Journal of Machine Learning Research 6 (2005) 1169–1198 Submitted 1/05; Published 7/05

Semigroup Kernels on Measures

Marco Cuturi MARCO.CUTURI@ENSMP.FR
Ecole des Mines de Paris
35 rue Saint Honoré
77305 Fontainebleau, France;
Institute of Statistical Mathematics
4-6-7 Minami-Azabu, Minato-Ku, Tokyo, Japan

Kenji Fukumizu FUKUMIZU@ISM.AC.JP
Institute of Statistical Mathematics
4-6-7 Minami-Azabu, Minato-Ku, Tokyo, Japan

Jean-Philippe Vert JEAN-PHILIPPE.VERT@ENSMP.FR
Ecole des Mines de Paris
35 rue Saint Honoré
77305 Fontainebleau, France

Editor: John Lafferty

Abstract

We present a family of positive definite kernels on measures, characterized by the fact that the value of the kernel between two measures is a function of their sum. These kernels can be used to derive kernels on structured objects, such as images and texts, by representing these objects as sets of components, such as pixels or words, or more generally as measures on the space of components. Several kernels studied in this work make use of common quantities defined on measures such as entropy or generalized variance to detect similarities. Given an a priori kernel on the space of components itself, the approach is further extended by restating the previous results in a more efficient and flexible framework using the "kernel trick". Finally, a constructive approach to such positive definite kernels through an integral representation theorem is proved, before presenting experimental results on a benchmark experiment of handwritten digits classification to illustrate the validity of the approach.

Keywords: kernels on measures, semigroup theory, Jensen divergence, generalized variance, reproducing kernel Hilbert space

1. Introduction

The challenge of performing classification or regression tasks over complex and non vectorial objects is an increasingly important problem in machine learning, motivated by diverse applications such as bioinformatics or multimedia document processing. The kernel method approach to such problems (Schölkopf and Smola, 2002) is grounded on the choice of a proper similarity measure, namely a positive definite (p.d.) kernel defined between pairs of objects of interest, to be used alongside with kernel methods such as support vector machines (Boser et al., 1992). While natural similarities defined through dot-products and related distances are available when the objects lie in a Hilbert space, there is no standard dot-product to compare strings, texts, videos, graphs or other

©2005 Marco Cuturi, Kenji Fukumizu and Jean-Philippe Vert.

structured objects. This situation motivates the proposal of various kernels, either tuned and trained to be efficient on specific applications or useful in more general cases.

One possible approach to kernel design for such complex objects consists in representing them by sets of basic components easier to manipulate, and designing kernels on such sets. Such basic components can typically be subparts of the original complex objects, obtained by exhaustive enumeration or random sampling. For example, a very common way to represent a text for applications such as text classification and information retrieval is to break it into words and consider it as a bag of words, that is, a finite set of weighted terms. Another possibility is to extract all fixed-length blocks of consecutive letters and represent the text by the vector of counts of all blocks (Leslie et al., 2002), or even to add to this representation additional blocks obtained by slight modifications of the blocks present in the text with different weighting schemes (Leslie et al., 2003). Similarly, a grey-level digitalized image can be considered as a finite set of points of $\mathbb{R}^3$ where each point (x,y,I) stands for the intensity I displayed on the pixel (x,y) in that image (Kondor and Jebara, 2003).

Once such a representation is obtained, different strategies have been adopted to design kernels on these descriptions of complex objects. When the set of basic components is finite, this representation amounts to encode a complex object as a finite-dimensional vector of counters, and any kernel for vectors can be then translated to a kernel for complex object through this feature representation (Joachims, 2002, Leslie et al., 2002, 2003). For more general situations, several authors have proposed to handle such weighted lists of points by first fitting a probability distribution to each list, and defining a kernel between the resulting distributions (Lafferty and Lebanon, 2002, Jebara et al., 2004, Kondor and Jebara, 2003, Hein and Bousquet, 2005). Alternatively, Cuturi and Vert (2005) use a parametric family of densities and a Bayesian framework to define a kernel for strings based on the mutual information between their sets of variable-length blocks, using the concept of mutual information kernels (Seeger, 2002). Finally, Wolf and Shashua (2003) recently proposed a formulation rooted in kernel canonical correlation analysis (Bach and Jordan, 2002, Melzer et al., 2001, Akaho, 2001) which makes use of the principal angles between the subspaces generated by the two sets of points to be compared when considered in a feature space.

We explore in this contribution a different direction to kernel design for weighted lists of basic components. Observing that such a list can be conveniently represented by a molecular measure on the set of basic components, that is a weighted sum of Dirac measures, or that the distribution of points might be fitted by a statistical model and result in a density on the same set, we formally focus our attention on the problem of defining a kernel between finite measures on the space of basic components. More precisely, we explore the set of kernels between measures that can be expressed as a function of their sum, that is:

$$k(\mu,\mu') = \varphi(\mu+\mu'). \tag{1}$$

The rationale behind this formulation is that if two measures or sets of points μ and μ' overlap, then it is expected that the sum $\mu+\mu'$ is more concentrated and less scattered than if they do not. As a result, we typically expect φ to quantify the dispersion of its argument, increasing when it is more concentrated. This setting is therefore a broad generalization of the observation by Cuturi and Vert (2005) that a valid kernel for strings, seen as bags of variable-length blocks, is obtained from the compression rate of the *concatenation* of the two strings by a particular compression algorithm.

The set of measures endowed with the addition is an Abelian semigroup, and the kernel (1) is exactly what Berg et al. (1984) call a *semigroup kernel*. The main contribution of this paper is to present several valid positive definite (p.d.) semigroup kernels for molecular measures or

densities. As expected, we prove that several functions φ that quantify the dispersion of measures through their entropy or through their variance matrix result in valid p.d. kernels. Using entropy to compare two measures is not a new idea (Rao, 1987) but it was recently restated within different frameworks (Hein and Bousquet, 2005, Endres and Schindelin, 2003, Fuglede and Topsøe, 2004). We introduce entropy in this paper slightly differently, noting that it is a semigroup negative definite function defined on measures. On the other hand, the use of generalized variance to derive a positive definite kernel between measures as proposed here is new to our knowledge. We further show how such kernels can be applied to molecular measures through regularization operations. In the case of the kernel based on the spectrum of the variance matrix, we show how it can be applied implicitly for molecular measures mapped to a reproducing kernel Hilbert space when a p.d. kernel on the space of basic components is provided, thanks to an application of the "kernel trick".

Besides these examples of practical relevance, we also consider the question of characterizing *all* functions φ that lead to a p.d. kernel through (1). Using the general theory of semigroup kernels we state an integral representation of such kernels and study the semicharacters involved in this representation. This new result provides a constructive characterization of such kernels, which we briefly explore by showing that Bayesian mixtures over exponential models can be seen as natural functions φ that lead to p.d. kernels, thus making the link with the particular case treated by Cuturi and Vert (2005).

This paper is organized as follows. We first introduce elements of measure representations of weighted lists and define the semigroup formalism and the notion of semigroup p.d. kernel in Section 2. Section 3 contains two examples of semigroup p.d. kernels, which are however usually not defined for molecular measures: the entropy kernel and the inverse generalized variance (IGV) kernel. Through regularization procedures, practical applications of such kernels on molecular measures are proposed in Section 4, and the approach is further extended by kernelizing the IGV through an a priori kernel defined itself on the space of components in Section 5. Section 6 contains the general integral representation of semigroup kernels and Section 7 makes the link between p.d. kernels and Bayesian posterior mixture probabilities. Finally, Section 8 contains an empirical evaluation of the proposed kernels on a benchmark experiment of handwritten digits classification.

2. Notations and Framework: Semigroup Kernels on Measures

In this section we set up the framework and notations of this paper, in particular the idea of semigroup kernel on the semigroup of measures.

2.1 Measures on Basic Components

We model the space of basic components by a Hausdorff space $(\mathcal{X}, \mathcal{B}, \nu)$ endowed with its Borel σ-algebra and a Borel dominant measure ν. A positive Radon measure μ is a positive Borel measure which satisfies $(i)\,\mu(C) < +\infty$ for any compact subset $C \subseteq \mathcal{X}$ and $(ii)\,\mu(B) = \sup\{\mu(C) | C \subseteq B, C$ compact $\}$ for any $B \in \mathcal{B}$ (see for example Berg et al. (1984) for the construction of Radon measures on Hausdorff spaces). The set of positive bounded (i.e., $\mu(\mathcal{X}) < +\infty$) Radon measures on $\mathcal{X}$ is denoted by $M_+^b(\mathcal{X})$. We introduce the subset of $M_+^b(\mathcal{X})$ of molecular (or atomic) measures $\mathrm{Mol}_+(\mathcal{X})$, namely measures such that

$$\mathrm{supp}(\mu) \stackrel{\mathrm{def}}{=} \{x \in \mathcal{X} \,|\, \mu(U) > 0, \text{ for all open subset } U \text{ s.t. } x \in U\}$$

is finite, and we denote by $\delta_x \in \mathrm{Mol}_+(\mathcal{X})$ the molecular (Dirac) measure of weight 1 on x. For a molecular measure μ, an *admissible base* of μ is a finite list γ of weighted points of $\mathcal{X}$, namely $\gamma = (x_i, a_i)_{i=1}^d$, where $x_i \in \mathcal{X}$ and $a_i > 0$ for $1 \leq i \leq d$, such that $\mu = \sum_{i=1}^d a_i \delta_{x_i}$. We write in that case $|\gamma| = \sum_{i=1}^d a_i$ and $l(\gamma) = d$. Reciprocally, a measure μ is said to be the image measure of a list of weighted elements γ if the previous equality holds. Finally, for a Borel measurable function $f \in \mathbb{R}^{\mathcal{X}}$ and a Borel measure μ, we write $\mu[f] = \int_{\mathcal{X}} f d\mu$.

2.2 Semigroups and Sets of Points

We follow in this paper the definitions found in Berg et al. (1984), which we now recall. An *Abelian semigroup* $(\mathcal{S}, +)$ is a nonempty set S endowed with an *associative* and *commutative composition* $+$ and a neutral element 0. Referring further to the notations used in Berg et al. (1984), note that we will only use auto-involutive semigroups in this paper, and will hence not discuss other semigroups which admit different involutions.

A function $\varphi : \mathcal{S} \to \mathbb{R}$ is called a *positive definite* (resp. *negative definite*, n.d.) function on the semigroup $(S, +)$ if $(s,t) \mapsto \varphi(s+t)$ is a p.d. (resp. n.d.) kernel on $\mathcal{S} \times \mathcal{S}$. The symmetry of the kernel being ensured by the commutativity of $+$, the positive definiteness is equivalent to the fact that the inequality

$$\sum_{i,j=1}^{N} c_i c_j \varphi(x_i + x_j) \geq 0$$

holds for any $N \in \mathbb{N}, (x_1, \ldots, x_N) \in \mathcal{S}^N$ and $(c_1 \ldots, c_n) \in \mathbb{R}^N$. Using the same notations, and adding the additional condition that $\sum_{i=1}^n c_i = 0$ yields the definition of negative definiteness as φ satisfying now

$$\sum_{i,j=1}^{N} c_i c_j \varphi(x_i + x_j) \leq 0.$$

Hence semigroup kernels are real-valued functions φ defined on the set of interest $\mathcal{S}$, the similarity between two elements s,t of $\mathcal{S}$ being just the value taken by that function on their composition, namely $\varphi(s+t)$.

Recalling our initial goal to quantify the similarity between two complex objects through finite weighted lists of elements in $\mathcal{X}$, we note that $(\mathcal{P}(\mathcal{X}), \cup)$ the set of subsets of $\mathcal{X}$ equipped with the usual union operator $\cup$ is a semigroup. Such a semigroup might be used as a feature representation for complex objects by mapping an object to the set of its components, forgetting about the weights. The resulting representation would therefore be an element of $\mathcal{P}(\mathcal{X})$. A semigroup kernel k on $\mathcal{P}(\mathcal{X})$ measuring the similarity of two sets of points $A, B \in \mathcal{P}(\mathcal{X})$ would use the value taken by a given p.d. function φ on their union, namely $k(A,B) = \varphi(A \cup B)$. However we put aside this framework for two reasons. First, the union composition is idempotent (i.e., for all A in $\mathcal{P}(\mathcal{X})$, we have $A \cup A = A$) which as noted in Berg et al. (1984, Proposition 4.4.18) drastically restricts the class of possible p.d. functions. Second, such a framework defined by sets would ignore the frequency (or weights) of the components described in lists, which can be misleading when dealing with finite sets of components. Other problematic features would include the fact that $k(A,B)$ would be constant when $B \subset A$ regardless of its characteristics, and that comparing sets of very different sizes should be difficult.

In order to overcome these limitations we propose to represent a list of weighted points $z = (x_i, a_i)_{i=1}^d$, where for $1 \leq i \leq d$ we have $x_i \in \mathcal{X}$ and $a_i > 0$, by its image measure $\delta_z = \sum_{i=1}^d a_i \delta_{x_i}$, and

focus now on the Abelian semigroup $(M_+^b(\mathcal{X}),+)$ to define kernels between lists of weighted points. This representation is richer than the one suggested in the previous paragraph in the semigroup $(\mathcal{P}(\mathcal{X}),\cup)$ to consider the merger of two lists. First it performs the union of the supports; second the sum of such molecular measures also adds the weights of the points common to both measures, with a possible renormalization on those weights. Two important features of the original list are however lost in this mapping: the order of its elements and the original frequency of each element within the list as a weighted singleton. We assume for the rest of this paper that this information is secondary compared to the one contained in the image measure, namely its unordered support and the *overall* frequency of each point in that support. As a result, we study in the following sections p.d. functions on the semigroup $(M_+^b(\mathcal{X}),+)$, in particular on molecular measures, in order to define kernels on weighted lists of simple components.

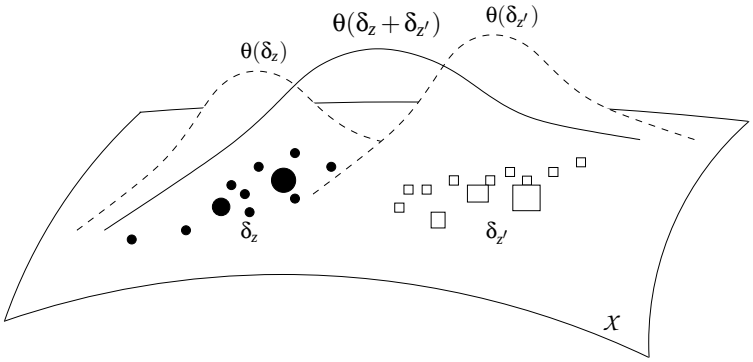


Figure 1: Measure representations of two lists z and z'. Each element of z (resp. z') list is represented by a black circle (resp. a white square), the size of which represents the associated weight. Five measures of interest are represented: the image measures δ_z and $\delta_{z'}$ of those weighted finite lists, the smoothed density estimates $\theta(\delta_z)$ and $\theta(\delta_{z'})$ of the two lists of points, and the smoothed density estimate $\theta(\delta_z+\delta_{z'})$ of the union of both lists.

Before starting the analysis of such p.d. functions, it should however be pointed out that several interesting semigroup p.d. kernels on measures are not directly applicable to molecular measures. For example, the first function we study below is only defined on the set of absolutely continuous measures with finite entropy. In order to overcome this limitation and be able to process complex objects in such situations, it is possible to think about alternative strategies to represent such objects by measures, as illustrated in Figure 1:

- The molecular measures δ_z and $\delta_{z'}$ as the image measures corresponding to the two weighted sets of points of z and z', where dots and squares represent the different weights applied on each points;

- Alternatively, smoothed estimates of these distributions obtained for example by non-parametric or parametric statistical density estimation procedures, and represented by $\theta(\delta_z)$ and $\theta(\delta_{z'})$ in Figure 1. Such estimates can be considered if a p.d. kernel is only defined for absolutely continuous measures. When this mapping takes the form of estimation among a given family

of densities (through maximum likelihood for instance) this can also be seen as a prior belief assumed on the distribution of the objects;

- Finally, a smoothed estimate of the sum $\delta_z + \delta_{z'}$ corresponding to the merging of both lists, represented by $\theta(\delta_z + \delta_{z'})$, can be considered. Note that $\theta(\delta_z + \delta_{z'})$ might differ from $\theta(\delta_z) + \theta(\delta_{z'})$.

A kernel between two lists of points can therefore be derived from a p.d. function on $(M_+^b(\mathcal{X}), +)$ in at least three ways:

$$k(z,z') = \begin{cases} \varphi(\delta_z + \delta_{z'}), & \text{using } \varphi \text{ directly on molecular measures,} \\ \varphi\left(\theta(\delta_z) + \theta(\delta_{z'})\right), & \text{using } \varphi \text{ on smoothed versions of the molecular measures,} \\ \varphi\left(\theta(\delta_z + \delta_{z'})\right), & \text{evaluating } \varphi \text{ on a smoothed version of the sum.} \end{cases}$$

The positive definiteness of φ on $M_+^b(\mathcal{X})$ ensures positive definiteness of k only in the first two cases. The third expression can be seen as a special case of the first one, where we highlight the usage of a preliminary mapping on the sum of two measures; in that case $\varphi \circ \theta$ should in fact be p.d. on $(M_+^b(\mathcal{X}), +)$, or at least $(\mathrm{Mol}_+(\mathcal{X}), +)$. Having defined the set of representations on which we will focus in this paper, namely measures on a set of components, we propose in the following section two particular cases of positive definite functions that can be computed through an addition between the considered measures. We then show how those quantities can be computed in the case of molecular measures in Section 4.

3. The Entropy and Inverse Generalized Variance Kernels

In this section we present two basic p.d. semigroup kernels for measures, motivated by a common intuition: the kernel between two measures should increase when the sum of the measures gets more "concentrated". The two kernels differ in the way they quantify the concentration of a measure, using either its entropy or its variance. They are therefore limited to a subset of measures, namely the subset of measures with finite entropy and the subset of sub-probability measures with non-degenerated variance, but are extended to a broader class of measures, including molecular measures, in Section 4.

3.1 Entropy Kernel

We consider the subset of $M_+^b(\mathcal{X})$ of absolutely continuous measures with respect to the dominant measure ν, and identify in this section a measure with its corresponding density with respect to ν. We further limit the subset to the set of non-negative valued ν-measurable functions on $\mathcal{X}$ with finite sum, such that

$$M_+^h(\mathcal{X}) \stackrel{\text{def}}{=} \{ f : \mathcal{X} \to \mathbb{R}^+ \,|\, f \text{ is } \nu\text{-measurable}, |h(f)| < \infty, |f| < \infty \}$$

where we write for any measurable non-negative valued function g,

$$h(g) \stackrel{\text{def}}{=} -\int_{\mathcal{X}} g \ln g \, d\nu,$$

(with $0\ln 0 = 0$ by convention) and $|g| \stackrel{\text{def}}{=} \int_{\mathcal{X}} g\, d\nu$, consistently with the notation used for measures. If g is such that $|g| = 1$, $h(g)$ is its differential entropy. Using the following inequalities,

$$(a+b)\ln(a+b) \leq a\ln a + b\ln b + (a+b)\ln 2, \text{ by convexity of } x \mapsto x\ln x,$$
$$(a+b)\ln(a+b) \geq a\ln a + b\ln b,$$

we have that $(M_+^h(\mathcal{X}),+)$ is an Abelian semigroup since for f, f' in $M_+^h(\mathcal{X})$ we have that $h(f+f')$ is bounded by integrating pointwise the inequalities above, the boundedness of $|f+f'|$ being also ensured. Following Rao (1987) we consider the quantity

$$J(f,f') \stackrel{\text{def}}{=} h(\frac{f+f'}{2}) - \frac{h(f)+h(f')}{2}, \tag{2}$$

known as the *Jensen divergence* (or Jensen-Shannon divergence) between f and f', which as noted by Fuglede and Topsøe (2004) can be seen as a symmetrized version of the Kullback-Leibler (KL) divergence D, since

$$J(f,f') = \frac{1}{2}D(f||\frac{f+f'}{2}) + \frac{1}{2}D(f'||\frac{f+f'}{2}).$$

The expression of Equation (2) fits our framework of devising semigroup kernels, unlike the direct use of the KL divergence (Moreno et al., 2004) which is neither symmetric nor negative definite. As recently shown in Endres and Schindelin (2003) and Österreicher and Vajda (2003), $\sqrt{J}$ is a metric on $M_+^h(\mathcal{X})$ which is a direct consequence of J's negative definiteness proven below, through Berg et al. (1984, Proposition 3.3.2) for instance. The Jensen-Divergence was also recently reinterpreted as a special case of a wider family of metrics on $M_+^b(\mathcal{X})$ derived from a particular family of Hilbertian metrics on $\mathbb{R}_+$ as presented in Hein and Bousquet (2005). The comparison between two densities f, f' is in that case performed by integrating pointwise the squared distance $d^2(f(x), f'(x))$ between the two densities over $\mathcal{X}$, using for d a distance chosen among a suitable family of metrics in $\mathbb{R}_+$ to ensure that the final value is independent of the dominant measure ν. The considered family for d is described in Fuglede and Topsøe (2004) through two parameters, a family of which the Jensen Divergence is just a special case as detailed in Hein and Bousquet (2005). The latter work shares with this paper another similarity, which lies in the "kernelization" of such quantities defined on measures through a prior kernel on the space of components, as will be reviewed in Section 5. However, of all the Hilbertian metrics introduced in Hein and Bousquet (2005), the Jensen-Divergence is the only one that can be related to the semigroup framework used throughout this paper.

Note finally that a positive definite kernel k is said to be infinitely divisible if $-\ln k$ is a negative definite kernel. As a consequence, any positive exponentiation $k^\beta, \beta > 0$ of an infinitely divisible kernel is a positive definite kernel.

Proposition 1 *h is a negative definite function on the semigroup $M_+^h(\mathcal{X})$. As a consequence e^{-h} is a positive definite function on $M_+^h(\mathcal{X})$ and its normalized counterpart, $k_h \stackrel{\text{def}}{=} e^{-J}$ is an infinitely divisible positive definite kernel on $M_+^h(\mathcal{X}) \times M_+^h(\mathcal{X})$.*

Proof It is known that the real-valued function $r : y \mapsto -y\ln y$ is n.d. on $\mathbb{R}_+$ as a semigroup endowed with addition (Berg et al., 1984, Example 6.5.16). As a consequence the function $f \mapsto r \circ f$ is n.d. on $M_+^h(\mathcal{X})$ as a pointwise application of r since $r \circ f$ is integrable w.r.t ν. For any real-valued n.d. kernel k and any real-valued function g, we have trivially that $(y,y') \mapsto k(y,y') + g(y) + g(y')$

remains negative definite. This allows first to prove that $h(\frac{f+f'}{2})$ is also n.d. through the identity $h(\frac{f+f'}{2}) = \frac{1}{2}h(f+f') + \frac{\ln 2}{2}(|f|+|f'|)$. Subtracting the normalization factor $\frac{1}{2}(h(f)+h(f'))$ gives the negative definiteness of J. This finally yields the positive definiteness of k_h as the exponential of the negative of a n.d. function through Schoenberg's theorem (Berg et al., 1984, Theorem 3.2.2). ■

Note that only e^{-h} is a semigroup kernel strictly speaking, since e^{-J} involves a normalized sum (through the division by 2) which is not associative. While both e^{-h} and e^{-J} can be used in practice on non-normalized measures, we name more explicitly $k_h = e^{-J}$ the *entropy kernel*, because what it indeed quantifies when f and f' are normalized (i.e., such that $|f| = |f'| = 1$) is the difference of the average of the entropy of f and f' from the entropy of their average. The subset of absolutely continuous *probability* measures on $(\mathcal{X}, \nu)$ with finite entropies, namely $\left\{ f \in M_+^h(\mathcal{X}), \text{ s.t.} \, |f| = 1 \right\}$ is not a semigroup since it is not closed by addition, but we can nonetheless define the restriction of J and hence k_h on it to obtain a p.d. kernel on probability measures inspired by semigroup formalism.

3.2 Inverse Generalized Variance Kernel

We assume in this subsection that $\mathcal{X}$ is an Euclidian space of dimension n endowed with Lebesgue's measure ν. Following the results obtained in the previous section, we propose under these restrictions a second semigroup p.d. kernel between measures which uses generalized variance. The generalized variance of a measure, namely the determinant of its variance matrix, is a quantity homogeneous to a volume in $\mathcal{X}$. This volume can be interpreted as a typical volume occupied by a measure when considering only its second order moments, making it hence a useful quantification of its dispersion. Besides being easy to compute in the case of molecular measures, this quantity is also linked to entropy if we consider that for normal laws $\mathcal{N}(m, \Sigma)$ the following relation holds:

$$\frac{1}{\sqrt{\det \Sigma}} \propto e^{-h(\mathcal{N}(m,\Sigma))}.$$

Through this observation, we note that considering the Inverse of the Generalized Variance (IGV) of a measure is equivalent to considering the value taken by e^{-2h} on its maximum likelihood normal law. We will put aside this interpretation in this section, before reviewing it with more care in Section 7.

Let us define the variance operator on measures μ with finite first and second moment of $M_+^b(\mathcal{X})$ as

$$\Sigma(\mu) \stackrel{\text{def}}{=} \mu[xx^\top] - \mu[x]\mu[x]^\top.$$

Note that $\Sigma(\mu)$ is always a positive semi-definite matrix when μ is a sub-probability measure, that is when $|\mu| \leq 1$, since

$$\Sigma(\mu) = \mu[(x - \mu[x])\,(x - \mu[x])^\top] + (1 - |\mu|)\,\mu[x]\mu[x]^\top.$$

We call $\det\Sigma(\mu)$ the generalized variance of a measure μ, and say a measure μ is *non-degenerated* if $\det\Sigma(\mu)$ is non-zero, meaning that $\Sigma(\mu)$ is of full rank. The subset of $M_+^b(\mathcal{X})$ of such measures with total weight equal to 1 is denoted by $M_+^\nu(\mathcal{X})$; $M_+^\nu(\mathcal{X})$ is convex through the following proposition:

Proposition 2 *$M_+^\nu(\mathcal{X}) \stackrel{\text{def}}{=} \left\{ \mu \in M_+^b(\mathcal{X}) : |\mu| = 1, \det\Sigma(\mu) > 0 \right\}$ is a convex set, and more generally for $\lambda \in [0,1)$, $\mu' \in M_+^b(\mathcal{X})$ such that $|\mu'| = 1$ and $\mu \in M_+^\nu(\mathcal{X})$, $(1-\lambda)\mu + \lambda\mu' \in M_+^\nu(\mathcal{X})$.*

Proof We use the following identity,

$$\Sigma\left((1-\lambda)\mu+\lambda\mu'\right) = (1-\lambda)\Sigma(\mu)+\lambda\Sigma(\mu')+\lambda(1-\lambda)\left(\mu[x]-\mu'[x]\right)\left(\mu[x]-\mu'[x]\right)^\top,$$

to derive that $\Sigma((1-\lambda)\mu+\lambda\mu')$ is a (strictly) positive-definite matrix as the sum of two positive semi-definite matrices and a strictly positive definite matrix $\Sigma(\mu)$. ■

$M_+^\nu(\mathcal{X})$ is not a semigroup, since it is not closed under addition. However we will work in this case on the mean of two measures in the same way we used their standard addition in the semigroup framework of $M_+^b(\mathcal{X})$.

Proposition 3 *The real-valued kernel k_ν defined on elements μ,μ' of $M_+^\nu(\mathcal{X})$ as*

$$k_\nu(\mu,\mu') = \frac{1}{\det\Sigma(\frac{\mu+\mu'}{2})}$$

is positive definite.

Proof Let y be an element of $\mathcal{X}$. For any $N\in\mathbb{N}$, any $c_1,...,c_N\in\mathbb{R}$ such that $\sum_i c_i = 0$ and any $\mu_1,...,\mu_N\in M_+^\nu(\mathcal{X})$ we have

$$\begin{aligned}\sum_{i,j} c_i c_j y^\top\Sigma(\frac{\mu_i+\mu_j}{2})y &= \sum_{i,j} c_i c_j y^\top\left(\frac{1}{2}\mu_i[xx^\top]+\frac{1}{2}\mu_j[xx^\top]-\right.\\ &\left.\frac{1}{4}\left(\mu_i[x]\mu_i[x]^\top+\mu_j[x]\mu_j[x]^\top+\mu_j[x]\mu_i[x]^\top+\mu_i[x]\mu_j[x]^\top\right)\right)y\\ &= -\frac{1}{4}\sum_{i,j} c_i c_j y^\top\left(\mu_j[x]\mu_i[x]^\top+\mu_i[x]\mu_j[x]^\top\right)y\\ &= -\frac{1}{2}\left(\sum_i c_i y^\top\mu_i[x]\right)^2 \leq 0,\end{aligned}$$

making thus the function $\mu,\mu'\mapsto y^\top\Sigma(\frac{\mu+\mu'}{2})y$ negative-definite for any $y\in\mathcal{X}$. Using again Schoenberg's theorem (Berg et al., 1984, Theorem 3.2.2) we have that $\mu,\mu'\mapsto e^{-y^\top\Sigma(\frac{\mu+\mu'}{2})y}$ is positive definite and so is the sum $\frac{1}{(2\pi)^{\frac{n}{2}}}\int_{\mathcal{X}} e^{-y^\top\Sigma(\frac{\mu+\mu'}{2})y}\nu(dy)$ which is equal to $1/\sqrt{\det\Sigma(\frac{\mu+\mu}{2})}$ ensuring thus the positive-definiteness of k_ν as its square. ■

Both entropy and IGV kernels are defined on subsets of $M_+^b(\mathcal{X})$. Since we are most likely to use them on molecular measures or smooth measures (as discussed in Section 2.2), we present in the following section practical ways to apply them in that framework.

4. Semigroup Kernels on Molecular Measures

The two positive definite functions defined in Sections 3.1 and 3.2 cannot be applied in the general case to $\mathrm{Mol}_+(\mathcal{X})$ which as exposed in Section 2 is our initial goal. In the case of the entropy kernel, molecular measures are generally not absolutely continuous with respect to ν (except on

finite spaces), and they have therefore no entropy; we solve this problem by mapping them into $M_+^h(\mathcal{X})$ through a smoothing kernel. In the case of the IGV, the estimates of variances might be poor if the number of points in the lists is not large enough compared to the dimension of the Euclidean space; we perform in that case a regularization by adding a unit-variance correlation matrix to the original variance. This regularization is particularly important to pave the way to the kernelized version of the IGV kernel presented in the next section, when $\mathcal{X}$ is not Euclidian but simply endowed with a prior kernel κ.

The application of both the entropy kernel and the IGV kernel to molecular measures requires a previous renormalization to set the total mass of the measures to 1. This technical renormalization is also beneficial, since it allows a consistent comparison of two weighted lists even when their size and total mass is very different. All molecular measures in this section, and equivalently all admissible bases, will hence be supposed to be normalized such that their total weight is 1, and $\mathrm{Mol}_+^1(\mathcal{X})$ denotes the subset of $\mathrm{Mol}_+(\mathcal{X})$ of such measures.

4.1 Entropy Kernel on Smoothed Estimates

We first define the Parzen smoothing procedure which allows to map molecular measures onto measures with finite entropy:

Definition 4 *Let κ be a probability kernel on $\mathcal{X}$ with finite entropy, i.e., a real-valued function defined on $\mathcal{X}^2$ such that for any $x \in \mathcal{X}$, $\kappa(x,\cdot) : y \mapsto \kappa(x,y)$ satisfies $\kappa(x,\cdot) \in M_+^h(\mathcal{X})$ and $|\kappa(x,\cdot)| = 1$. The κ-Parzen smoothed measure of μ is the probability measure whose density with respect to ν is $\theta_\kappa(\mu)$, where*

$$\begin{aligned} \theta_\kappa : \mathrm{Mol}_+^1(\mathcal{X}) &\longrightarrow M_+^h(\mathcal{X}) \\ \mu &\mapsto \sum_{x \in \mathrm{supp}\,\mu} \mu(x)\kappa(x,\cdot). \end{aligned}$$

Note that for any admissible base $(x_i, a_i)_{k=1}^d$ of μ we have that $\theta_\kappa(\mu) = \sum_{i=1}^d a_i \kappa(x_i,\cdot)$. Once this mapping is defined, we use the entropy kernel to propose the following kernel on two molecular measures μ and μ',

$$k_h^\kappa(\mu,\mu') = e^{-J(\theta_\kappa(\mu),\theta_\kappa(\mu'))}.$$

As an example, let $\mathcal{X}$ be an Euclidian space of dimension n endowed with Lebesgue's measure, and κ the isotropic Gaussian RBF kernel on that space, namely

$$\kappa(x,y) = \frac{1}{(2\pi\sigma)^{\frac{n}{2}}} e^{-\frac{\|x-y\|^2}{2\sigma^2}}.$$

Given two weighted lists z and z' of components in $\mathcal{X}$, $\theta_\kappa(\delta_z)$ and $\theta_\kappa(\delta_{z'})$ are thus mixtures of Gaussian distributions on $\mathcal{X}$. The resulting kernel computes the entropy of $\theta_\kappa(\delta_z)$ and $\theta_\kappa(\delta_{z'})$ taken separately and compares it with that of their mean, providing a positive definite quantification of their overlap.

4.2 Regularized Inverse Generalized Variance of Molecular Measures

In the case of a molecular measure μ defined on an Euclidian space $\mathcal{X}$ of dimension n, the variance $\Sigma(\mu)$ is simply the usual empirical estimate of the variance matrix expressed in an orthonormal basis

of $\mathcal{X}$:

$$\Sigma(\mu) = \mu[xx^\top] - \mu[x]\mu[x]^\top \quad = \sum_{i=1}^{d} a_i x_i x_i^\top - \left(\sum_{i=1}^{d} a_i x_i\right)\left(\sum_{i=1}^{d} a_i x_i\right)^\top,$$

where we use an admissible base $\gamma = (x_i, a_i)_{i=1}^d$ of μ to give a matrix expression of $\Sigma(\mu)$, with all points x_i expressed as column vectors. Note that this matrix expression, as would be expected from a function defined on measures, does not depend on the chosen admissible base. Given such an admissible base, let $X_\gamma = [x_i]_{i=1..d}$ be the $n \times d$ matrix made of all column vectors x_i and Δ_γ the diagonal matrix of weights of γ taken in the same order $(a_i)_{1\leq 1\leq d}$. If we write I_d for the identity matrix of rank d and $\mathbb{1}_{d,d}$ for the $d \times d$ matrix composed of ones, we have for any base γ of μ that:

$$\Sigma(\mu) = X_\gamma(\Delta_\gamma - \Delta_\gamma \mathbb{1}_{d,d}\Delta_\gamma)X_\gamma^\top,$$

which can be rewritten as

$$\Sigma(\mu) = X_\gamma(I_d - \Delta_\gamma \mathbb{1}_{d,d})\Delta_\gamma(I_d - \mathbb{1}_{d,d}\Delta_\gamma)X_\gamma^\top,$$

noting that $(\Delta_\gamma \mathbb{1}_{d,d})^2 = \Delta_\gamma \mathbb{1}_{d,d}$ since $\operatorname{trace}\Delta_\gamma = 1$.

The determinant of $\Sigma(\mu)$ can be equal to zero when the size of the support of μ is smaller than n, the dimension of $\mathcal{X}$, or more generally when the linear span of the points in the support of μ does not cover the whole space $\mathcal{X}$. This problematic case is encountered in Section 5 when we consider kernelized versions of the IGV, using an embedding of $\mathcal{X}$ into a functional Hilbert space of potentially infinite dimension. Mapping an element of $\mathrm{Mol}_+^1(\mathcal{X})$ into $M_+^v(\mathcal{X})$ by adding to it any element of $M_+^v(\mathcal{X})$ through Proposition 2 would work as a regularization technique; for an arbitrary $\rho \in M_+^v(\mathcal{X})$ and a weight $\lambda \in [0,1)$ we could use the kernel defined as

$$\mu, \mu' \mapsto \frac{1}{\det\Sigma\left(\lambda\frac{\mu+\mu'}{2} + (1-\lambda)\rho\right)}.$$

We use in this section a different strategy inspired by previous works (Fukumizu et al., 2004, Bach and Jordan, 2002) further motivated in the case of covariance operators on infinite dimensional spaces as shown by Cuturi and Vert (2005). The considered regularization consists in modifying directly the matrix $\Sigma(\mu)$ by adding a small diagonal component ηI_n where $\eta > 0$ so that its spectrum never vanishes. When considering the determinant of such a regularized matrix $\Sigma(\mu) + \eta I_n$ this is equivalent to considering the determinant of $\frac{1}{\eta}\Sigma(\mu) + I_n$ up to a factor η^n, which will be a more suitable expression in practice. We thus introduce the regularized kernel k_v^η defined on measures $(\mu, \mu') \in M_+^b(\mathcal{X})$ with finite second moment as

$$k_v^\eta(\mu, \mu') \stackrel{\text{def}}{=} \frac{1}{\det\left(\frac{1}{\eta}\Sigma\left(\frac{\mu+\mu'}{2}\right) + I_n\right)}.$$

It is straightforward to prove that the regularized function k_v^η is a positive definite kernel on the measures of $M_+^b(\mathcal{X})$ with finite second-order moments using the same proof used in Proposition 3. If we now introduce

$$K_\gamma \stackrel{\text{def}}{=} \left[x_i^\top x_j\right]_{1\leq i,j\leq d},$$

for the $d \times d$ matrix of dot-products associated with the elements of a base γ, and

$$\tilde{K}_\gamma \stackrel{\text{def}}{=} \left[(x_i - \sum_{k=1}^{d} a_k x_k)^\top (x_j - \sum_{k=1}^{d} a_k x_k) \right]_{1 \leq i,j \leq d} = (I_d - \mathbb{1}_{d,d}\Delta_\gamma) K_\gamma (I_d - \Delta_\gamma \mathbb{1}_{d,d}),$$

for its centered expression with respect to the mean of μ, we have the following result:

Proposition 5 *Let $\mathcal{X}$ be an Euclidian space of dimension n. For any $\mu \in \mathrm{Mol}_+^1(\mathcal{X})$ and* any *admissible base γ of μ we have*

$$\det\left(\frac{1}{\eta}\tilde{K}_\gamma \Delta_\gamma + I_{l(\gamma)}\right) = \det\left(\frac{1}{\eta}\Sigma(\mu) + I_n\right).$$

Proof We omit the references to μ and γ in this proof to simplify matrix notations, and write $d = l(\gamma)$. Let $\tilde{X}$ be the $n \times d$ matrix $[x_i - \sum_{j=1}^{d} a_j x_j]_{i=1..d}$ of centered column vectors enumerated in γ, namely $\tilde{X} = X(I_d - \Delta \mathbb{1}_{d,d})$. We have

$$\Sigma = \tilde{X}\Delta\tilde{X}^\top,$$
$$\tilde{K}\Delta = \tilde{X}^\top\tilde{X}\Delta.$$

Through the singular value decomposition of $\tilde{X}\Delta^{\frac{1}{2}}$, it is straightforward to see that the non-zero elements of the spectrums of matrices $\tilde{K}\Delta, \Delta^{\frac{1}{2}}\tilde{X}^\top\tilde{X}\Delta^{\frac{1}{2}}$ and Σ are identical. Thus, regardless of the difference between n and d, we have

$$\det\left(\frac{1}{\eta}\tilde{K}\Delta + I_d\right) = \det\left(\frac{1}{\eta}\Delta^{\frac{1}{2}}\tilde{X}^\top\tilde{X}\Delta^{\frac{1}{2}} + I_d\right) = \det\left(\frac{1}{\eta}\tilde{X}\Delta\tilde{X}^\top + I_n\right) = \det\left(\frac{1}{\eta}\Sigma + I_n\right),$$

where the addition of identity matrices only introduces an offset of 1 for all eigenvalues. ■

Given two measures $\mu, \mu' \in \mathrm{Mol}_+^1(\mathcal{X})$, the following theorem can be seen as a regularized equivalent of Proposition 3 through an application of Proposition 5 to $\mu'' = \frac{\mu+\mu'}{2}$.

Theorem 6 *Let $\mathcal{X}$ be an Euclidian space. The kernel k_v^η defined on two measures μ, μ' of $\mathrm{Mol}_+^1(\mathcal{X})$ as*

$$k_v^\eta(\mu, \mu') = \frac{1}{\det\left(\frac{1}{\eta}\tilde{K}_\gamma\Delta_\gamma + I_{l(\gamma)}\right)},$$

where γ is any admissible base of $\frac{\mu+\mu'}{2}$, is p.d. and independent of the choice of γ.

Given two objects z, z' and their respective molecular measures δ_z and $\delta_{z'}$, the computation of the IGV for two such objects requires in practice an admissible base of $\frac{\delta_z+\delta_{z'}}{2}$ as seen in Theorem 6. This admissible base can be chosen to be of the cardinality of the support of the mixture of δ_z and $\delta_{z'}$, or alternatively be the simple merger of two admissible bases of z and z' with their weights divided by 2, without searching for overlapped points between both lists. This choice has no impact on the final value taken by the regularized IGV-kernel and can be arbitrated by computational considerations.

If we now take a practical look at the IGV's definition, we note that it can be applied but to cases where the component space $\mathcal{X}$ is Euclidian, and only if the studied measures can be summarized efficiently by their second order moments. These limitations do not seem very realistic in practice,

since $\mathcal{X}$ may not have a vectorial structure, and the distribution of the components may not even be well represented by Gaussians in the Euclidian case. We propose to bypass this issue and introduce the usage of the IGV in a more flexible framework by using the kernel trick on the previous quantities, since the IGV of a measure can be expressed only through the dot-products between the elements of the support of the considered measure.

5. Inverse Generalized Variance on the RKHS Associated with a Kernel κ

As with many quantities defined by dot-products, one is tempted to replace the usual dot-product matrix $\tilde{K}$ of Theorem 6 by an alternative Gram-matrix obtained through a p.d. kernel κ defined on $\mathcal{X}$. The advantage of such a substitution, which follows the well known "kernel trick" principle (Schölkopf and Smola, 2002), is multiple as it first enables us to use the IGV kernel on any non-vectorial space endowed with a kernel, thus in practice on any component space endowed with a kernel; second, it is also useful when $\mathcal{X}$ is a dot-product space where a non-linear kernel can however be used (e.g., using Gaussian kernel) to incorporate into the IGV's computation higher-order moment comparisons. We prove in this section that the inverse of the regularized generalized variance, computed in Proposition 5 through the centered dot-product matrix $\tilde{K}_\gamma$ of elements of any admissible base γ of μ, is still a positive definite quantity if we replace $\tilde{K}_\gamma$ by a centered Gram-matrix $\tilde{\mathcal{K}}_\gamma$, computed through an a priori kernel κ on $\mathcal{X}$, namely

$$\begin{aligned}\mathcal{K}_\gamma &= [\kappa(x_i,x_j)]_{1\leq i,j\leq d}\\ \tilde{\mathcal{K}}_\gamma &= (I_d - \mathbb{1}_{d,d}\Delta_\gamma)\,\mathcal{K}_\gamma(I_d - \Delta_\gamma \mathbb{1}_{d,d}).\end{aligned}$$

This substitution follows also a general principle when considering kernels on measures. The "kernelization" of a given kernel defined on measures to take into account a prior similarity on the components, when computationally feasible, is likely to improve its overall performance in classification tasks, as observed in Kondor and Jebara (2003) but also in Hein and Bousquet (2005) under the "Structural Kernel" appellation. The following theorem proves that this substitution is valid in the case of the IGV.

Theorem 7 *Let $\mathcal{X}$ be a set endowed with a p.d. kernel κ. The kernel*

$$k_\kappa^\eta(\mu,\mu') = \frac{1}{\det\left(\frac{1}{\eta}\tilde{\mathcal{K}}_\gamma\Delta_\gamma + I_{l(\gamma)}\right)}, \tag{3}$$

defined on two elements μ,μ' in $\mathrm{Mol}_+^1(\mathcal{X})$ is positive definite, where γ is any *admissible base of $\frac{\mu+\mu'}{2}$.*

Proof Let $N\in\mathbb{N}$, $\mu_1,..,\mu_N\in\mathrm{Mol}_+^1(\mathcal{X})$ and $(c_i)_{i=1}^N\in\mathbb{R}^N$. Let us now study the quantity $\sum_{i=1}^N c_i c_j k_\kappa^\eta(\mu_i,\mu_j)$. To do so we introduce by the Moore-Aronszajn theorem (Berlinet and Thomas-Agnan, 2003, p.19) the reproducing kernel Hilbert space Ξ with reproducing kernel κ indexed on $\mathcal{X}$. The usual mapping from $\mathcal{X}$ to Ξ is denoted by ϕ, that is $\phi:\mathcal{X}\ni x\mapsto\kappa(x,\cdot)$. We define

$$\mathcal{Y}\stackrel{\text{def}}{=}\mathrm{supp}\left(\sum_{i=1}^N\mu_i\right)\subset\mathcal{X},$$

the finite set which numbers all elements in the support of the N considered measures, and

$$\Upsilon\stackrel{\text{def}}{=}\mathrm{span}\,\phi(\mathcal{Y})\subset\Xi,$$

the linear span of the elements in the image of $\mathcal{Y}$ through ϕ. Υ is a vector space whose finite dimension is upper-bounded by the cardinality of $\mathcal{Y}$. Endowed with the dot-product inherited from Ξ, we further have that Υ is Euclidian. Given a molecular measure $\mu \in \mathrm{Mol}_+^1(\mathcal{Y})$, let $\phi(\mu)$ denote the image measure of μ in Υ, namely $\phi(\mu) = \sum_{x\in\mathcal{Y}} \mu(x)\delta_{\phi(x)}$. One can easily check that any admissible base $\gamma = (x_i, a_i)_{i=1}^d$ of μ can be used to provide an admissible base $\phi(\gamma) \stackrel{\text{def}}{=} (\phi(x_i), a_i)_{i=1}^d$ of $\phi(\mu)$. The weight matrices Δ_γ and $\Delta_{\phi(\gamma)}$ are identical and we further have $\tilde{\mathcal{K}}_\gamma = \tilde{K}_{\phi(\gamma)}$ by the reproducing property, where $\tilde{K}$ is defined by the dot-product of the Euclidian space Υ induced by κ. As a result, we have that $k_\kappa^\eta(\mu_i,\mu_j) = k_v^\eta(\phi(\mu_i),\phi(\mu_j))$ where k_v^η is defined on $\mathrm{Mol}_+^1(\Upsilon)$, ensuring the non-negativity

$$\sum_{i=1}^N c_i c_j k_\kappa^\eta(\mu_i,\mu_j) = \sum_{i=1}^N c_i c_j k_v^\eta(\phi(\mu_i),\phi(\mu_j)) \geq 0$$

and hence positive-definiteness of k_κ^η. ■

As bserved in the experimental section, the kernelized version of the IGV is more likely to be successful to solve practical tasks since it incorporates meaningful information on the components. Before observing these practical improvements, we provide a general study of the family of semigroup kernels on $M_+^b(\mathcal{X})$ by casting the theory of integral representations of positive definite functions on a semigroup (Berg et al., 1984) in the framework of measures, providing new results and possible interpretations of this class of kernels.

6. Integral Representation of Positive Definite Functions on a Set of Measures

In this section we study a general characterization of *all* p.d. functions on the whole semigroup $(M_+^b(\mathcal{X}),+)$, including thus measures which are not normalized. This characterization is based on a general integral representation theorem valid for any semigroup kernel, and is similar in spirit to the representation of p.d. functions obtained on Abelian groups through Bochner's theorem (Rudin, 1962). Before stating the main results in this section we need to recall basic definitions of semicharacters and exponentially bounded function (Berg et al., 1984, chap. 4).

Definition 8 *A real-valued function* ρ *on an Abelian semigroup* $(S,+)$ *is called a* semicharacter *if it satisfies the following properties:*

(i) $\rho(0) = 1$

(ii) $\forall s,t \in S, \rho(s+t) = \rho(s)\rho(t)$.

It follows from the previous definition and the fact that $M_+^b(\mathcal{X})$ is *2-divisible* (i.e., $\forall \mu \in M_+^b(\mathcal{X}), \exists \mu' \in M_+^b(\mathcal{X})$ s.t. $\mu = 2\mu'$) that semicharacters are nonnegative valued since it suffices to write that $\rho(\mu) = \rho(\frac{\mu}{2})^2$. Note also that semicharacters are trivially positive definite functions on S. We denote by S^* the set of semicharacters on $M_+^b(\mathcal{X})$, and by $\hat{S} \subset S^*$ the set of bounded semicharacters. S^* is a Hausdorff space when endowed with the topology inherited from $\mathbb{R}^S$ having the topology of pointwise convergence. Therefore we can consider the set of Radon measures on S^*, namely $M_+^b(S^*)$.

Definition 9 *A function* $f : M_+^b(\mathcal{X}) \to \mathbb{R}$ *is called* exponentially bounded *if there exists a function* $\alpha : M_+^b(\mathcal{X}) \to \mathbb{R}_+$ *(called an* absolute value*) satisfying* $\alpha(0) = 1$ *and* $\alpha(\mu+\mu') \leq \alpha(\mu)\alpha(\mu')$ *for*

$\mu, \mu' \in M_+^b(\mathcal{X})$, *and a constant* $C > 0$ *such that:*

$$\forall \mu \in M_+^b(\mathcal{X}), \quad f(\mu) \leq C\alpha(\mu).$$

We can now state two general integral representation theorems for p.d. functions on semigroups (Berg et al., 1984, Theorems 4.2.5 and 4.2.8). These theorems being valid on any semigroup, they hold in particular on the particular semigroup $(M_+^b(\mathcal{X}), +)$.

Theorem 10
- *A function* $\varphi : M_+^b(\mathcal{X}) \to \mathbb{R}$ *is p.d. and exponentially bounded if and only if it has an integral representation:*

$$\varphi(s) = \int_{S^*} \rho(s) d\omega(\rho),$$

with $\omega \in M_+^c(S^*)$ *(the set of Radon measures on* S^* *with compact support).*

- *A function* $\varphi : M_+^b(\mathcal{X}) \to \mathbb{R}$ *is p.d. and bounded if and only if it has an integral representation of the form:*

$$\varphi(s) = \int_{\hat{S}} \rho(s) d\omega(\rho),$$

with $\omega \in M_+(\hat{S})$.

In both cases, if the integral representation exists, then there is uniqueness of the measure ω *in* $M_+(S^*)$.

In order to make these representations more constructive, we need to study the class of (bounded) semicharacters on $(M_+^b(\mathcal{X}), +)$. Even though we are not able to provide a complete characterization, even of bounded semicharacters, the following proposition introduces a large class of semicharacters, and completely characterizes the *continuous* semicharacters. For matters related to continuity of functions defined on $M_+^b(\mathcal{X})$, we will consider the weak topology of $M_+^b(\mathcal{X})$ which is defined in simple terms through the *portmanteau* theorem (Berg et al., 1984, Theorem 2.3.1). Note simply that if μ_n converges to μ in the weak topology then for any *bounded* measurable and continuous function f we have that $\mu_n[f] \to \mu[f]$. We further denote by $C(\mathcal{X})$ the set of continuous real-valued functions on $\mathcal{X}$ and by $C^b(\mathcal{X})$ its subset of bounded functions. Both sets are endowed with the topology of pointwise convergence. For a function $f \in \mathbb{R}^{\mathcal{X}}$ we write ρ_f for the function $\mu \mapsto e^{\mu[f]}$ when the integral is well defined.

Proposition 11 *A semicharacter* $\rho : M_+^b(\mathcal{X}) \to \mathbb{R}$ *is continuous on* $(M_+^b(\mathcal{X}), +)$ *endowed with the weak topology if and only if there exists* $f \in C^b(\mathcal{X})$ *such that* $\rho = \rho_f$. *In that case,* ρ *is a bounded semicharacter on* $M_+^b(\mathcal{X})$ *if and only if* $f \leq 0$.

Proof For a continuous and bounded function f, the semicharacter ρ_f is well-defined. If a sequence μ_n in $M_+^b(\mathcal{X})$ converges to μ weakly, we have $\mu_n[f] \to \mu[f]$, which implies the continuity of ρ_f. Conversely, suppose ρ is weakly continuous. Define $f : \mathcal{X} \to [-\infty, \infty)$ by $f(x) = \log \rho(\delta_x)$. If a sequence x_n converges to x in $\mathcal{X}$, obviously we have $\delta_{x_n} \to \delta_x$ in the weak topology, and

$$\rho(\delta_{x_n}) \quad \to \quad \rho(\delta_x),$$

which means the continuity of f. To see the boundedness of f, assume the contrary. Then, we can find $x_n \in \mathcal{X}$ such that either of $0 < f(x_n) \to \infty$ or $0 > f(x_n) \to -\infty$ holds. Let $\beta_n = |f(x_n)|$. Because the measure $\frac{1}{\beta_n}\delta_{x_n}$ converges weakly to zero, the continuity of ρ means

$$\rho\left(\tfrac{1}{\beta_n}\delta_{x_n}\right) \to 1,$$

which contradicts with the fact $\rho(\frac{1}{\beta_n}\delta_{x_n}) = e^{\frac{1}{\beta_n} f(x_n)} = e^{\pm 1}$. Thus, ρ_f is well-defined, weakly continuous on $M_+^b(\mathcal{X})$ and equal to ρ on the set of molecular measures. It is further equal to ρ on $M_+^b(\mathcal{X})$ through the denseness of molecular measures in $M_+^b(\mathcal{X})$, both in the weak and the pointwise topology (Berg et al., 1984, Proposition 3.3.5). Finally suppose now that ρ_f is bounded and that there exists x in $\mathcal{X}$ such that $f(x) > 0$. By $\rho_f(n\delta_x) = e^{nf(x)}$ which diverges with n we see a contradiction. The converse is straightforward. ■

Let ω be a bounded nonnegative Radon measure on the Hausdorff space of continuous real-valued functions on $\mathcal{X}$, namely $\omega \in M_+^b(C(\mathcal{X}))$. Given such a measure, we first define the subset M_ω of $M_+^b(\mathcal{X})$ as

$$M_\omega = \{\mu \in M_+^b(\mathcal{X}) \mid \sup_{f \in \operatorname{supp}\omega} \mu[f] < +\infty\}.$$

M_ω contains the null measure and is a semigroup.

Corollary 12 *For any bounded Radon measure $\omega \in M_+^b(C(\mathcal{X}))$, the following function φ is a p.d. function on the semigroup $(M_\omega, +)$:*

$$\varphi(\mu) = \int_{C(\mathcal{X})} \rho_f(\mu)\, d\omega(f). \tag{4}$$

If $\operatorname{supp}\omega \subset C^b(\mathcal{X})$ then φ is continuous on M_ω endowed with the topology of weak convergence.

Proof For $f \in \operatorname{supp}\omega$, ρ_f is a well defined semicharacter on M_ω and hence positive definite. Since

$$\varphi(\mu) \leq |\omega| \sup_{f \in \operatorname{supp}\omega} \mu[f]$$

is bounded, φ is well defined and hence positive definite. Suppose now that $\operatorname{supp}\omega \subset C^b(\mathcal{X})$ and let μ_n be a sequence of M_ω converging weakly to μ. By the bounded convergence theorem and continuity of all considered semicharacters (since all considered functions f are bounded) we have that:

$$\lim_{n\to\infty} \varphi(\mu_n) = \int_{C(\mathcal{X})} \lim_{n\to\infty} \rho_f(\mu_n)\, d\omega(f) = \varphi(\mu).$$

and hence φ is continuous w.r.t the weak topology. ■

When the measure ω is chosen in such a way that the integral (4) is tractable or can be approximated, then a valid p.d. kernel for measures is obtained; an example involving mixtures over exponential families is provided in Section 7.

Before exploiting this constructive representation, a few remarks should be pointed out. When using non-bounded functions (as is the case when using expectation or second-order moments of measures) the continuity of the integral φ is left undetermined to our knowledge, even when its existence is ensured. However, when $\mathcal{X}$ is compact we have that $C(\mathcal{X}) = C^b(\mathcal{X})$ and hence continuity

on M_ω of any function φ constructed through corollary 12. Conversely, there exist continuous p.d. functions on $(M_+^b(\mathcal{X}),+)$ that can not be represented in the form (4). Although any continuous p.d. function can necessarily be represented as an integral of semicharacters by Theorem 10, the semicharacters involved in the representation are not necessarily continuous as in (4). An example of such a continuous p.d. function written as an integral of non-continuous semicharacters is exposed in Appendix A. It is an open problem to our knowledge to fully characterize continuous p.d. functions on $(M_+^b(\mathcal{X}),+)$.

7. Projection on Exponential Families through Laplace's Approximation

The constructive approach presented in corollary 12 can be used in practice to define kernels by restricting the space $C(\mathcal{X})$ to subspaces where computations are tractable. A natural way to do so is to consider a vector space of finite dimension s of $C(\mathcal{X})$, namely the span of a free family of s non-constant functions $f_1,...,f_s$ of $C(\mathcal{X})$, and define a measure on that subspace by applying a measure on the weights associated with each function. The previous integral representation (4) would then take the form:

$$\varphi(\mu) = \int_\Theta e^{\mu[\sum_{i=1}^s \theta_i f_i]}\, \omega(d\theta),$$

where ω is now a bounded measure on a compact subset $\Theta \subseteq \mathbb{R}^s$ and μ is such that $\mu[f_i] < +\infty$ for $1 \leq i \leq s$. The subspace of $C(\mathcal{X})$ considered in this section is however slightly different, in order to take advantage of the natural benefits of exponential densities generated by all functions $f_1,...,f_s$. Following Amari and Nagaoka (2001, p.69), this requires the definition of the cumulant generating function of ν with respect to $f_1,...,f_s$ as

$$\psi(\theta) \stackrel{\text{def}}{=} \log \nu[e^{\sum_{i=1}^s \theta_i f_i}],$$

such that for each $\theta \in \Theta$,

$$p_\theta \stackrel{\text{def}}{=} \exp\left(\sum_{i=1}^s \theta_i f_i - \psi(\theta)\right) \nu,$$

is a probability density, which defines an exponential family of densities on $\mathcal{X}$ as θ varies in Θ. Rather than the direct span of functions $f_1,...,f_s$ on Θ, this is equivalent to considering the hypersurface $\{\sum_{i=1}^s \theta_i f_i - \psi(\theta)\}$ in $\text{span}\{f_1,..,f_s,-1\}$. This yields the following expression:

$$\varphi(\mu) = \int_\Theta e^{\mu[\sum_{i=1}^s \theta_i f_i - \psi(\theta)]}\, \omega(d\theta).$$

Following the notations of Amari and Nagaoka (2001) the η-parameters (or expectation parameters) of μ are defined as

$$\hat{\eta}_i \stackrel{\text{def}}{=} \frac{1}{|\mu|}\mu[f_i],\ 1 \leq i \leq s,$$

and $\hat{\theta}$ stands for the θ-parameters of $\hat{\eta}$. We assume in the following approximations that $\hat{\theta} \in \Theta$ and recall two identities (Amari and Nagaoka, 2001, Chapters 3.5 & 3.6):

$$\chi(\theta) \stackrel{\text{def}}{=} \sum_{i=1}^s \theta_i \eta_i - \psi(\theta) = -h(\theta),\ \text{the dual potential},$$

$$D(\theta||\theta') = \psi(\theta) + \chi(\theta') - \sum_{i=1}^s \theta_i \eta'_i,\ \text{the KL divergence},$$

where we used the abbreviations $h(\theta) = h(p_\theta)$ and $D(\theta||\theta') = D(p_\theta||p_{\theta'})$. We can then write

$$\begin{aligned}\mu[\sum_{i=1}^{s}\theta_i f_i - \psi(\theta)] &= |\mu|\left(\sum_{i=1}^{s}\theta_i\hat{\eta}_i - \psi(\theta)\right)\\ &= |\mu|\left(\sum_{i=1}^{s}\hat{\theta}_i\hat{\eta}_i - \psi(\hat{\theta}) + \sum_{i=1}^{s}(\theta_i-\hat{\theta}_i)\hat{\eta}_i + \psi(\hat{\theta}) - \psi(\theta)\right)\\ &= -|\mu|\left(h(\hat{\theta}) + D(\hat{\theta}||\theta)\right),\end{aligned}$$

to obtain the following factorized expression,

$$\varphi(\mu) = e^{-|\mu|h(\hat{\theta})}\int_\Theta e^{-|\mu|D(\hat{\theta}||\theta)}\omega(d\theta). \tag{5}$$

The quantity $e^{-|\mu|h(\hat{\theta})}$ was already evoked in Section 3.2 when multivariate normal distributions were used to express the IGV kernel. When $\mathcal{X}$ is an Euclidian space of dimension n, this is indeed equivalent to defining $s = n + n(n+1)/2$ base functions, more precisely $f_i = x_i$ and $f_{ij} = x_i x_j$, and dropping the integral of Equation (5). Note that such functions are not bounded and that M_ω corresponds here to the set of measures with finite first and second order moments.

The integral of Equation (5) cannot be computed in a general case. The use of conjugate priors can however yield exact calculations, such as in the setting proposed by Cuturi and Vert (2005). In their work $\mathcal{X}$ is a finite set of short sequences formed over an alphabet, functions f_i are all possible indicator functions of $\mathcal{X}$ and ω is an additive mixture of Dirichlet priors. The kernel value is computed through a factorization inspired by the context-tree weighting algorithm (Willems et al., 1995). In the general case a numerical approximation can also be derived using Laplace's method (Dieudonné, 1968) under the assumption that $|\mu|$ is large enough. To do so, first notice that

$$\begin{aligned}\frac{\partial D(\hat{\theta}||\theta)}{\partial\theta_i}\,|_{\theta=\hat{\theta}} = \frac{\partial\psi}{\partial\theta_i}\,|_{\theta=\hat{\theta}} - \hat{\eta}_i = 0,\\ \frac{\partial D(\hat{\theta}||\theta)}{\partial\theta_i\partial\theta_j} = \frac{\partial\psi}{\partial\theta_i\partial\theta_j} = g_{ij}(\theta),\end{aligned}$$

where $G_\theta = [g_{ij}(\theta)]$ is the Fisher information matrix computed in θ and hence a p.d. matrix. The following approximation then holds:

$$\varphi(\mu) \underset{|\mu|\to\infty}{\sim} e^{-|\mu|h(\hat{\theta})}\int_{\mathbb{R}^s}\omega(\hat{\theta})e^{-\frac{|\mu|}{2}(\theta-\hat{\theta})^\top G_{\hat{\theta}}(\theta-\hat{\theta})}d\theta = e^{-|\mu|h(\hat{\theta})}\left(\frac{2\pi}{|\mu|}\right)^{\frac{s}{2}}\frac{\omega(\hat{\theta})}{\sqrt{\det G_{\hat{\theta}}}}$$

which can be simplified by choosing ω to be Jeffrey's prior (Amari and Nagaoka, 2001, p.44), namely

$$\omega(d\theta) = \frac{1}{V}\sqrt{\det G_\theta}\,d\theta, \qquad \text{where } V = \int_\Theta\sqrt{\det G_\theta}\,d\theta.$$

Up to a multiplication by V this provides an approximation of φ by $\tilde{\varphi}$ as

$$\varphi(\mu) \underset{|\mu|\to\infty}{\sim} \tilde{\varphi}(\mu) \overset{\text{def}}{=} e^{-|\mu|h(\hat{\theta})}\left(\frac{2\pi}{|\mu|}\right)^{\frac{s}{2}}.$$

The η-coordinates of μ are independent of the total weight $|\mu|$, hence $\tilde{\varphi}(2\mu) = \tilde{\varphi}(\mu)^2(\frac{|\mu|}{4\pi})^{\frac{s}{2}}$. This identity can be used to propose a renormalized kernel for two measures as

$$k(\mu,\mu') \stackrel{\text{def}}{=} \frac{\tilde{\varphi}(\mu+\mu')}{\sqrt{\tilde{\varphi}(2\mu)\tilde{\varphi}(2\mu')}} = \frac{e^{-(|\mu+\mu'|)h(p_{\mu+\mu'})}}{e^{-|\mu|h(p_\mu)-|\mu'|h(p_{\mu'})}} \left(\frac{2\sqrt{|\mu||\mu'|}}{|\mu|+|\mu'|}\right)^{\frac{s}{2}}.$$

where p_μ stands for $p_{\hat{\theta}_\mu}$. When μ and μ' are normalized such that their total weight coincides and is equal to β, we have that

$$k(\mu,\mu') = e^{-2\beta\left(h(p_{\mu''})-\frac{h(p_\mu)+h(p_{\mu'})}{2}\right)}, \tag{6}$$

where $\mu'' = \mu+\mu'$. From Equation (6), we see that β can be tuned in practice and thought of as a width parameter. It should be large enough to ensure the consistency of Laplace's approximation and thus positive definiteness, while not too large at the same time to avoid diagonal dominance issues. In the case of the IGV kernel this tradeoff can however be put aside since the inverse of the IGV is directly p.d. as was proved in Proposition 3. However and to our knowledge this assertion does not stand in a more general case when the functions $f_1,...,f_s$ are freely chosen.

8. Experiments on Images of the MNIST Database

We present in this section experimental results and discussions on practical implementations of the IGV kernels on a benchmark experiment of handwritten digits classification. We focus more specifically on the kernelized version of the IGV and discuss its performance with respect to other kernels. The entropy kernel performed very poorly in the series of experiments presented here, besides requiring a time consuming Monte Carlo computation, which is why we do not consider it in this section. We believe however that in more favourable cases, notably when the considered measures are multinomials, the entropy kernel and its structural variants (Hein and Bousquet, 2005) may provide good results.

8.1 Linear IGV Kernel

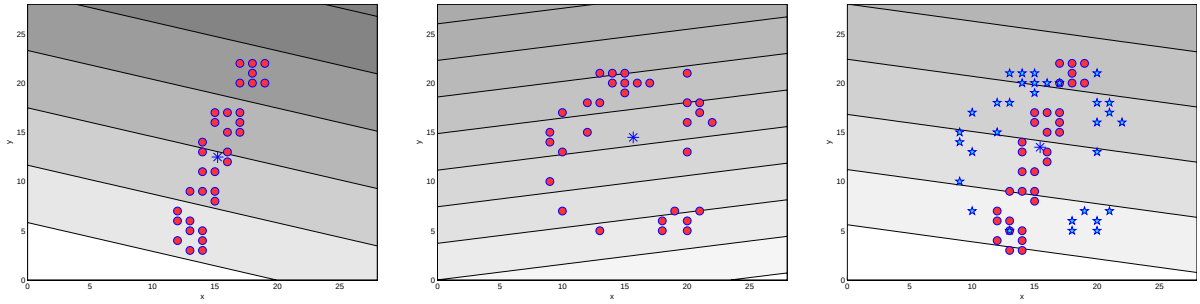
Following the previous work of Kondor and Jebara (2003), we have conducted experiments on 500 and 1000 images (28×28 pixels) taken from the MNIST database of handwritten digits (black shapes on a white background), with 50 (resp. 100) images for each digit. To each image z we randomly associate a set of d distinct points which are black (intensity superior to 190) in the image. In this case the set of components is $\{1,..,28\}\times\{1,..,28\}$ which we map onto points with coordinates between 0 and 1, thus defining $\mathcal{X} = [0,1]^2$. The linear IGV kernel as described in Section 3.2 is equivalent to using the linear kernel $\kappa((x_1,y_1),(x_2,y_2)) = x_1x_2+y_1y_2$ on a non-regularized version of the kernelized-IGV. It also boils down to fitting Gaussian bivariate-laws on the points and measuring the similarity of two measures by performing variance estimation on the samples taken first separately and then together. The resulting variances can be diagonalized to obtain three diagonal variance matrices, which can be seen as performing PCA on the sample,

$$\Sigma(\mu) = \begin{pmatrix} \Sigma_{1,1} & 0 \\ 0 & \Sigma_{2,2} \end{pmatrix}, \quad \Sigma(\mu') = \begin{pmatrix} \Sigma'_{1,1} & 0 \\ 0 & \Sigma'_{2,2} \end{pmatrix}, \qquad \Sigma(\mu'') = \begin{pmatrix} \Sigma''_{1,1} & 0 \\ 0 & \Sigma''_{2,2} \end{pmatrix}.$$

and the value of the kernel is computed through

$$k_\nu(\mu,\mu') = \frac{\sqrt{\Sigma_{1,1}\Sigma_{2,2}\,\Sigma'_{1,1}\Sigma'_{2,2}}}{\Sigma''_{1,1}\Sigma''_{2,2}}.$$

This ratio is for instance equal to 0.3820 for two handwritten digits in the case shown in Figure 2. The linear IGV manages a good discrimination between ones and zeros. Indeed, ones are shaped



$\Sigma_{1,1} = 0.0552$
$\Sigma_{2,2} = 0.0013$

$\Sigma'_{1,1} = 0.0441$
$\Sigma'_{2,2} = 0.0237$

$\Sigma''_{1,1} = 0.0497$
$\Sigma''_{2,2} = 0.0139$

Figure 2: Weighted PCA of two different measures and their mean, with their first principal component shown. Below are the variances captured by the first and second principal components, the generalized variance being the product of those two values.

as sticks, and hence usually have a strong variance carried by their first component, followed by a weak second component. On the other hand, the variance of zeros is more equally distributed between the first and second axes. When both weighted sets of points are united, the variance of the mean of both measures has an intermediary behaviour in that respect, and this suffices to discriminate numerically both images. However this strategy fails when using numbers which are not so clearly distinct in shape, or more precisely whose surface cannot be efficiently expressed in terms of Gaussian ellipsoids. To illustrate this we show in Figure 3 the Gram matrix of the linear IGV on 60 images, namely 20 zeros, 20 ones and 20 twos. Though images of ones can be efficiently discriminated from the two other digits, we clearly see that this is not the case between zeros and twos, whose support may seem similar if we try to capture them through Gaussian laws. In practice, the results obtained with the linear IGV on this particular task where so unadapted to the learning goal that the SVM's trained based on that methodology did not converge in most cases, which is why we discarded it.

8.2 Kernelized IGV

Following previous works (Kondor and Jebara, 2003, Wolf and Shashua, 2003) and as suggested in the initial discussion of Section 5, we use in this section a Gaussian kernel of width σ to incorporate a prior knowledge on the pixels, and equivalently to define the reproducing kernel Hilbert space Ξ by using

$$\kappa((x_1,y_1),(x_2,y_2)) = e^{-\frac{(x_1-x_2)^2+(y_1-y_2)^2}{2\sigma^2}}.$$

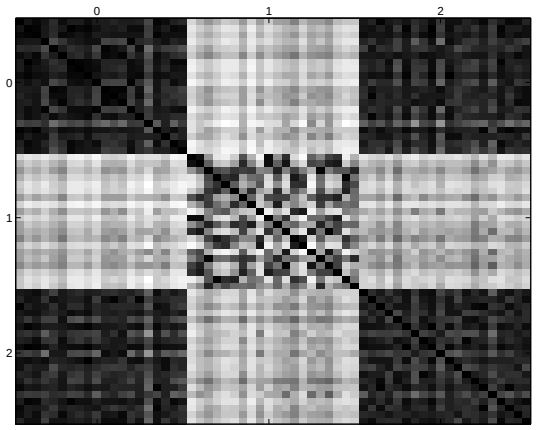


Figure 3: Normalized Gram matrix computed with the linear IGV kernel of twenty images of "0", "1" and "2" displayed in that order. Darker spots mean values closer to 1, showing that the restriction to "0" and "1" yields good separation results, while "0" and "2" can hardly be discriminated using variance analysis.

As pointed out by Kondor and Jebara (2003), the pixels are no longer seen as points but rather as functions (Gaussian bells) defined on the components space $[0,1]^2$. To illustrate this approach we show in Figure 4 the first four eigenfunctions of three measures μ_1, μ_0 and $\frac{\mu_1+\mu_0}{2}$ built from the image of a handwritten "1" and "0" with their corresponding eigenvalues, as well as for images of "2" and "0" in Figure 5.

Setting σ, the width of κ, to define the functions contained in the RKHS Ξ is not enough to fully characterize the values taken by the kernelized IGV. We further need to define η, the regularization parameter, to control the weight assigned to smaller eigenvalues in the spectrum of Gram matrices. Both parameters are strongly related, since the value of σ controls the range of the typical eigenvalues found in the spectrum of Gram matrices of admissible bases, whereas η acts as a scaling parameter for those eigenvalues as can be seen in Equation (3). Indeed, using a very small σ value, which means Ξ is only defined by peaked Gaussian bells around each pixels, yields diagonally dominant Gram matrices very close to the identity matrix. The resulting eigenvalues for $\tilde{\mathcal{K}}\Delta$ are then all very close to $\frac{1}{d}$, the inverse of the amount of considered points. On the contrary, a large value for σ yields higher values for the kernel, since all points would be similar to each other and Gram matrices would turn close to the matrix $\mathbb{1}_{d,d}$ with a single significant eigenvalue and all others close to zero. We address these issues and study the robustness of the final output of the k-IGV kernel in terms of classification error by doing preliminary experiments where both η and σ vary freely.

8.3 Experiments on the SVM Generalization Error

To study the behaviour and the robustness of the IGV kernel under different parameter settings, we used two ranges of values for η and σ:

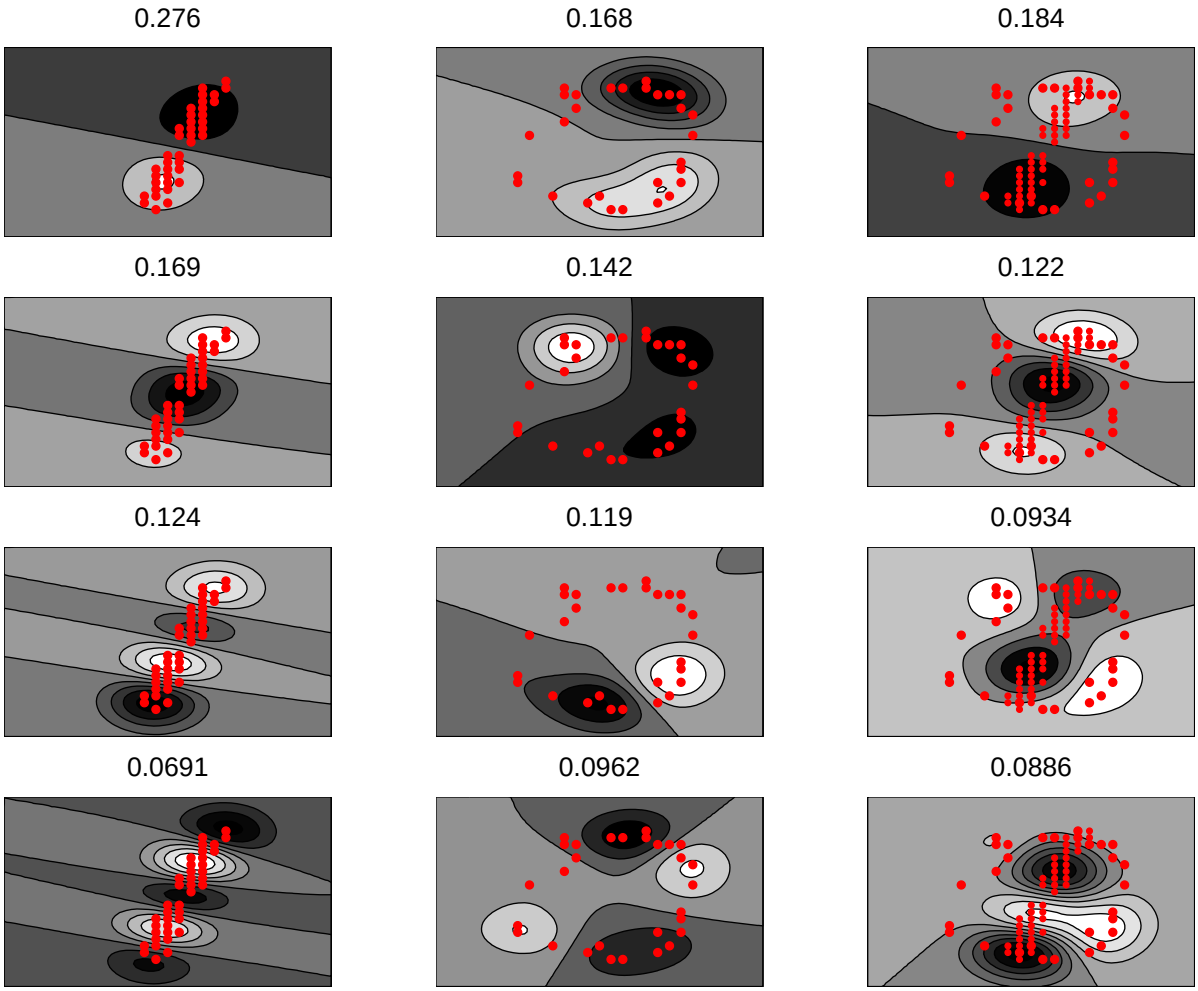


Figure 4: The four first eigenfunctions of respectively three empirical measures μ_1 (first column), μ_0 (second column) and $\frac{\mu_1+\mu_0}{2}$ (third column), displayed with their corresponding eigenvalues, using $\eta = 0.01$ and $\sigma = 0.1$.

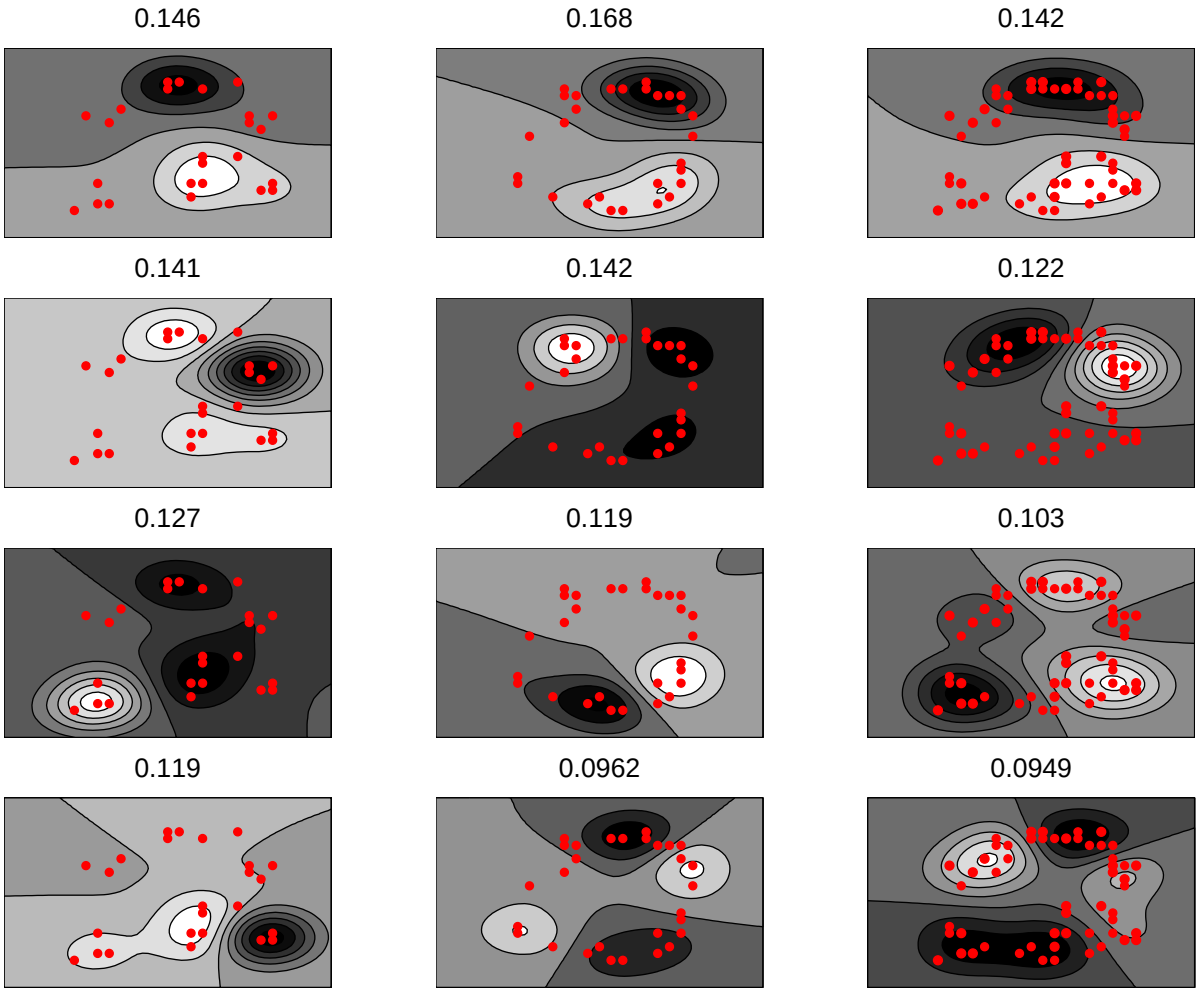


Figure 5: Same representation as in Figure 4, with μ_2, μ_0 and $\frac{\mu_2+\mu_0}{2}$.

$$\eta \in 10^{-2} \times \{0.1, 0.3, 0.5, 0.8, 1, 1.5, 2, 3, 5, 8, 10, 20\}$$
$$\sigma \in \{0.05, 0.1, 0.12, 0.15, 0.18, 0.20, 0.25, 0.3\}.$$

For each kernel k_κ^η defined by a (σ, η) couple, we trained 10 binary SVM classifiers (each one trained to recognize each digit versus all other digits) on a training fold of our 500 images dataset such that the proportion of each class was kept to be one tenth of the total size of the training set. Using then the test fold, our decision for each submitted image was determined by the highest SVM score proposed by the 10 trained binary SVM's. To determine train and test points, we led a 3-fold cross validation, namely randomly splitting our total dataset into 3 balanced subsets, using successively 2 subsets for training and the remaining one for testing (that is roughly 332 images for training and 168 for testing). The test error was not only averaged on those cross-validations folds but also on 5 different fold divisions. All the SVM experiments in this experimental section were run using the spider[1] toolbox. Most results shown here did not improve by choosing different soft margin C parameters, we hence just set $C = \infty$ as suggested by default by the authors of the toolbox.

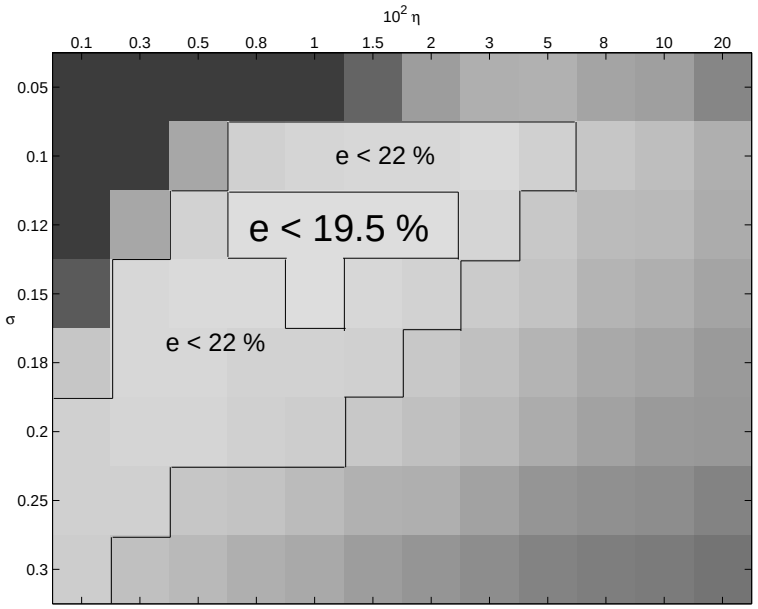


Figure 6: Average test error (displayed as a grey level) of different SVM handwritten character recognition experiments using 500 images from the MNIST database (each seen as a set of 25 to 30 randomly selected black pixels), carried out with 3-fold (2 for training, 1 for test) cross validations with 5 repeats, where parameters η (regularization) and σ (width of the Gaussian kernel) have been tuned to different values.

The error rates are graphically displayed in Figure 6 using a grey-scale plot. Note that for this benchmark the best testing errors were reached using a σ value of 0.12 with an η parameter within 0.008 and 0.02, this error being roughly 19.5%. All values below and on the right side of this zone

1. see http://www.kyb.tuebingen.mpg.de/bs/people/spider/

are below 32.5%, which is the value reached on the lower right corner. All standard deviations with respect to multiple cross-validations of those results were inferior to 2.3%, the whole region under 22% being under a standard deviation of 1%. Those preliminary tests show that the IGV kernel has an overall robust performance within what could be considered as a sound range of values for both η and σ. Note that the optimal range of parameter found in this experiment only applies to the specific sampling procedure that was used in this case (25 to 30 points), and may not be optimal for larger matrices. However the stability observed here led us to discarding further tuning of σ and η when the amount of sampled points is different. We simply applied $\sigma = 0.1$ and $\eta = 0.01$ for the remaining of the experimental section.

As in Kondor and Jebara (2003), we also compared the results obtained with the k-IGV to the standard RBF kernel performed on the images seen as binary vectors of $\{0,1\}^{28\times 28}$ further normalized so that their components sums up to 1. Using the same range for σ that was previously tested, we applied the formula $k(z,z') = e^{-\frac{\|z-z'\|}{2\sigma^2}}$. Since the RBF kernel is grounded on the exact overlapping between two images we expect it to perform poorly with few pixels and significantly better when d grows, while we expect the k-IGV to capture more quickly the structure of the images with fewer pixels through the kernel κ. This is illustrated in Figure 7 where the k-IGV outperforms significantly the RBF kernel, reaching with a sample of less than 30 points a performance the RBF kernel only reaches above 100 points. Taking roughly all black points in the images, by setting $d = 200$ for instance, the RBF kernel error is still 17.5%, an error the IGV kernel reaches with roughly 35 points.

Finally, we compared the kernelized-version of the Bhattacharrya kernel (k-B) proposed in Kondor and Jebara (2003), the k-IGV, the polynomial kernel and the RBF kernel by using a larger database of the first 1,000 images in MNIST (100 images for each of the 10 digits), selecting randomly $d = 40, 50, 60, 70$ and 80 points and performing the cross-validation methodology previously detailed. The polynomial kernel was performed seeing the images as binary vectors of $\{0,1\}^{28\times 28}$ and applying the formula $k_{b,d}(z,z') = (z\cdot z' + b)^d$. We followed the observations of Kondor and Jebara (2003) concerning parameter tuning for the k-B kernel but found out that it performed better using the same set of parameters used for the k-IGV. The results presented in Table 1 of the k-IGV kernel show a consistent improvement over all other kernels for this benchmark of 1000 images, under all sampling schemes.

We did not use the kernel described by Wolf and Shashua (2003) in our experiments because of its poor scaling properties for a large amount of considered points. Indeed, the kernel proposed by Wolf and Shashua (2003) takes the form of the product of d cosines values where d is the cardinality of the considered sets of points, thus yielding negligible values in practice when d is large as in our case. Their SVM experiments were limited to 6 or 7 points while we mostly consider lists of more than 40 points here. This problem of poor scaling which in practice produces a diagonal-dominant kernel led us to discarding this method in our comparison. All semigroup kernels presented in this paper are grounded on statistical estimation, which makes their values stable under variable sizes of samples through renormalization, a property shared with the work of Kondor and Jebara (2003). Beyond a minimal amount of points needed to perform sound estimation, the size of submitted samples influences positively the accuracy of the k-IGV kernel. A large sample size can lead however to computational problems since the value of the k-IGV-kernel requires not only the computation of the centered Gram-matrix $\mathcal{K}$ and a few matrix multiplications, but also the computation of a determinant, an operation which can quickly become prohibitive since it has a complexity of $O(d^{2.3})$ where d is the size of the considered Gram matrix. Although we did not opti-

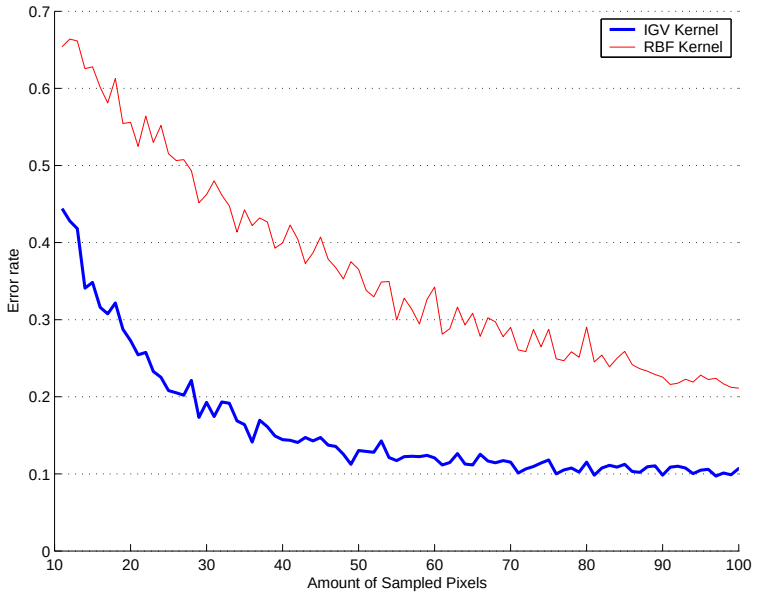


Figure 7: Average test error with RBF ($\sigma = 0.2$) and k-IGV ($\sigma = 0.1$ and $\eta = 0.01$) kernels led on 90 different samplings of 500 images. The curves show an overall trend that both kernels perform better when they are given more points to compute the similarity between two images. If we consider $d = 200$, the RBF kernel error is 0.175, that is 17.5%, a threshold the IGV kernel reaches with slightly more than 35 points. Each sampling corresponds to a different amount of sampled points d, those samplings being ordered increasingly with d. Each sampling has been performed independently which explains the bumpiness of those curves.

mize the computations of both k-B and k-IGV kernels (by storing precomputed values for instance or using numerical approximations in the computation of the determinant), this computational cost in the case of a naive implementation, illustrated by the running times displayed in Table 1, remains an issue that needs to be addressed in practical applications.

Sample Size	Gaussian $\sigma = 0.1$	Polynomial $b = 10; d = 4$	k-B $\eta = 0.01; \sigma = 0.1$	k-IGV $\eta = 0.01; \sigma = 0.1$
40 pixels	32.2 (1)	31.3 (1.5)	19.1 (1500)	16.2 (1000)
50 ”	28.5 (1)	26.3 (1.5)	17.1 (2500)	14.7 (1400)
60 ”	24.5 (1)	22.0 (1.5)	15.8 (3600)	14.6 (2400)
70 ”	22.2 (1)	19.5 (1.5)	15.1 (4100)	13.1 (2500)
80 ”	20.3 (1)	17.4 (1.5)	14.5 (5500)	12.8 (3200)

Table 1: SVM Error rate in percents of different kernels used on a benchmark test of recognizing digits images, where only 40 to 80 black points where sampled from the original images. The 1,000 images where randomly split into 3 balanced sets to perform cross validation (2 for training and 1 for testing), the error being first averaged over 5 such splits, the whole process being repeated again over 3 different random samples of points. Running times are indicated in minutes.

9. Conclusion

We presented in this work a new family of kernels between measures. Such kernels are defined through prior functions which should ideally quantify the concentration of a measure. Once such a function is properly defined, the kernel computation goes through the evaluation of the function on the two measures to be compared and on their mixture. As expected when dealing with concentration of measures, two intuitive tools grounded on information theory and probability, namely entropy and variance, prove to be useful to define such functions. Their expression is however still complex in terms of computational complexity, notably for the k-IGV kernel. Computational improvements or numerical simplifications should be brought forward to ensure a feasible implementation for large-scale tasks involving tens of thousands of objects.

An attempt to define and understand the general structure of p.d. functions on measures was also presented, through a representation as integrals of elementary functions known as semicharacters. We are investigating further theoretical properties and characterizations of both semicharacters and positive definite functions on measures. The choice of alternative priors on semicharacters to propose other meaningful kernels, with convenient properties on molecular measures for instance, is also a subject of future research. As for practical applications, these kernels can be naturally applied on complex objects seen as molecular measures. We also expect to perform further experiments to measure the performance of semigroup kernels on a diversified sample of challenging tasks, including cases where the space of components is not a vector space, notably when the considered measures are multinomials on a finite component space endowed with a kernel.

Acknowledgments

The authors would like to thank Francis Bach and Jérémie Jakubowicz for fruitful discussions, Imre Risi Kondor for sharing his code with us, anonymous reviewers for their comments which improved the quality of the paper and Xavier Dupré for his help on the MNIST database. MC acknowledges a JSPS doctoral short-term grant which made his stay in Japan possible. KF was supported by JSPS KAKENHI 15700241 and a research grant from the Inamori Foundation. JPV is supported by NHGRI NIH award R33 HG003070 and by the ACI "Nouvelles Interfaces des Mathématiques" of the French Ministry of Research. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

Appendix A : an Example of Continuous Positive Definite Function Given by Noncontinuous Semicharacters

Let $\mathcal{X}$ be the unit interval $[0,1]$ hereafter. For any t in $\mathcal{X}$, a semicharacter on $M_+^b(\mathcal{X})$ is defined by

$$\rho_{h_t}(\mu) = e^{\mu([0,t])},$$

where $h_t(x) = I_{[0,t]}(x)$ is the index function of the interval $[0,t]$. Note that ρ_{h_t} is *not* continuous for $t \in [0,1)$ by Proposition 11.

For $\mu \in M_+^b(\mathcal{X})$, the function $t \mapsto \mu([0,t))$ is bounded and non-decreasing, thus, Borel-measurable, since the discontinuous points are countable at most. A positive definite function on $M_+^b(\mathcal{X})$ is defined by

$$\varphi(\mu) = \int_0^1 \rho_{h_t}(\mu) dt.$$

This function is continuous, while it is given by the integral of noncontinuous semicharacters.

Proposition *The positive definite function* φ *is continuous and exponentially bounded.*

Proof Suppose μ_n converges to μ weakly in $M_+^b(\mathcal{X})$. We write $F_n(t) = \mu_n([0,t])$ and $F(t) = \mu([0,t])$. Because μ_n and μ are finite measures, the weak convergence means

$$F_n(t) \quad \to \quad F(t)$$

for any continuous point of F. Since the set of discontinuous points of F is at most countable, the above convergence holds almost everywhere on X with Lebesgue measure. From the weak convergence, we have $F_n(1) \to F(1)$, which means there exists $M > 0$ such that $\sup_{t \in \mathcal{X}, n \in \mathbb{N}} F_n(t) < M$. By the bounded convergence theorem, we obtain

$$\lim_{n \to \infty} \varphi(\mu_n) = \lim_{n \to \infty} \int_0^1 e^{F_n(t)} dt = \int_0^1 e^{F(t)} dt = \varphi(\mu).$$

For the exponential boundedness, by taking an absolute value $\alpha(\mu) = e^{\mu(X)}$, we have

$$|\varphi(\mu)| \leq \int_0^1 \alpha(\mu) dt = \alpha(\mu).$$

■

References

Shotaro Akaho. A kernel method for canonical correlation analysis. In *Proceedings of International Meeting on Psychometric Society (IMPS2001)*, 2001.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. AMS vol. 191, 2001.

Francis Bach and Michael Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic Analysis on Semigroups*. Springer-Verlag, 1984.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2003.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th annual ACM workshop on Computational Learning Theory*, pages 144–152. ACM Press, 1992.

Marco Cuturi and Jean-Philippe Vert. The context-tree kernel for strings. *Neural Networks*, 2005. In press.

Marco Cuturi and Jean-Philippe Vert. Semigroup kernels on finite sets. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 329–336. MIT Press, Cambridge, MA, 2005.

Jean Dieudonné. *Calcul Infinitésimal*. Hermann, Paris, 1968.

Dominik M. Endres and Johannes E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860, 2003.

Bent Fuglede and Flemming Topsøe. Jensen-shannon divergence and hilbert space embedding. In *Proc. of the Internat. Symposium on Information Theory*, page 31, 2004.

Kenji Fukumizu, Francis Bach, and Michael Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability measures. January 2005.

Tony Jebara, Risi Kondor, and Andrew Howard. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844, 2004.

Thorsten Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms*. Kluwer Academic Publishers, Dordrecht, 2002. ISBN 0-7923-7679-X.

Risi Kondor and Tony Jebara. A kernel between sets of vectors. In *Proceedings of the International Conference on Machine Learning*, 2003.

John Lafferty and Guy Lebanon. Information diffusion kernels. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.

Christina Leslie, Eleazar Eskin, and William Stafford Noble. The spectrum kernel: a string kernel for svm protein classific ation. In *Proceedings of the Pacific Symposium on Biocomputing 2002*, pages 564–575. World Scientific, 2002.

Christina Leslie, Eleazar Eskin, Jason Weston, and William Stafford Noble. Mismatch string kernels for svm protein classification. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *Advances in Neural Information Processing Systems 15*, Cambridge, MA, 2003. MIT Press.

Thomas Melzer, Michael Reiter, and Horst Bischof. Nonlinear feature extraction using generalized canonical correlation analysis. In *Proceedings of International Conference on Artificial Neural Networks (ICANN)*, pages 353–360, 2001.

Pedro J. Moreno, Purdy P. Ho, and Nuno Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

Ferdinand Österreicher and Igor Vajda. A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics*, 55:639–653, 2003.

C. R. Rao. Differential metrics in probability spaces. In S.-I. Amari, O.E. Barndorff-Nielsen, R.E. Kass, S.L. Lauritzen, and C.R. Rao, editors, *Differential Geometry in Statistical Inference*, Hayward,CA, 1987. Institute of Mathematical Statistics.

Walter Rudin. *Fourier Analysis on Groups*. John Wiley & sons, 1962.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization , Optimization, and Beyond*. MIT Press, Cambridge, MA, 2002.

Matthias Seeger. Covariance kernels from bayesian generative models. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 905–912, Cambridge, MA, 2002. MIT Press.

F. M. J. Willems, Y. M. Shtarkov, and Tj. J. Tjalkens. The context-tree weighting method: basic properties. *IEEE Transactions on Information Theory*, pages 653–664, 1995.

Lior Wolf and Amnon Shashua. Learning over sets using kernel principal angles. *Journal of Machine Learning Research*, 4:913–931, 2003.