysis*. Springer, New York, 1985.

[27] J. A. Costa and A. O. Hero. Learning intrinsic dimension and intrinsic entropy of high-dimensional datasets. In *Proc. of European Sig. Proc. Conference (EUSIPCO)*, 2004.

[28] A. Devinatz. On measurable positive definite operator functions. *J. London Math. Soc.*, 35:417–424, 1960.

[29] M. Deza and M. Laurent. *Geometry of Cuts and Metrics*. Springer, New York, 1997.

[30] R. M. Dudley. Universal Donsker classes and metric entropy. *Ann. Prob.*, 15:1306–1326, 1987.

[31] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Trans. Inf. Theory*, 49:1858–1860, 2003.

[32] K. Falconer. *Fractal Geometry*. Wiley, Hoboken, NJ, 2nd edition, 2003.

[33] B. Fuglede and F. Topsøe. Jensen-Shannon divergence and Hilbert space embedding. In *Proc. of IEEE Symp. on Inf. Theory*, 2004.

[34] B. Fuglede. Spirals in Hilbert space: with an application in information theory. *Exp. Mathematicae*, 23:23–45, 2005.

[35] K. Fukunaga. An algorithm for finding intrinsic dimensionality of data. *IEEE Trans. on Computers*, 20:176–183, 1971.

[36] E. Giné and J. Zinn. Marcinkiewicz type laws of large numbers and convergence of moments for U-statistics. In *Probability in Banach spaces, 8, Brunswick, 1991*, pages 273–291. Birkhäuser Boston, 1992.

[37] T. Graepel, R. Herbrich, B. Schölkopf, A. Smola, P. Bartlett, K.R. Müller, K. Obermayer, and R. Williamson. Classification on proximity data with LP-machines. In *International Conference on Artificial Neural Networks*, pages 304–309, 1999.

[38] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D*, 9:189–208, 1983.

[39] W. Greblicki, A. Krzyzak, and M. Pawlak. Distribution-free pointwise consistency of kernel regression estimate. *Ann. Stat.*, 12:1570–1575, 1984.

[40] M. Hein, J.-Y. Audibert, and U. von Luxburg. From graphs to manifolds - weak and strong pointwise consistency of graph Laplacians. In P. Auer and R. Meir, editors, *Proc. of the 18th Conf. on Learning Theory (COLT)*, Berlin, 2005. Springer.

[41] M. Hein and J.-Y. Audibert. Intrinsic dimensionality estimation of submanifolds in $\mathbb{R}^d$. In L. De Raedt and S. Wrobel, editors, *Proc. of the 22nd Int. Conf. on Machine Learning (ICML)*, 2005.

[42] M. Hein, O. Bousquet, and B. Schölkopf. Maximal margin classification for metric spaces. *J. Comput. System Sci.*, 71:333–359, 2005.

[43] M. Hein and O. Bousquet. Maximal margin classification for metric spaces. In B. Schölkopf and M. Warmuth, editors, *Proc. of the 16th Ann. Conf. on Learning Theory (COLT)*, New-York, 2003. Springer.

[44] M. Hein and O. Bousquet. Kernels, associated structures and generalizations. Technical Report TR-127, Max Planck Institute for Biological Cybernetics, 2004.

[45] M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In Z. Ghahramani and R. Cowell, editors, *Proc. of the 10th Int. Workshop on Art. Int. and Stat. (AISTATS)*, 2005.

[46] M. Hein, T. N. Lal, and O. Bousquet. Hilbertian metrics on probability measures and their application in SVM's. In C. E. Rasmussen, H. H. Bülthoff, M. Giese, and B. Schölkopf, editors, *26th Pattern Recognition Symposium (DAGM)*, Berlin, 2004. Springer.

[47] H. Hendriks, J.H.M. Janssen, and F.H. Ruymgaart. Strong uniform convergence of density estimators on compact Euclidean manifolds. *Statist. Prob. Lett.*, 16:305–311, 1993.

[48] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[49] S. Janson. *Gaussian Hilbert Spaces*. Cambridge University Press, Cambridge, 1997.

[50] T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *J. Mach. Learn. Res.*, 5:819–844, 2004.

[51] T. Jebara and R. Kondor. Bhattacharyya and expected likelihood kernels. In B. Schölkopf and M. Warmuth, editors, *Proc. of the 16th Ann. Conf. on Learning Theory (COLT)*. Springer, 2003.

[52] J. Jost. *Riemannian Geometry and Geometric Analysis*. Springer-Verlag, Berlin, third edition, 2002.

[53] B. Kegl. Intrinsic dimension estimation using packing numbers. In S. Thrun, S. Becker, and K. Obermayer, editors, *Adv. in Neur. Inf. Proc. Syst. (NIPS)*, volume 15. MIT Press, 2003.

[54] W. Klingenberg. *Riemannian Geometry*. De Gruyter, Berlin, 1982.

[55] R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proc. of the 19th Int. Conf. on Machine Learning (ICML)*, 2002.

[56] M. Krein. Hermitian-positive kernels on homogeneous spaces I and II. *Amer. Math. Soc. Translations Ser. 2*, 34:69–164, 1963. Original: Ukrain. Mat. Z., 1, 64-98,(1949), and 2, 10-59,(1950).

[57] J. Lafferty and G. Lebanon. Diffusion kernels on statistical manifolds. *J. Mach. Learn. Res.*, 6:129–163, 2005.

[58] S. S. Lafon. *Diffusion Maps and Geometric Harmonics*. PhD thesis, Yale University, 2004. http://www.math.yale.edu /˜sl349 /publications /dissertation.pdf.

[59] R. Latala and K. Oleszkiewicz. On the best constant in the Khintchine-Kahane inequality. *Studia Math.*, 109:101–104, 1994.

[60] J. M. Lee. *Riemannian Manifolds*. Springer, New York, 1997.

[61] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory*, 37:145–151, 1991.

[62] M. N. Lukić and J. H. Beder. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Trans. Am. Math. Soc.*, 353:3945–3969, 2001.

[63] C. McDiarmid. On the method of bounded differences. In *Surveys in combinatorics, (Norwich, 1989)*, pages 148–188. Cambridge University Press, Cambridge, 1989.

[64] P. McDonald and R. Meyers. Diffusions on graphs, poisson problems and spectral geometry. *Trans. Amer. Math. Soc.*, 354:5111–5136, 2002.

[65] S. Mendelson. Geometric parameters in learning theory. In V. D. Milman and G. Schechtman, editors, *Geometric aspects of functional analysis*, pages 193–235, Berlin, 2004. Springer.

[66] C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17:177–204, 2005.

[67] P. J. Moreno, P. P. Hu, and N. Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In S. Thrun, S. Becker, and K. Obermayer, editors, *Adv. in Neur. Inf. Proc. Syst. (NIPS)*, volume 16, Cambridge, MA, 2003. MIT Press.

[68] H. Niemi. Stochastic processes as Fourier transforms of stochastic measures. *Ann. Acad. Sci. Fenn. Ser. A I*, 591, 1975.

[69] D. Nolan and D. Pollard. U-processes, rates of convergence. *Ann. Stat.*, 15:780–799, 1987.

[70] K. R. Parthasarathy and K. Schmidt. *Positive definite kernels, continuous tensor products, and central limit theorems of probability theory*, volume 272 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1972.

[71] E. Pekalska, P. Paclik, and R.P.W. Duin. A generalized kernel approach to dissimilarity-based classification. *J. Mach. Learn. Res.*, 2:175–211, 2001.

[72] B. Pelletier. Kernel density estimation on Riemannian manifolds. *Statist. Prob. Lett.*, 73:297–304, 2005.

[73] M. Reed and B. Simon. *Methods of modern mathematical physics. Vol. 1: Functional Analysis*. Academic Press, San Diego, 1980.

[74] S. Rosenberg. *The Laplacian on a Riemannian manifold*. Cambridge University Press, 1997.

[75] W. Rudin. *Functional Analysis*. McGraw Hill, 1991.

[76] H. H. Schaefer and M. P. Wolff. *Topological Vector Spaces*. Springer, New York, 1999. Second edition.

[77] T. Schick. *Analysis on ∂-manifolds of bounded geometry, Hodge-De Rham Isomorphsim and L^2-Index theorem*. PhD thesis, Universität Mainz, 1996.

[78] T. Schick. Manifolds with boundary of bounded geometry. *Math. Nachr.*, 223:103–120, 2001.

[79] I. J. Schoenberg. Metric spaces and completely monotone functions. *Ann. Math.*, 39:811–841, 1938.

[80] I. J. Schoenberg. Metric spaces and positive definite functions. *Trans. Amer. Math. Soc.*, 44:522–536, 1938.

[81] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[82] B. Schölkopf. The kernel trick for distances. In T. G. Dietterich, T. K. Leen, and V. Tresp, editors, *Adv. in Neur. Inf. Proc. Syst. (NIPS)*, volume 13, Cambridge, MA, 2000. MIT Press.

[83] L. Schwartz. Sous-espaces hilbertiens et noyaux associés. *Journal d'Analyse, Jerusalem*, XIII:115–256, 1964.

[84] R. Serfling. *Approximation Theorems in Mathematical Statistics*. Wiley, New York, 1980.

[85] A.J. Smola and R. Kondor. Kernels and regularization on graphs. In B. Schölkopf and M. Warmuth, editors, *Proc. of the 16th Ann. Conf. on Learning Theory (COLT)*, Lecture Notes in Computer Science. Springer, 2003.

[86] O. G. Smolyanov, H. von Weizsäcker, and O. Wittich. Brownian motion on a manifold as limit of stepwise conditioned standard Brownian motions. In *Stochastic processes, physics and geometry: new interplays, II*, volume 29, pages 589–602, Providence, RI, 2000. AMS.

[87] O. G. Smolyanov, H. von Weizsäcker, and O. Wittich. Chernoff's theorem and discrete time approximations of Brownian motion on manifolds. Preprint, available at http://lanl.arxiv.org/abs/math.PR/0409155, 2004.

[88] P. Sorjonen. Pontrjaginräume mit einem reproduzierenden Kern. *Ann. Acad. Sci. Fenn. Ser. A I*, 594, 1975.

[89] I. Steinwart. Support vector machines are universally consistent. *J. Complexity*, 18:768–791, 2002.

[90] I. Steinwart. Consistency of support vector machines and other regularized kernel machines. *IEEE Trans. Inf. Theory*, 51:128–142, 2005.

[91] C. Suquet. Une topologie prehilbertienne sur l'espace des mesures à signes bornées. *Publ. Inst. Statist. Univ. Paris*, 35:51–77, 1990.

[92] C. Suquet. Convergences stochastiques de suites de mesures alatoires signées considérées comme variables aléatoires hilbertiennes. *Publ. Inst. Statist. Univ. Paris*, 37:71–99, 1993.

[93] C. Suquet. Distances euclidiennes sur les mesures signées et application à des théorèmes de Berry-Esséen. *Bull. Belg. Math. Soc. Simon Stevin*, 2:161–181, 1995.

[94] F. Takens. On the numerical determination of the dimension of an attractor. In *Dynamical systems and bifurcations*, volume 58, pages 99–106, 1985.

[95] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[96] J. Theiler. Estimating fractal dimension. *J. Opt. Soc. Am. A*, 7:1055–1073, 1990.

[97] F. Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Inform. Th.*, 46:1602–1609, 2000.

[98] F. Topsøe. Jenson-Shannon divergence and norm-based measures of discrimination and variation. Preprint, 2003.

[99] V. Vapnik. *The nature of statistical learning theory*. Spinger, New York, second edition, 2000.

[100] N. T. Varopoulos. Brownian motion and random walks on manifolds. *Ann. Inst. Fourier*, 34:243–269, 1984.

[101] C. Villani. *Topics in Optimal Transportation*. AMS, Providence, RI, 2003.

[102] U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. Technical Report TR-134, Max Planck Institute for Biological Cybernetics, 2004.

[103] U. von Luxburg and O. Bousquet. Distance-based classification with Lipschitz functions. *J. Mach. Learn. Res.*, 5:669–695, 2004.

[104] U. von Luxburg. *Statistical Learning with Similarity and Dissimilarity functions*. PhD thesis, Max Planck Institute for Biological Cybernetics/Fakultät für Elektrotechnik und Informatik, Technische Universität Berlin, 2004.

[105] H. Wendland. Piecewise polynomial, positive definite and compactly supported radial basis functions of minimal degree. *Adv. Comp. Math.*, 4:389–396, 1995.

[106] W. Woess. *Random Walks on Infinite Graphs and Groups*. Cambridge University Press, Cambridge, 2000.

[107] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Adv. in Neur. Inf. Proc. Syst. (NIPS)*, volume 16. MIT Press, 2004.

[108] D. Zhou, B. Schölkopf, and T. Hofmann. Semi-supervised learning on directed graphs. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Adv. in Neur. Inf. Proc. Syst. (NIPS)*, volume 17. MIT Press, 2005.

[109] D. Zhou, B. Xiao, H. Zhou, and R. Dai. Global geometry of SVM classifiers. Technical Report Technical Report 30-5-02, AI Lab, Institute of Automation, Chinese Academy of Sciences, 2002.

[110] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, CMU CALD, 2002.