Hilbertian Metrics and Positive Definite Kernels on Probability Measures

Matthias Hein and Olivier Bousquet
Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany
{*first.last*}*@tuebingen.mpg.de*

Abstract

We investigate the problem of defining Hilbertian metrics resp. positive definite kernels on probability measures, continuing the work in [5]. This type of kernels has shown very good results in text classification and has a wide range of possible applications. In this paper we extend the two-parameter family of Hilbertian metrics of Topsøe such that it now includes all commonly used Hilbertian metrics on probability measures. This allows us to do model selection among these metrics in an elegant and unified way. Second we investigate further our approach to incorporate similarity information of the probability space into the kernel. The analysis provides a better understanding of these kernels and gives in some cases a more efficient way to compute them. Finally we compare all proposed kernels in two text and two image classification problems.

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. 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 [8] and colors [1].
- 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 [8, 7]. 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 $\mathcal{M}_+^1(\mathcal{X})$[1]. In a first section we will summarize the close connection between Hilbertian metrics and pd kernels so that in general statements for one category can be easily transferred to the other one.
We will consider two types of kernels on probability measures. The first one is general covariant. 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 further investigate types of kernels on probability measures which incorporate such a similarity measure, see [5]. This will yield on the one hand a better understanding of these kernels 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).

[1] $\mathcal{M}_+^1(\mathcal{X})$ denotes the set of positive measures μ on $\mathcal{X}$ with $\mu(\mathcal{X}) = 1$

2 Hilbertian Metrics versus Positive Definite Kernels

It is a well-known fact 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 [11] 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[2].

Definition 2.1 *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 2.1 *[2] 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.*

Similar to pd kernels one can also characterize cpd kernels. Namely one can write all cpd kernels in the form: $k(x,y) = -\frac{1}{2} \left\| \phi_x - \phi_y \right\|_{\mathcal{H}}^2 + f(x) + f(y)$. 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 [2, 3.2] and [11] 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 one can show, see [4], that 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 2.1, one can always define a corresponding pd kernel. Nevertheless for the convenience of the reader we will often explicitly state the corresponding pd kernels.

[2] A (semi)-metric $d(x,y)$ (A semi-metric $d(x,y)$ fulfills the conditions of a metric except that $d(x,y) = 0$ does not imply $x = y$.) 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.

3 γ-homogeneous Hilbertian Metrics and Positive Definite Kernels on $\mathbb{R}_+$[3]

The class of Hilbertian metrics on probability measures we consider in this paper are based on a pointwise comparison of the densities $p(x)$ with 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[4]. The class of γ-homogeneous Hilbertian metrics on $\mathbb{R}_+$ was recently characterized by Fuglede:

Theorem 3.1 (Fuglede [3]) *A symmetric function $d : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ with $d(x,y) = 0 \iff x = y$ is a γ-homogeneous, continuous 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*

$$d^2(x,y) = \int_{\mathbb{R}_+} \left| x^{(\gamma + i\lambda)} - y^{(\gamma + i\lambda)} \right|^2 d\rho(\lambda) \qquad (1)$$

Using Lemma 2.1 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 3.1 *A symmetric function $k : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ with $k(x,x) = 0 \iff x = 0$ is a 2γ-homogeneous continuous 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*

$$k(x,y) = \int_{\mathbb{R}} x^{(\gamma + i\lambda)} y^{(\gamma - i\lambda)} \, d\kappa(\lambda) \qquad (2)$$

Proof: If k has the form given in (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 by applying theorem 3.1, where κ is the symmetrized version of ρ around the origin, together with lemma 2.1 and $k(x,y) = \langle \phi_x, \phi_y \rangle = \frac{1}{2} \left(-d^2(x,y) + d^2(x,0) + d^2(y,0) \right)$. □

At first glance Theorem 3.1, 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.

[3] $\mathbb{R}_+$ is the positive part of the real line with 0 included

[4] A symmetric function k is γ-homogeneous if $k(c\,x, c\,y) = c^\gamma k(x,y)$ for all $c \in \mathbb{R}_+$

Recently Topsøe and Fuglede proposed an interesting two-parameter family of Hilbertian metrics on $\mathbb{R}_+$ [13, 3]. 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 3.2 *The function $d : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ defined as:*

$$d^2_{\alpha|\beta}(x,y) = \frac{2^{\frac{1}{\beta}} \left(x^\alpha + y^\alpha\right)^{\frac{1}{\alpha}} - 2^{\frac{1}{\alpha}} \left(x^\beta + y^\beta\right)^{\frac{1}{\beta}}}{2^{\frac{1}{\alpha}} - 2^{\frac{1}{\beta}}} \qquad (3)$$

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:

$$\lim_{\alpha\to\beta} d^2_{\alpha|\beta}(x,y) = \frac{\beta^2 2^{1/\beta}}{\log(2)} \frac{\partial}{\partial\beta} \left(\frac{x^\beta + y^\beta}{2}\right)^{(1/\beta)} =$$

$$\frac{\left(x^\beta + y^\beta\right)^{\frac{1}{\beta}}}{\log(2)} \left[\frac{x^\beta}{x^\beta + y^\beta} \log\left(\frac{2x^\beta}{x^\beta + y^\beta}\right) + \frac{y^\beta}{x^\beta + y^\beta} \log\left(\frac{2y^\beta}{x^\beta + y^\beta}\right)\right]$$

Note that $d^2_{\alpha|\beta} = d^2_{\beta|\alpha}$. We need the following lemmas in the proof:

Lemma 3.1 *[2, 2.10] If $k : \mathcal{X}\times\mathcal{X}$ is cpd and $k(x,x) \leq 0$, $\forall x \in \mathcal{X}$ then $-(-k)^\gamma$ is also cpd for $0 < \gamma \leq 1$.*

Lemma 3.2 *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 [2] 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. [10]. Therefore $\lim_{t\to 0} 1/(t-k) = -1/k$ is positive definite if k is strictly negative. □

We can now prove Theorem 3.2:

Proof: The proof for the symmetry, the limit $\alpha \to \beta$ and the parameter range $1 \leq \alpha \leq \infty$, $1/2 \leq \beta \leq \alpha$ can be found in [3]. We prove that $-d^2_{\alpha|\beta}$ is cpd for $1 \leq \alpha \leq \infty$, $-\infty \leq \beta \leq -1$. 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 $k(x,y) \leq 0$, $\forall\, x, y \in \mathcal{X}$. Therefore by Lemma 3.1, $-(x^\alpha + y^\alpha)^{1/\alpha}$ is cpd for $1 \leq \alpha < \infty$. 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 $1 \leq \beta \leq \infty$ which is cpd as we have shown and strictly negative if we restrict k to $\{x \in \mathbb{R} \,|\, x > 0\} \times \{x \in \mathbb{R} \,|\, x > 0\}$. Then all conditions for lemma 3.2 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 the first part with $(2^{(1/\alpha - 1/\beta)} - 1)^{-1}$ and the second one with $(1 - 2^{(1/\beta - 1/\alpha)})^{-1}$ and adding them gives the result. □

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[5] our definition will at first depend on the choice of the dominating measure. This dependence would restrict the applicability of our approach. For example if we had $\mathcal{X} = \mathbb{R}^n$ and chose 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. 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 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 $(dP/d\mu)(x)$ [6] of P with respect to the dominating measure μ.

Proposition 4.1 *Let P and Q be two probability measures on $\mathcal{X}$, μ an arbitrary dominating measure[7] 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*

$$D^2_{\mathcal{M}_+^1(\mathcal{X})}(P,Q) := \int_{\mathcal{X}} d^2_{\mathbb{R}_+}(p(x), q(x)) d\mu(x)\,, \qquad (4)$$

is a Hilbertian metric on $\mathcal{M}_+^1(\mathcal{X})$. $D_{\mathcal{M}_+^1(\mathcal{X})}$ is independent of the dominating measure μ.

For a proof, see [5]. Note that if we use an arbitrary metric on $\mathbb{R}_+$ in the above proposition, we also get a Hilbertian metric. But this metric would only be defined on the set of measures dominated by a certain measure μ and not on $\mathcal{M}_+^1(\mathcal{X})$. Moreover it would also depend on the choice of the dominating measure μ.

[5] 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.

[6] 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.

[7] Such a dominating measure always exists take e.g. $M = (P+Q)/2$

We can now apply this principle of building covariant Hilbertian metrics on $\mathcal{M}_+^1(\mathcal{X})$ and use the family of 1/2-homogeneous Hilbertian metrics $d^2_{\alpha|\beta}$ on $\mathbb{R}_+$ from the previous section. This yields as special cases the following well-known measures on $\mathcal{M}_+^1(\mathcal{X})$.

$$\begin{aligned}
D^2_{1|-1}(P,Q) &= \int_{\mathcal{X}} \frac{(p(x)-q(x))^2}{p(x)+q(x)} d\mu(x),\\
D^2_{\frac{1}{2}|1}(P,Q) &= \int_{\mathcal{X}} (\sqrt{p(x)}-\sqrt{q(x)})^2 d\mu(x),\\
D^2_{1|1}(P,Q) &= \frac{1}{\log(2)} \int_{\mathcal{X}} p(x) \log\left[\frac{2p(x)}{p(x)+q(x)}\right]\\
&\quad + q(x)\log\left[\frac{2q(x)}{p(x)+q(x)}\right] d\mu(x),\\
D^2_{\infty|1}(P,Q) &= \int_{\mathcal{X}} |p(x)-q(x)| d\mu(x). \qquad (5)
\end{aligned}$$

$D^2_{1|-1}$ 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 [7]. 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. Instead of the work in [9] 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. Note that the family $d^2_{\alpha|\beta}$ is designed in such a way that the maximal distance of $D^2_{\alpha|\beta}$ is $2, \forall\, \alpha, \beta$. For completeness we also give the corresponding pd kernels on $\mathcal{M}_+^1(\mathcal{X})$, where we take in Lemma 2.1 the zero measure as x_0 in $\mathcal{M}_+^1(\mathcal{X})$. This choice seems strange at first since we are dealing with probability measures. But in fact the whole framework presented in this paper can easily be extended to all finite, positive measures on $\mathcal{X}$. For this set the zero measure is a natural choice of the origin.

$$\begin{aligned}
K_{1|-1}(P,Q) &= \int_{\mathcal{X}} \frac{p(x)q(x)}{p(x)+q(x)} d\mu(x),\\
K_{\frac{1}{2}|1}(P,Q) &= \int_{\mathcal{X}} \sqrt{p(x)q(x)} d\mu(x),\\
K_{1|1}(P,Q) &= \frac{-1}{\log(2)} \int_{\mathcal{X}} p(x)\log\left(\frac{p(x)}{p(x)+q(x)}\right)\\
&\quad + q(x)\log\left(\frac{q(x)}{p(x)+q(x)}\right) d\mu(x),\\
K_{\infty|1}(P,Q) &= \int_{\mathcal{X}} \min\{p(x),q(x)\} d\mu(x). \qquad (6)
\end{aligned}$$

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 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 probability space. 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.

5.1 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 "close" or "far" their support is. Obviously if our training set consists only of disjoint measures learning is not possible with covariant metrics. We have proposed in [5] 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 like e.g. for Riemannian manifolds where there exists a natural volume measure. Such a preferred measure is then a natural choice for the dominating measure, so that theoretically 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, discrete 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 5.1 (Structural Kernel I) *Let k be a bounded PD kernel on $\mathcal{X}$ and $\hat{k}$ a bounded PD kernel on $\mathbb{R}_+$. Then*

$$K_I(P,Q) = \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\, \hat{k}(p(x),q(y))\, d\mu(x)\, d\mu(y) \qquad (7)$$

is a pd kernel on $\mathcal{M}_+^1(\mathcal{X},\mu)\times\mathcal{M}_+^1(\mathcal{X},\mu)$.

We refer to [5] for the proof. Note that this kernel can easily be extended to all bounded, signed measures as it is in general true for all metrics resp. kernels in this paper. This structural kernel generalizes previous work done by Suquet, see [12], 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 (7) 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.

Proposition 5.1 *The kernel $K_I(P,Q)$ can be equivalently written as the inner product in $L_2(T\times S,\omega\otimes\kappa)$:*

$$K_I(P,Q) = \int_T\int_S \phi_P(t,\lambda)\overline{\phi_Q(t,\lambda)}\,d\kappa(\lambda)\,d\omega(t)$$

for some sets T,S with the feature map:

$$\phi : \mathcal{M}_+^1(\mathcal{X},\mu) \to L_2(T\times S,\omega\otimes\kappa),$$
$$P \to \phi_P(t,\lambda) = \int_{\mathcal{X}} \Gamma(x,t)\Psi(p(x),s)d\mu(x).$$

where

$$k(x,y) = \int_T \Gamma(x,t)\overline{\Gamma(y,t)}d\omega(t),$$
$$\hat{k}(p(x),q(x)) = \int_S \Psi(p(x),s)\overline{\Psi(p(y),s)}d\kappa(s).$$

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. [6]. For the product of two positive definite kernels we have such a representation on the set $T\times S$. Since for any finite measure space $(\mathcal{Y},\mu)$ one has $L_2(\mathcal{Y},\mu)\subset L_1(\mathcal{Y},\mu)$ we can apply Fubini's theorem and interchange the integration order. The definition of the feature map $\Phi_P(t,\lambda)$ then follows easily. □

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 on $\mathbb{R}^n$ is a classical result due to Bochner:

Theorem 5.2 *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, $\phi_P(t) = \int_{\mathbb{R}^n} e^{i\langle x,t\rangle}p(x)d\mu(x) = E_P e^{i\langle x,t\rangle}$. Finally the structural kernel can in this case be equivalently written as $K_I(P,Q) = \int_{\mathbb{R}^n} E_P e^{i\langle x,t\rangle}E_Q e^{i\langle x,t\rangle}d\omega(t)$. That means the kernel is in this case nothing else than the inner product between the characteristic functions of the measures in $L_2(\mathbb{R}^n,\omega)$[8]. Moreover the computational cost has decreased dramatically, 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.

5.2 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 of disjoint measures will not correspond to their 'closeness' in $\mathcal{X}$.

Theorem 5.3 (Structural Kernel II) *Let $s:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be a non-negative function, $\hat{k}$ a one-homogeneous pd kernel on $\mathbb{R}_+$ and μ a dominating*

[8] Note that ω is not the Lebesgue measure.

measure of P and Q. Then

$$K_{II}(P,Q) = \int_{\mathcal{X}^2} s(x,y)\hat{k}(p(x),q(x))\hat{k}(p(y),q(y))d\mu(x)\,d\mu(y), \tag{8}$$

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 $\hat{k}^2$ positive definite on $(\mathbb{R}_+ \times \mathbb{R}_+) \times (\mathbb{R}_+ \times \mathbb{R}_+)$. Since $s(x,y)$ is also 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 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 $\int_{\mathcal{X}} |f(x)|d\mu(x) \leq \int_{\mathcal{X}} \sqrt{\hat{k}(p(x),p(x))\hat{k}(q(x),q(x))}d\mu(x) = \kappa(\mathbb{R})^2 \int_{\mathcal{X}} \sqrt{p(x)q(x)}d\mu(x) \leq \kappa(\mathbb{R})^2$, where we have used the representation of one-homogeneous kernels. A bounded pd kernel $s(x,y)$ defines a positive definite integral operator $I : L_1(\mathcal{X},\mu) \to L_\infty(\mathcal{X},\mu)$, $(Ig)(x) = \int_{\mathcal{X}} s(x,y)g(y)d\mu(y)$. 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 5.2 *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 $s(x,y)$. The proposition then states that 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.

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 [10]. 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 [8] 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 [8]. 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 [1]. The Corel14 has 14 classes each with 100 examples. As reported in [1] 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 [14].
- *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 multi-class 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[9]:

$$K_{\alpha|1,\lambda}(P,Q) = e^{-D^2_{\alpha,1}(P,Q)/\lambda} \tag{9}$$

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 (9) 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) 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 1.

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.

[9] It is well-known that this transform yields a positive definite kernel iff D is a Hilbertian metric, see e.g. [2].

7 Conclusion

We went on with the work started in [5] on Hilbertian metrics resp. pd kernels on $\mathcal{M}^1_+(\mathcal{X})$. 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 further structural kernels 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. 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.

Acknowledgements

We would like to thank Guy Lebanon for kindly providing us with the WebKB and Reuters data set in preprocessed from. Furthermore we are thankful to Flemming Topsøe and Bent Fuglede for providing us with preprints of their papers [13, 3].

References

[1] O. Chapelle, P. Haffner, and V. Vapnik. SVMs for histogram-based image classification. *IEEE Trans. on Neural Networks*, 10:1055–1064, 1999.

[2] J. P. R. Christensen C. Berg and P. Ressel. *Harmonic Analysis on Semigroups*. Springer, New York, 1984.

[3] B. Fuglede. Spirals in Hilbert space. With an application in information theory. To appear in Expositiones Mathematicae, 2004.

Table 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

[4] M. Hein, O. Bousquet, and B. Schölkopf. Maximal margin classification for metric spaces. *Journal of Computer and System Sciences*, to appear.

[5] M. Hein, T. N. Lal, and O. Bousquet. Hilbertian metrics on probability measures and their application in SVM's. In *26th Pattern Recognition Symposium (DAGM)*. Springer, 2004.

[6] S. Janson. *Gaussian Hilbert Spaces*. Cambridge University Press, Cambridge, 1997.

[7] T. Jebara and R. Kondor. Bhattacharyya and expected likelihood kernels. In *16th Annual Conference on Learning Theory (COLT)*, 2003.

[8] J. Lafferty and G. Lebanon. Diffusion kernels on statistical manifolds. Technical Report CMU-CS-04-101, School of Computer Science, Carnegie Mellon University, Pittsburgh, 2004.

[9] P. J. Moreno, P. P. Hu, and N. Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. *NIPS*, 16, 2003.

[10] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[11] B. Schölkopf. The kernel trick for distances. *NIPS*, 13, 2000.

[12] 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.

[13] F. Topsøe. Jenson-Shannon divergence and norm-based measures of discrimination and variation. Preprint, 2003.

[14] H. Wendland. Piecewise polynomial, positive definite and compactly supported radial basis functions of minimal degree. *Adv. Comp. Math.*, 4:389–396, 1995.