Convolution Kernels on Discrete Structures
UCSC-CRL-99-10

David Haussler
Department of Computer Science
University of California at Santa Cruz
Santa Cruz, CA 95064
email: haussler@cse.ucsc.edu
URL: http://www.cse.ucsc.edu/ haussler

July 8, 1999

Abstract

We introduce a new method of constructing kernels on sets whose elements are discrete structures like strings, trees and graphs. The method can be applied iteratively to build a kernel on a infinite set from kernels involving generators of the set. The family of kernels generated generalizes the family of radial basis kernels. It can also be used to define kernels in the form of joint Gibbs probability distributions. Kernels can be built from hidden Markov random fields, generalized regular expressions, pair-HMMs, or ANOVA decompositions. Uses of the method lead to open problems involving the theory of infinitely divisible positive definite functions. Fundamentals of this theory and the theory of reproducing kernel Hilbert spaces are reviewed and applied in establishing the validity of the method.

1 Introduction

Many problems in statistics and pattern recognition demand that discrete structures likes strings, trees, and graphs be classified or clustered based on

similarity. To do this, it is desirable to have a method to extract real-valued features $\phi_1(x), \phi_2(x), \ldots$ from any structure x in a class X of discrete structures. If finitely many features are extracted, the feature extraction process can be represented by a mapping from X into d-dimensional Euclidean space $\Re^d$, and if infinitely many features are extracted, by a mapping into the Hilbert space of all square-summable sequences, l_2. In the latter case we get an infinite series representation of x, much like the Fourier series representation of a function in L_2. Here we present some methods for defining series representations for discrete structures using a general type of kernel function we call a *convolution kernel*.

Because we use a kernel formulation, the series representations we develop are implicit. That is, rather than giving an explicit formula for $\{\phi_n(x)\}_{n\geq 1}$, we will give a formula for the inner product, or *kernel*, $K(x,y) = \sum_n \phi_n(x)\phi_n(y)$ that can be computed for any structures $x, y \in X$. Of course, this serves to define a distance between structures x and y in the standard manner: $d(x,y) = \sqrt{K(x,x) - 2K(x,y) + K(y,y)}$. We discuss other methods of defining a distance from a kernel as well. There is considerable recent work showing that most standard classification, clustering and regression methods can be "kernelized", that is, they can be accomplished without ever explicitly representing the feature vector $\{\phi_n(x)\}_{n\geq 1}$, relying instead only on indirect computations of the kernel $K(x,y)$ or the distance $d(x,y)$ [28, 31, 2, 13, 23, 30, 17] (see also the bibliography at http://svm.first.gmd.de.) The kernels and corresponding distance functions we construct are suitable for all such methods. In particular, there is a 1-1 correspondence between kernels and Gaussian processes defined on the set X [3, 32, 21]. We do not pursue this avenue in this paper, but the kernels we develop can be plugged directly into Gaussian process methods.

Convolution kernels are obtained from other kernels by a certain sum over products that can be viewed as a generalized convolution (Section 2). That it is possible to construct kernels in this way follows from some simple closure properties of the class of positive definite functions, which are the abstract embodiments of kernels (Sections 2 and 7). Convolution kernels generalize the classes of radial basis and simple exponential kernels (Section 2.3) and the class of ANOVA kernels (Section 2.4). They can be used to represent joint probability distributions on pairs of structures from a set X (Section 3). Classical methods of using convolutions and generating functions to analyse

discrete distributions can be extended to convolution kernels (Section 3.2). By normalizing a positive convolution kernel we obtain a probability distribution on $X \times X$ that we call a *Gibbs kernel* (Section 3.5). These kernels may have promising applications in areas where structures can be modeled generatively by Hidden Markov Random Fields (HMRFs) [12, 18, 5].

Convolution kernels can be applied iteratively to build a kernel on a infinite set from kernels involving generators of the set. We introduce a class of generalized regular expressions to define kernels in this manner (Section 4). We give an example by developing a kernel on finite strings that describes the relationship between two strings that are derived from a common ancestor under the operations of insertion, deletion and substitution of letters (Section 4.4). This and similar kernels are related to the pair-HMMs defined in [4]. This provides a new angle on the old field of syntactic pattern recognition, developed by Kung-Sun Fu and his colleagues [9, 10, 11].

Attempts to control the "width" parameter in generalized radial basis kernels derived from convolution kernels lead us to the important notion of infinitely divisible kernels, which we review (Section 6). Some open problems are mentioned in this regard. We also review the theory of reproducing kernel Hilbert spaces [22, 32, 33] (Section 7), and use it to derive several results mentioned in earlier sections.

2 Convolution kernels

2.1 Kernels

Let X be a set and $K : X \times X \to \Re$, where $\Re$ denotes the real numbers[1] and $\times$ denotes set product. We say K is a *kernel on* $X \times X$ if K is symmetric, i.e. for any x and $y \in X$, $K(x, y) = K(y, x)$, and K is *positive definite*, in the sense that for any $N \geq 1$ and any $x_1, \ldots, x_N \in X$, the matrix K defined by $K_{ij} = K(x_i, x_j)$ is positive definite, i.e. $\sum_{ij} c_i c_j K_{ij} \geq 0$ for all $c_1, \ldots, c_N \in \Re$. Equivalently, a symmetric matrix is positive definite if all its eigenvalues are nonnegative, see, e.g. [29].

[1] Many authors consider the more general case of complex-valued kernels. The relationship between the definitions used for that case and the ones used here for the real case is discussed in [1], section 1.6, page 68. Virtually all of the results extend naturally to the complex case.

It is readily verified that if each $x \in X$ is represented by the sequence[2] $\phi(x) = \{\phi_n(x)\}_{n\geq 1}$ such that $K : X \times X \to \Re$ is the l_2 inner product $K(x,y) = \sum_n \phi_n(x)\phi_n(y) = \langle \phi(x), \phi(y) \rangle$, then K is a kernel, because for any $x_1, \ldots, x_N \in X$ and $c_1, \ldots, c_N \in \Re$,

$$\begin{aligned} \sum_{i,j=1}^{N} c_i c_j K(x_i, x_j) &= \sum_{i,j=1}^{N} c_i c_j \langle \phi(x_i), \phi(x_j) \rangle \\ &= \left\langle \sum_{i=1}^{N} c_i \phi(x_i), \sum_{j=1}^{N} c_j \phi(x_j) \right\rangle \\ &= \left\langle \sum_{i=1}^{N} c_i \phi(x_i), \sum_{i=1}^{N} c_i \phi(x_i) \right\rangle \geq 0 \end{aligned} \tag{1}$$

(see Equation (24) in Section 7.)

It turns out that under reasonable assumptions on X and K, which nearly always hold in practice, any kernel K can be represented as $K(x,y) = \sum_n \phi_n(x)\phi_n(y)$ for some choice of functions $\{\phi_n\}$ [22]. We give a proof of this in Section 7 (Theorem 5). In particular, this is true for all kernels on $X \times X$ for a countable set X, and more generally, it is true whenever X is a separable metric space and K is a continuous function on $X \times X$ (see Section 7.) Thus, in some sense, choosing a kernel on $X \times X$ is the same a choosing a series $\phi(x)$ in l_2 of "feature values" to represent each $x \in X$.

The class of kernels on a set $X \times X$ has wonderful closure properties that can be used to great advantage. In particular, it is readily verified that this class is closed under addition, multiplication by a positive constant and pointwise limits (see e.g. [1]). Hence they form a closed convex cone [1]. It is also well-known that the class is closed under product, i.e. if $K_1(x,y)$ and $K_2(x,y)$ are kernels, then $K(x,y) = K_1(x,y)K_2(x,y)$ is a kernel. This is equivalent to the fact that positive definite matrices are closed under Schur product, i.e. element-wise product $[A \cdot B]_{i,j} = A_{i,j}B_{i,j}$. (see e.g. [1], Theorem 1.12, page 69).

Because kernels are closed under product, it is easy to see that they are also closed under *tensor product*, i.e. if $K_1(x,y)$ is a kernel on $X \times X$ and $K_2(u,v)$ is a kernel on $U \times U$ then $K_1 \otimes K_2((x,u),(y,v)) = K_1(x,y)K_2(u,v)$ is a kernel on $(X \times U) \times (X \times U)$ ([1], Corollary 1.13, page 70). Similarly,

[2]Note that since $\sum_n \phi_n^2(x) = K(x,x) < \infty$, $\phi(x) \in l^2$ for all x.

since they are closed under sum, they are also closed under *direct sum*, i.e. $K_1 \oplus K_2((x,u),(y,v)) = K_1(x,y) + K_2(u,v)$ is a kernel on $(X \times U) \times (X \times U)$. Going in the other direction, if $K((x,u),(y,v))$ is a kernel on $(X \times X) \times (X \times X)$, then the *diagonal projection* $K^{\Delta}(x,y) = K((x,x),(y,y))$ is a kernel on $X \times X$. It is clear that $(K_1 \otimes K_2)^{\Delta} = K_1 K_2$ and $(K_1 \oplus K_2)^{\Delta} = K_1 + K_2$.

Lastly, it is easy to see that if $S \subseteq X$ and K is a kernel on $S \times S$, then K may be extended to a kernel on $X \times X$ by defining $K(x,y) = 0$ if either x or y is not in S. This follows directly from the definition of a positive definite function. We call this the *zero extension* of K.

2.2 R-Convolution kernels

Suppose $x \in X$ is a composite structure and $x_1, \ldots, x_D$ are its "parts", where x_d is in the set X_d for each $1 \leq d \leq D$, and D is a positive integer. Throughout this paper we assume that $X, X_1, \ldots, X_D$ are nonempty, separable metric spaces. This includes the special case that $X, X_1, \ldots, X_D$ are countable sets (see Section 7.) This countable case is the primary focus of the paper.

We can represent the relation "$x_1, \ldots, x_d$ are the parts of x" by a relation R on the set $X_1 \times \cdots \times X_D \times X$, where $R(x_1, \ldots, x_D, x)$ is true iff $x_1, \ldots, x_D$ are the parts of x. For brevity, let $\vec{x} = x_1, \ldots, x_D$, and denote $R(x_1, \ldots, x_D, x)$ by $R(\vec{x}, x)$. Let $R^{-1}(x) = \{\vec{x} : R(\vec{x}, x)\}$. We say R is *finite* if $R^{-1}(x)$ is finite for all $x \in X$. Here are some examples:

1. If x is a D-tuple in $X = X_1 \times \cdots \times X_D$, and each component of $x \in X$ is a part of x, then $R(\vec{x}, x)$ iff $\vec{x} = x$.

2. If $X_1 = X_2 = X$, where X is the set of all finite strings over a finite alphabet $\mathcal{A}$, then we can define $R(x_1, x_2, x)$ iff $x_1 \circ x_2 = x$, where $x_1 \circ x_2$ denotes the concatenation of strings x_1 and x_2.

3. Continuing the previous example, if the alphabet $\mathcal{A}$ has only one letter, then a finite string can be represented by the nonnegative integer n that is its length, so $X_1 = X_2 = X = \{0, 1, \ldots\}$ and $R(n_1, n_2, n)$ iff $n_1 + n_2 = n$.

4. If $X_1 = \ldots = X_D = X$, where X is the set of all D-degree ordered and rooted trees, then we can define $R(\vec{x}, x)$ iff $x_1, \ldots, x_D$ are the D subtrees of the root of the tree $x \in X$.

Note that examples 2 and 3 show it is possible that a given object x may be decomposable into parts in multiple ways. Examples 2-4 demonstrate how the relation between part and structure can be used iteratively to define more complex structures in X when $X_1 = \cdots = X_D = X$ for an infinite set X.

Suppose $x, y \in X$ and for some decompositions of x and y, $\vec{x} = x_1, \ldots, x_D$ are the parts of x, and $\vec{y} = y_1, \ldots, y_D$ are the parts of y. Suppose further that for each $1 \leq d \leq D$, we have a kernel K_d on X_d that we can use to measure the similarity $K_d(x_d, y_d)$ between the part x_d and the part y_d. If X_d is uncountable, then we assume K_d is continuous. Then we define the similarity $K(x, y)$ between x and y as the following generalized convolution

$$K(x,y) = \sum_{\vec{x} \in R^{-1}(x), \vec{y} \in R^{-1}(y)} \prod_{d=1}^{D} K_d(x_d, y_d) \tag{2}$$

This defines a symmetric function on $S \times S$, where $S = \{x : R^{-1}(x) \text{ is not empty }\}$. We define *R-convolution* of $K_1, \ldots, K_D$, denoted $K_1 \star \cdots \star K_D(x, y)$, to be the zero extension of K to $X \times X$. We refer to K as a *finite convolution* if R is finite.

Theorem 1 *If $K_1, \ldots, K_D$ are kernels on $X_1 \times X_1, \ldots, X_D \times X_D$, respectively, and R is a finite relation on $X_1 \times \cdots \times X_D \times X$, then $K_1 \star \ldots \star K_D$ is a kernel on $X \times X$.*

To prove this theorem we need

Lemma 1 *Let K be a kernel on a set $U \times U$ and for all finite, nonempty $A, B \subseteq U$ define $K'(A, B) = \sum_{x \in A, y \in B} K(x, y)$. Then K' is a kernel on the product of the set of all finite, nonempty subsets of U with itself.*

The proof of this lemma is given in Section 7.

Proof of the theorem: Let U denote $X_1 \times \cdots \times X_D$. Since $K_1, \ldots, K_D$ are kernels by assumption, it is clear from the closure of kernels under tensor product that

$$\tilde{K}(\vec{x}, \vec{y}) = \prod_{d=1}^{D} K_d(x_d, y_d)$$

is a kernel on $U \times U$.

Since R is finite, by Lemma 1, $\tilde{K}'(R^{-1}(x), R^{-1}(y))$ is a kernel on the product of the set of all nonempty $R^{-1}(x)$ such that $x \in X$ with itself. Since

$K_1 \star \cdots \star K_D(x, y)$ is the zero extension of $K(x, y) = \tilde{K}'(R^{-1}(x), R^{-1}(y))$, it follows that it is a kernel on $X \times X$. $\square$

In the case that X is uncountable, in what follows we assume it is a separable metric space with a metric defined such that K is a continuous function.

2.3 Example: radial basis and simple exponential kernels

In example 1 in Section 2, because there is only one way to decompose each x, the R-convolution kernel reduces to

$$K_1 \star \cdots \star K_D(x, y) = \prod_{d=1}^{D} K_d(x_d, y_d).$$

For each $1 \leq d \leq D$, let $f_d : X_d \to \Re$, $\sigma_d > 0$, and

$$K_d(x, y) = e^{-(f_d(x) - f_d(y))^2 / 2\sigma_d^2}.$$

It is well-known that K_d is a kernel (see, e.g., [1], Section 1.10 on page 69 and Theorem 2.2 on page 74). Then

$$K_1 \star \cdots \star K_D(x, y) = e^{-\sum_{d=1}^{D} (f_d(x_d) - f_d(y_d))^2 / 2\sigma_d^2}. \tag{3}$$

Kernels of this form are called *radial basis kernels* [28]. In radial basis kernels, each function f_d is used to extract a primitive real-valued feature from the component x_d of x. These features are then used to define a kernel K that in fact maps x implicitly into an infinite dimensional feature space. Such kernels have proven quite useful in practice [27].

Continuing with Example 1 from Section 2, using the same primitive features $\{f_d(x_d) : 1 \leq d \leq D\}$, we can define the *simple exponential kernel*

$$K_1 \star \cdots \star K_D(x, y) = e^{\sum_{d=1}^{D} f_d(x_d) f_d(y_d) / \sigma_d^2} = \prod_{d=1}^{D} K_d(x_d, y_d),$$

where here

$$K_d(x, y) = e^{f_d(x) f_d(y) / \sigma_d^2},$$

which is also a kernel for any real-valued function f_d (see, e.g., [1], Corollary 1.14, page 70). This is closely related to the radial basis kernel defined above.

Indeed, if $\tilde{K}$ is the radial basis kernel above and K is the simple exponential kernel, then it is easily verified that

$$\tilde{K}(x,y) = \frac{K(x,y)}{\sqrt{K(x,x)}\sqrt{K(y,y)}} \tag{4}$$

We will introduce a more general version of the radial basis kernels in Section 5 below.

2.4 Example: ANOVA kernels

Here is a quite different type of R-convolution kernel that is used in practice and called an *analysis of variance* (ANOVA) kernel [31, 30]. Let $X = S^n$ for some set S and $K^{(i)}$ be a kernel on $S \times S$ for each $1 \leq i \leq n$. For $1 \leq D \leq n$, the ANOVA kernel of order D is defined by[3]

$$K(x,y) = \sum_{1 \leq i_1 < \cdots < i_D \leq n} \prod_{d=1}^{D} K^{(i_d)}(x_{i_d}, y_{i_d}).$$

For each $1 \leq d \leq D$, let $X_d = S \times \{1, \ldots, n\}$, and let $\tilde{K}((s,i),(t,j)) = K^{(i)}(s,t)$ if $i = j$ and 0 else. It is readily verified that $\tilde{K}$ is a kernel if the $K^{(i)}$ are. Let $K_d = \tilde{K}$ for all $1 \leq d \leq D$. Define $R((s_1,i_1), \ldots, (s_D,i_D), x)$ iff $s_d = x_{i_d}$ for $1 \leq d \leq D$ and $i_1 < \cdots < i_D$. Since the cardinality of $R^{-1}(x)$ is $\binom{n}{D}$, R is finite. Clearly $K(x,y)$ is the R-convolution of $K_1, \ldots, K_D$.

If $D = n$, then it is clear that $K = K^{(1)} \otimes \cdots \otimes K^{(n)}$. At the other extreme, if $D = 1$, then $K = K^{(1)} \oplus \cdots \oplus K^{(n)}$. Thus by playing with the definition of the "parts of" relation R in ANOVA kernels, we get a spectrum of kernels from direct sum to tensor product. We can play further with this definition to get a spectrum from (normal) sums to products. Define everything the same as above, except let $X = S$, and $R((s_1,i_1), \ldots, (s_D,i_D), x)$ iff $s_d = x$ for $1 \leq d \leq D$ and $i_1 < \cdots < i_D$. We call the resulting kernel K a *diagonal projection ANOVA kernel*. For such a kernel, it is easily verified that if $D = n$ then

$$K = (K^{(1)} \otimes \cdots \otimes K^{(n)})^{\Delta} = K^{(1)} \cdots K^{(n)}$$

[3]Typically all $K^{(i)}$ are the same and the superscript is dropped.

and if $D = 1$

$$K = (K^{(1)} \oplus \cdots \oplus K^{(n)})^{\Delta} = K^{(1)} + \cdots + K^{(n)}.$$

Hence R convolutions generalize both products and sums of kernels. Interesting variations on the radial basis and simple exponential kernels discussed in the previous sections are possible using diagonal projection ANOVA kernels in place of the simple products used there. Although these kernels can have an exponential number of terms, e.g. when $D = n/2$, there is a recursive formula that allows them to be computed efficiently, which is critical for their practical utility [31].

3 P-Kernels

We say a kernel K is *positive* if $K(x, y) \geq 0$ for all x, y. If K is a positive kernel and $\sum_{x,y} K(x, y) = 1$, then K is a probability distribution on $X \times X$, and is called a *P-kernel*.

3.1 Closure properties

The class of positive kernels is closed under addition, multiplication, multiplication by a positive scalar, pointwise limits, and R-convolution with a finite relation R. These closure properties are clear: since we already know that each closure property holds for the class of all kernels, to verify that they hold for the class of positive kernels, it suffices to notice that they preserve positivity.

Let us say that the relation R is *is a function* if for every $\vec{x}$ there is one x such that $R(\vec{x}, x)$. In Examples 1-4 in Section 2, the relation R is a function.

Theorem 2 *The class of P-kernels is closed under convex combination and R-convolution for a finite function R.*

Proof: The closure of P-kernels under convex combination is clear for finite convex combinations, since kernels are closed under addition and multiplication by a positive constant, and convex combination preserves the property that the kernel is positive and sums to 1. To verify this closure property for infinite convex combinations of the form $K(x, y) = \sum_n p_n K_n(x, y)$, where

each K_n is a P-kernel, $p_n > 0$, and $\sum_n p_n = 1$, we can additionally use the closure of kernels under pointwise limits, since $0 \leq K_n(x, y) \leq 1$ for all n, x, y.

To see that the second closure property holds, assume the kernel K_d is a probability distribution on $X_d \times X_d$ for $1 \leq d \leq D$ and let Q be the product distribution on $X_1 \times \cdots \times X_D \times X_1 \times \cdots \times X_D$ defined by $Q(\vec{x}, \vec{y}) = \prod_{d=1}^{D} K_d(x_d, y_d)$. The R-convolution K is the zero extension of the image of this distribution under the function R, and hence is a probability distribution on $X \times X$. $\square$

3.2 Simple R-convolutions and generating functions

For each $1 \leq d \leq D$, let $g_d : X_d \to \Re$. We say that $g : X \to \Re$ is the *simple R-convolution of* $g_1, \ldots, g_D$ if

$$g(x) = \sum_{\vec{x} \in R^{-1}(x)} \prod_{d=1}^{D} g_d(x_d) \tag{5}$$

whenever $R^{-1}(x)$ is not empty and $g(x) = 0$ otherwise. We denote this convolution by $g = g_1 \star \cdots \star g_D$.

As demonstrated above for P-kernels, it is easily verified that if each g_d is a probability distribution on X and R is a function, then $g_1 \star \cdots \star g_D$ is a probability distribution on X. Thus, since the class of probability distributions on X is also clearly closed under convex combinations, it has the same closure properties as those given for P-kernels in Theorem 2, but using simple convolutions.

Simple convolutions of probability distributions are illustrated by classical convolutions of discrete random variables. As in Example 3 of Section 2, let $D = 2$, $X_1 = X_2 = X = \{0, 1, \ldots\}$ and $R(n_1, n_2, n)$ iff $n_1 + n_2 = n$. If $\mathbf{X}$ is a random variable taking values in X with distribution $g(n) = P(\mathbf{X} = n)$, and $\mathbf{Y}$ is a random variable with distribution $h(n) = P(\mathbf{Y} = n)$, then

$$g \star h(n) = P(\mathbf{X} + \mathbf{Y} = n).$$

The *generating function* for $\mathbf{X}$ is defined by $G(s) = \sum_{n=0}^{\infty} g(n) s^n$ where s is a formal variable, and similarly, the generating function for $\mathbf{Y}$ may be defined by $H(s) = \sum_{n=0}^{\infty} h(n) s^n$. Then the generating function for $\mathbf{X} + \mathbf{Y}$ is clearly $G(s)H(s)$. So convolution of distributions with this relation R corresponds

to multiplication of generating functions. By differentiating the generating function, one obtains the moments of the distribution (see, e.g., [6]).

Other kinds of convolutions can be used to represent combinatorial counting problems, because if $g_d(x_d) = 1$ for all x_d, then $g_1 \star \cdots \star g_D(x)$ is the cardinality of $R^{-1}(x)$. As an example, let $D = 2$, $X_1 = X_2 = X = \{1, 2, \ldots\}$, $R(n_1, n_2, n)$ iff $n = n_1 n_2$, and $g_1(n) = g_2(n) = 1$ for all n. Then $g_1 \star g_2(n) = n + 1 - \phi(n)$, where $\phi(n)$ is Euler's totient function, which counts the number of non negative integers less than n that are relatively prime to n.

3.3 Independent and diagonal kernels

We say the kernel K is *independent* if there is a function $g : X \to \Re$ such that $K(x, y) = g(x)g(y)$. It is clear that if K is an independent P-kernel, then K is a product of two independent and identical distributions on X.

Convolutions of independent kernels decompose into an independent kernel consisting of the product of two simple convolutions:
if $K_d(x_d, y_d) = g_d(x_d)g_d(y_d)$ for all $1 \leq d \leq D$, then

$$K(x, y) = K_1 \star \cdots \star K_D(x, y) = (g_1 \star \cdots \star g_D(x)) (g_1 \star \cdots \star g_D(y)) . \quad (6)$$

Hence convolutions of independent P-kernels under a function R are again independent P-kernels.

We say the kernel K is *diagonal* if there is a (necessarily positive) function g such that $K(x, y) = g(x)\delta(x, y)$, where the δ function is defined by $\delta(x, y) = 0$ if $x \neq y$ and $\delta(x, x) = 1$. The *identity kernel* is the diagonal kernel $K(x, y) = \delta(x, y)$. For a function R, a convolution of diagonal kernels is a diagonal kernel of simple convolutions:
if $K_d(x_d, y_d) = g_d(x_d)\delta(x_d, y_d)$, then

$$K(x, y) = K_1 \star \cdots \star K_D(x, y) = g_1 \star \cdots \star g_D(x)\delta(x, y). \quad (7)$$

Hence convolutions of diagonal P-kernels under a function R are again diagonal P-kernels.

3.4 Positive simple convolutions, Gibbs distributions, and hidden Markov random fields

We say g is *positive* if $g(x) \geq 0$ for all x. Like the class of positive kernels, the class of positive functions is closed under addition, multiplication, multi-

plication by a positive scalar, pointwise limits, and simple R-convolution for any relation R.

Suppose that $Z = \sum_{x \in X} g(x)$ is finite and nonzero. Then we may normalize g to a probability distribution

$$P(x) = \frac{g(x)}{Z} \tag{8}$$

If g is a simple R convolution, then we call $P(x)$ the *Gibbs distribution* for this R convolution, an refer to it as an *R-Gibbs distribution*. We call Z the *partition function*. The central example is a finite Markov random field, or more generally, a finite Markov random field with latent variables, which we will call a *Hidden Markov Random Field (HMRF)*.

A finite HMRF is defined as joint distribution on a finite set of visible random variables $V_1, \ldots, V_n$ and a finite set of unobserved (hidden, latent) random variables $U_1, \ldots, U_m$. We assume here that these variables have a finite range. The HMRF is defined in terms of the cliques of a graph on $n + m$ vertices representing these variables, along with auxiliary functions associated with these cliques. We give a brief definition here; details can be found in [18, 5].

For each $1 \leq d \leq D$, let C_d be a distinct subset of $\{V_1, \ldots, V_n\} \cup \{U_1, \ldots, U_m\}$. These will be the *cliques*. We assume all variables are contained in at least one clique. Let X_d denote the set of all possible (joint) assignments to the variables in C_d. We call X_d the *variable assignment set* for C_d. Let $U = (U_1, \ldots, U_m)$ and $V = (V_1, \ldots, V_n)$, and u, v denote assignments to the random vectors U and V respectively. Let $u^{(d)}$ denote the joint assignment u restricted to the variables in $C_d \cap \{U_1, \ldots, U_m\}$, and similarly for $v^{(d)}$. Thus $(u^{(d)}, v^{(d)}) \in X_d$ denotes the assignment to the variables in the clique C_d induced by the global joint assignments u and v.

For each $1 \leq d \leq D$, let $h_d : X_d \to \Re$ be a positive function. We call h_d the *compatibility function* for the clique C_d. Let

$$h(v) = \sum_u \prod_{d=1}^{D} h_d(u^{(d)}, v^{(d)}) \tag{9}$$

Then this defines a HMRF with Gibbs distribution on the visible variables V given by

$$P(v) = \frac{h(v)}{Z} \tag{10}$$

where

$$Z = \sum_v h(v).$$

Let X be the set of all assignments to V. Define $R(\vec{x}, x)$ iff there exist assignments u and v such that $x = v$ and $x_d = (u^{(d)}, v^{(d)})$ for all $1 \leq d \leq D$. We call R the *assignment checking relation* for the HMRF. Since the variables have a finite range, R is a finite relation. Then it is clear that

$$h(v) = h_1 \star \cdots \star h_D(v),$$

and hence $P(v)$ is the Gibbs distribution defined from the convolution of $h_1, \ldots, h_D$ under R.

3.5 Gibbs kernels

If K is a positive kernel and $Z = \sum_{x,y \in X} K(x, y)$ is finite and nonzero, then we may normalize K to a P-kernel

$$P(x, y) = \frac{K(x, y)}{Z} \tag{11}$$

If K is an R-convolution, then we call P an *R-Gibbs kernel*. For example, radial basis, simple exponential and positive ANOVA kernels can be used to generate Gibbs kernels in this way. More interesting is to use a hidden Markov random field.

Assume we have an HMRF with D cliques as in Section 3.4. Let X be the set of assignments to the observed variables, $X_1, \ldots, X_D$ be the variable assignment sets for the cliques, R be the assignment checking relation, $h_1, \ldots, h_D$ be the compatibility functions for the cliques, and $h = h_1 \star \cdots \star h_D$. Then, as given in Equation (10), the probability distribution defined by the HMRF is $P(x) = h(x)/Z$. Let us set $K_d(x_d, y_d) = h_d(x_d)h_d(y_d)$. Then from Equation (6) it follows that

$$K(x, y) = K_1 \star \times \star K_D(x, y) = h(x)h(y),$$

thus K is an independent positive kernel. Clearly the Gibbs kernel

$$P(x, y) = P(x)P(y).$$

So in this case the Gibbs kernel is just the product of two independent copies of the Gibbs distribution.

Alternatively, we can define $K_d(x_d, y_d) = h_d(x_d)\delta(x_d, y_d)$, obtaining the diagonal kernel

$$K(x, y) = K_1 \star \cdots \star K_D(x, y) = h(x)\delta(x, y)$$

using Equation (7). In this case

$$P(x, y) = P(x)\delta(x, y),$$

i.e. the HMRF appears on the diagonal of the Gibbs kernel P. Here the assignments x and y to the observed variables are completely correlated.

Now let $K_1, \ldots, K_D$ be any positive kernels on the clique variable assignments $X_1, \ldots, X_D$ of the HMRF. These kernels replace the compatibility functions $h_1, \ldots, h_D$, and can be defined, e.g. by convex combination, so that they interpolate between the two extremes above. Let

$$P(x, y) = \frac{K_1 \star \cdots \star K_D(x, y)}{Z}.$$

The Gibbs kernel P models a dependency between two assignments x and y to the visible variables of the HMRF. If the variables in V represent parts of an observed structure, then this Gibbs kernel provides a way of using the generative probability model inherent in a HMRF to define a notion of similarity between related structures. This idea will be further developed in a separate paper. (See also [16, 15] for an alternate way to do this.)

4 Iterated convolution kernels and generalized regular expressions

When X is countably infinite and $X_d = X$ for $1 \leq d \leq D$, as in the examples of kernels for strings and natural numbers given in Section 2 above, it is very useful to be able to build more complex kernels from simpler kernels using the

closure properties of kernels. This is most conveniently done for P-kernels; hence we restrict ourselves to that case in this section. Accordingly, we will assume that the relation R is a finite function, so that we may exploit both the closure under convex combination and the closure under R-convolution. The essence of these constructions is to exploit the recursive nature of the relation R by iterating the closure properties.

4.1 Iteration of a simple convolution

First let us define the finite iteration of a simple convolution. Let $X_1 = X_2 = X$, and R be a finite function. Let $x_1 \circ x_2$ be the (unique) x such that $R(x_1, x_2, x)$. We say that R is *associative* if $x_1 \circ (x_2 \circ x_3) = (x_1 \circ x_2) \circ x_3$ for all $x_1, x_2, x_3 \in X$. In this case $(X, \circ)$ is a *semigroup*. The relation R is associative in Examples 2 and 3 in Section 2. We assume associativity of R in what follows.

Let $g : X \to \Re$ be a probability distribution on X. Then we define $g^{(1)} = g$, and for every $r \geq 2$ we define $g^{(r)} = g \star g^{(r-1)}$, where $g \star h$ is the simple R-convolution defined in Equation (5). Technically, this is the left iteration of g, but it is easily verified that since R is associative, then this is the same as the right iteration $g^{(r)} = g^{(r-1)} \star g$, and so there is no loss of specificity in using this notation. It is clear that $g^{(r)}$ is a probability distribution on X for all $r \geq 1$.

Iteration of a simple convolution is used extensively in the application of generating functions. For example, if $\mathbf{X}_1, \ldots, \mathbf{X}_r$ are independent random variables with a geometric distribution $g(n) = q^n p$ where $p + q = 1$, and G is the generating function for this distribution, then

$$G(s) = \frac{p}{1 - qs}.$$

Thus $\mathbf{Y} = \sum_{i=1}^{r} \mathbf{X}_i$ has generating function

$$\left(\frac{p}{1 - qs} \right)^r,$$

and it follows that $\mathbf{Y}$ has the negative binomial distribution

$$g^{(r)}(n) = \left(\begin{array}{c} r + n - 1 \\ n \end{array} \right) q^n p^r.$$

These and other classical examples, as found in, e.g., [6], use iterated convolution under the semigroup $(\{0, 1, \ldots\}, +)$ defined by the associative relation in Example 3 of Section 2.

4.2 Infinite iteration of a simple convolution and probability distributions on regular languages

Let $0 \leq \gamma < 1$. We define the γ-infinite iteration of g by

$$g_\gamma^\star = (1-\gamma) \sum_{r=1}^{\infty} \gamma^{r-1} g^{(r)}.$$

We call this the (generalized) Kleene star operation for reasons discussed below[4]. Clearly, $g_\gamma^\star$ is a convex combination of the $g^{(r)}$ using a geometric distribution with parameter γ. Hence $g_\gamma^\star$ is a probability distribution on X.

Let X be the set of all finite strings over a finite alphabet $\mathcal{A}$ and $\circ$ be the operation of string concatenation, as in Example 2 of Section 2. Let ϵ denote the empty string. A subset of X is called a *language*. The operation of concatenation is extended to languages by defining

$$L_1 \circ L_2 = \{x \circ y : x \in L_1 \text{ and } y \in L_2\}.$$

The iteration of this operation is defined by $L^{(1)} = L$, and $L^{(r)} = L \circ L^{(r-1)}$, $r \geq 2$. The Kleene star operation is defined by

$$L^\star = \{\epsilon\} \cup \bigcup_{r \geq 1} L^{(r)}.$$

Finally, the *regular languages* are defined to be the smallest set of languages that contain $\{\epsilon\}$ and $\{a\}$ for all letters $a \in \mathcal{A}$, and are closed under union, concatenation and Kleene star [14].

The operations of convex combination, simple convolution, and γ-iterated convolution may be used to define a class of probability distributions on regular languages called *regular probability distributions* For any string x, we call the distribution $g_x(y) = \delta(x, y)$ the *indicator distribution* for x. Let g and h be two probability distributions on X. Corresponding to the operation of union, for any $0 < \gamma < 1$, we can form the (binary) convex combination $\gamma g +$

[4]It is actually more like the regular operator X^+ than $X^\star$.

$(1-\gamma)h$, which is clearly also a probability distribution on X. Corresponding to the operation of concatenation, we have the convolution $g \star h$. It is clear that $g_x \star g_y = g_{x \circ y}$. Finally, corresponding to Kleene star, we have the γ-iterated convolution $g^\star_\gamma$. The class of regular probability distributions on X is the smallest class of probability distributions that contains the indicator functions for the empty string and all letters of the alphabet $\mathcal{A}$, and is closed under binary convex combination, convolution, and γ-iterated convolution for any γ.

Notice that $\{\epsilon\} \cup \mathcal{A}$ forms the (minimal) set of *generators* for the semigroup $(X, \circ)$, in the sense that any element of X can be constructed by applying the operation $\circ$ finitely many times to these generators. The above definition of regular probability distributions is easily extended to any semigroup by defining it to be the smallest class of probability distributions that contains the indicator functions for the generators and is closed under binary convex combination, convolution, and γ-iterated convolution for any γ. We denote this set of distributions by $\mathcal{G}$.

4.3 Iterated convolution of P-kernels

Analogous operations can be defined for P-kernels. Let $K : X \times X \to \Re$ be a P-kernel, and R be a finite associative function representing an operation $\circ$. Then $K^{(r)}$ denotes the R-convolution of K with itself r times, and

$$K^\star_\gamma = (1-\gamma) \sum_{r=1}^{\infty} \gamma^{r-1} K^{(r)}$$

is the γ-infinite iteration of K. These are P-kernels on the semigroup $(X, \circ) \times (X, \circ)$.

Building on the set $\mathcal{G}$ of regular distributions on $(X, \circ)$, we define the set of *regular P-kernels* on $(X, \circ) \times (X, \circ)$, denoted $\mathcal{K}$, as the smallest class of P-kernels that contains the the kernel $K(x, y) = g(x)g(y)$ for every $g \in \mathcal{G}$ and is closed under binary convex combination, convolution, and γ-infinite convolution for any γ. If $(X, \circ)$ is the semigroup of strings over a finite alphabet with the operation of concatenation, then we call $\mathcal{K}$ the class of *regular string kernels*.

A fuller theory of regular P-kernels, along with their representations as machines and grammars, and their extension to stochastic context-free gram-

mar kernels is in preparation as a separate paper. Here we present a simple application of this theory.

4.4 Application of regular string kernels

Here we derive a regular string kernel that can be used to measure the similarity between strings, based on an underlying generative probability model for pairs of strings. This application will be discussed in detail, and experimental results presented, in a separate paper. In pattern recognition, a string may represent a sequence of elementary objects derived from the decomposition of a structured object. Each elementary object can be denoted by a letter in the finite alphabet $\mathcal{A}$. The set X of all objects is thus identified with the semigroup of all finite strings over the alphabet $\mathcal{A}$.

In many pattern recognition applications, we can not assume that all object strings from similar objects have the same length. This occurs, for example, when the strings consist of amino acids representing proteins, nucleic acids representing genes, or phonemes representing spoken words [26, 4, 9, 24]. In these contexts, some objects may be missing components that other similar objects have. However, we can align any two object strings so that their corresponding components are adjacent, using a special symbol '-' to indicate that a component is missing at a certain place in one of the strings. For example, using the alphabet $\mathcal{A} = \{A, B, C\}$, and the strings $x = BCABBCBAACACAACCCAAB$ and $y = BCCABBCABBAABACAACCAAB$,

```
x = BC-ABBC--BAACACAACCCAAB
y = BCCABBCABBAABACAACC-AAB
```

represents an alignment between x and y with four insertions or deletions of components and one substitution (a component of type 'C' exchanged with one of type 'B' in one place.)

One way to define a generative probability model that captures the essential properties of such string alignments is to model the strings x and y as having been derived from a common "ancestor" string z. For example, one choice would be to take $z = BCABBCBAACACAACCAAB$ and show the derivation of x and y from z as

```
z = BCABBCBAACACAACC-AAB
x = BCABBCBAACACAACCCAAB
```

and

```
z = BC-ABBC--BAACACAACCAAB
y = BCCABBCABBAABACAACCAAB
```

Here we have chosen z such that in going from z to either x or y, we only make insertions and substitutions, no deletions. It is always possible to choose such a z to represent a common ancestor of two aligned strings, so we will assume that the derivation is always done in that way.

In order to model this derivation process, we start by defining a kernel that models the substitution process on a single letter. We assume that given an ancestor letter $a \in \mathcal{A}$, $p(b|a)$ denotes the probability that this ancestor derives the letter b in the string x. We assume that this probability is the same for the string y, and that the derivation of the two letters, one in x and one in y, is independent, given the ancestor letter a in z. Finally, we assume that the probability of the ancestor letter a is given by $p(a)$.

For every $a \in \mathcal{A}$, the zero extension to all finite strings in X of the kernel K_a on $\mathcal{A} \times \mathcal{A}$ defined by $K_a(b, c) = p(b|a)p(c|a)$ is a regular string kernel by the basis case of the inductive definition. For any strings $x, y \in X$, define

$$K_1(x, y) = \sum_{a \in \mathcal{A}} p(a) K_a(x, y).$$

Since the class of regular string kernels is closed under finite convex combination, K_1 is a regular string kernel as well. K_1 models the substitution process for single letters, and is zero for all strings that are not single letters.

Next we model the insertion process. Between any two consecutive letters of the ancestor z, arbitrary stings can be inserted in the corresponding places in x and y. We assume that the inserted string in x is independent from the inserted string in y. Let g be any regular probability distribution on strings. For example, it is easy to see that the distribution in which the length of the string has a geometric distribution and the letters are independently chosen according to any fixed distribution on the alphabet is a regular probability distribution. We may take $K_2(x, y) = g(x)g(y)$ as a model of our insertion process, which again is a regular string kernel by the basis case of the inductive definition.

Finally, to generate a pair of strings x and y that are derived randomly from a common ancestor, we have to iterate the processes of insertion and

substitution. For some parameter $0 \leq \gamma < 1$, let

$$K(x,y) = \gamma K_2 \star (K_1 \star K_2)^{\star}_{\gamma} + (1-\gamma)K_2. \tag{12}$$

Clearly $K(x,y)$ is a regular string kernel, and it models the generative process we have defined. To see this more clearly, it is useful to look at some special cases.

- If $\gamma = 0$ then $K(x,y) = k_2(x,y) = g(x)g(y)$, i.e. K is an independent kernel, so the strings x and y are modeled as being independently generated according to the same underlying insertion process, with the empty common ancestor. The "similarity" $K(x,y)$ depends only on the magnitude of the individual probabilities $g(x)$ and $g(y)$.

- If $K_2(x,y) = g_\epsilon(x)g_\epsilon(y)$, where ϵ denotes the empty string, then $K_2(x,y) = 0$ unless both x and y are empty. Thus no insertions whatsoever are allowed. In this case it is easily verified that $K_1 \star K_2 = K_1$, and that if $x = x_1 \ldots x_s$ and $y = y_1, \ldots, y_t$, where $x_i, y_j \in \mathcal{A}$, then for $r \geq 1$, $K_1^{(r)}(x,y) = 0$ unless $r = s = t$, in which case $K_1^{(r)}(x,y) = \prod_{i=1}^{r} K_1(x_i, y_i)$. From this, it follows easily that $K(x,y) = 0$ if x and y have different lengths, else if they have the same length r then

$$K(x,y) = (1-\gamma)\gamma^r \prod_{i=1}^{r} K_1(x_i, y_i). \tag{13}$$

  Here we take the product to be 1 if $r = 0$. Thus in this case the kernel K decomposes into a geometric mixture of disjoint product distributions on pairs of strings of different lengths. Stings x and y are completely dissimilar ("orthogonal") if they have different lengths, else they are similar to the extent that their corresponding letters are similar (= "likely to have been derived from a common ancestor letter".) A further special case is obtained if $\mathcal{A} = \{0,1\}$, $K_1(0,0) = K_1(1,1) = \alpha$ and $K_1(0,1) = K_1(1,0) = \beta$ where $0 < \beta < \alpha < 1$ and $\alpha + \beta = 1/2$. Then if x and y have the same length r,

$$K(x,y) = (1-\gamma)(\gamma\alpha)^r \left(\frac{\alpha}{\beta}\right)^{-d_H(x,y)}, \tag{14}$$

  where $d_H(x,y)$ is the *Hamming distance* between the binary strings x and y, defined as the number of components in which they differ.

In general, K interpolates between the two extremes given in these cases, so strings that have similar overall structure, with a few likely insertions, deletions, and substitutions, are more similar under K than strings of comparable length that don't share this property.

It can be shown that regular string kernels can all be modeled by *pair-HMMs*, as defined in [4]. This means that there is an efficient dynamic programming algorithm to evaluate these kernels, which is a very important practical consideration. However, not all pair-HMMs define regular string kernels. For example, it is possible to define a pair-HMM that represents a distribution on pairs of strings that is not symmetric, and hence not a kernel. The regular string kernel K defined above is closely related to the joint probability distribution defined by the pairwise local alignment pair-HMM on page 86 (Figure 4.3) of [4]. That distribution incorporates a few bells and whistles that are easily accomplished by using a slightly more complex regular string kernel, with the exception of one feature: in the definition of that pair-HMM, between consecutive letters of a hypothetical ancestor string, a string can be inserted in the corresponding position in either x or y, but not in both. This kind of distribution also cannot be modeled by a kernel, because it fails to be positive definite. However, if one replaces this part of the model by an independent distribution on the insertions in x and y like those used at the beginning and ends of this pair-HMM from [4], then the distribution of the resulting pair-HMM is a regular string kernel.

Pair-HMMs have already proven useful in pattern recognition applications involving strings by virtue of the generative models they define. By developing them further into regular string kernels, we can take advantage of other kinds of pattern recognition and clustering methods that use an implicit feature-space representation, obtained from the kernel.

5 Generalized radial basis kernels and radial distances

In Section 3.4, we showed how to derive a Gibbs kernel from an arbitrary positive kernel by normalizing it. A different kind of normalization of a positive kernel K is to convert it into a *generalized radial basis* kernel, defined

in analogy with Equation (4) of Section 2.3 by

$$\tilde{K}(x,y) = \frac{K(x,y)}{\sqrt{K(x,x)}\sqrt{K(y,y)}} = e^{-((1/2)(\log K(x,x)+\log K(y,y))-\log K(x,y))}. \quad (15)$$

Here we assume $K(x,x) > 0$ for all x. If $K(x,x) = 0$, it is easy to see that we must have $K(x,y) = 0$ for all y, else K is not positive definite. Hence we can remove all x such that $K(x,x) = 0$ if necessary.

It is clear that $\tilde{K}$ is a kernel, since it is the product of K with the kernel $K'(x,y) = \sqrt{K^{-1}(x,x)}\sqrt{K^{-1}(y,y)}$. Furthermore, because K is positive definite, for any $x, y \in X$, the matrix

$$\begin{pmatrix} K(x,x) & K(x,y) \\ K(y,x) & K(y,y) \end{pmatrix}$$

is symmetric and has nonnegative eigenvalues, and hence its determinant is nonnegative. It follows that

$$|K(x,y)| \leq \sqrt{K(x,x)}\sqrt{K(y,y)} \quad (16)$$

for all $x, y \in X$, which can be viewed as a generalized Cauchy-Schwarz inequality, since any semi-inner product is a kernel (see Equation (24) in Section 7). Thus for a positive kernel K

$$0 \leq \tilde{K}(x,y) \leq 1 \quad (17)$$

and the "radial distance"

$$d^2(x,y) = \frac{1}{2}(\log K(x,x) + \log K(y,y)) - \log K(x,y) \quad (18)$$

is always nonnegative, is 0 when $x = y$, and is infinite when $K(x,y) = 0$. The normalization of the kernel values to the range $[0,1]$ can be quite important in practice.

As an example, consider the regular string kernel K defined in Equation (12) from Section 4.4. The square of the corresponding generalized radial basis kernel $\tilde{K}$ represents a kind of odds ratio, comparing the probability $K^2(x,y)$ that x and y would be generated together from a common ancestor on two independent occasions to the probability $K(x,x)K(y,y)$ that two

copies of x would be generated from a common ancestor on one occasion, and two copies of y would be generated from a common ancestor on an independent occasion. An analogous interpretation for $\tilde{K}$ can be given for any P-kernel K. For P-kernels K such as that in Equation (12), where the values of K rapidly get exponentially small as the size of x and y increase, moving from K to $\tilde{K}$ also has the advantage of normalizing the similarity measurement to remove some undesirable aspects of this length dependency, and helping to keep the values of K within a representable range. They still get extremely small, however, and in practice, one would prefer to deal only with the generalized radial distance $-\log \tilde{K}(x,y) = d^2(x,y)$.

In fact, there is a large literature on closely related types of distances between strings, going back to early work of Ulam and colleagues, and further developed by many others [26]. The earlier work did not derive these distances from probability models for insertions deletions and substitutions; this idea was introduced fairly recently [4]. Most theoretical discussions of such string distances have only been concerned with the question of whether or not a certain distance function $d(x,y)$ is a metric, i.e. if it is symmetric, satisfies the triangle inequality $d(x,y) \leq d(x,z) + d(z,y)$, and has the property that $d(x,x) = 0$. However, for many pattern recognition applications, this is not sufficient for d to be a useful distance [20]. For a distance d to be useful, we need to actually embed the metric space (X,d) in a finite dimensional Euclidean space $\Re^N$, or in the space of all infinite square summable sequences l_2, via some feature extraction mapping $\phi(x) = \{\phi_n(x)\}$ such that $d^2(x,y) = \sum_n (\phi(x) - \phi(y))^2$. This is called an *isometric embedding*. String distances cannot in general be isometrically embedded into a finite dimensional Euclidean space, hence Linial et al. develop general methods by which these distances, and in fact any general metric distance, can be *approximately* embedded in a finite dimensional Euclidean space, in the sense that the Euclidean distance between $\phi(x)$ and $\phi(y)$ is close to the original distance $d(x,y)$ for all x and y [20]. They apply these results to the problem of classifying protein sequences [19]. However, if (X,d) can be embedded in l_2, as mentioned in the introduction, we can still take advantage of most of the classical pattern recognition, clustering, regression and classification methods via the kernel formulation. Thus an interesting open question is the following.

Question 1 *If K is a regular string kernel as defined in (12), and $d^2(x,y) =$*

$\frac{1}{2}(\log K(x,x) + \log K(y,y)) - \log K(x,y)$, *when can* (X,d) *be isometrically embedded in* l_2*? More generally, if* K *is an arbitrary positive kernel on* $X \times X$*, and* $d^2(x,y) = \frac{1}{2}(\log K(x,x) + \log K(y,y)) - \log K(x,y)$*, when can* (X,d) *be isometrically embedded in* l_2*?*

If we allow $K(x,y) = 0$ for some x and y, then $d^2(x,y) = \infty$ for these x and y, thus (X,d) cannot be isometrically embedded in l_2. However, d^2 might still be a useful distance if it decomposed X into sets $X_1, X_2, \ldots$ such that $d^2(x,y) < \infty$ for $x, y \in X_n$, $d^2(x,y) = \infty$ for $x \in X_n$ and $y \in X_m$, $n \neq m$, and if (X_n, d) is isometrically embedded in l_2 for each X_n. In this case we say that (X,d) *can be isometrically embedded in a disjoint union of* l_2 *spaces*. More generally, we have the following question.

Question 2 *If* K *is an arbitrary positive kernel on* $X \times X$*, and* $d^2(x,y) = \frac{1}{2}(\log K(x,x) + \log K(y,y)) - \log K(x,y)$*, when can* (X,d) *be isometrically embedded in a disjoint union of* l_2 *spaces?*

It is interesting to consider this question in the extreme special cases for the regular string kernel discussed in Section 4.4. In particular, if $K(x,y) = g(x)g(y)$ for some g, which we will assume is strictly positive, then $\tilde{K}(x,y) = 1$ for all x, y, and hence $d^2(x,y) = 0$ for all x, y. Thus (X,d) can be isometrically embedded into a zero dimensional space. On the other hand, if, as in (14), $K(x,y) = 0$ for x and y of different lengths, and for x and y of length r,

$$K(x,y) = (1-\gamma)(\gamma\alpha)^r \left(\frac{\alpha}{\beta}\right)^{-d_H(x,y)},$$

then

$$d^2(x,y) = -\log \tilde{K}(x,y) = \log(\alpha/\beta) d_H(x,y)$$

if x and y have the same length, else $d^2(x,y) = \infty$. The Hamming distance $d_H(x,y)$ is clearly the squared Euclidean distance $\sum_i (x_i - y_i)^2$ in the case of binary sequences x and y. Thus in this case (x,d) can be isometrically embedded into the disjoint union of $\Re^N$ for $N \geq 0$.

Even if (X,d) cannot be isometrically embedded into a useful space, then, using kernel methods of pattern recognition, we can still work directly with the generalized radial basis function $\tilde{K}(x,y)$, which has the form

$$\tilde{K}(x,y) = e^{-d^2(x,y)}. \tag{19}$$

It is very important in practice to be able to scale the radial distance $d^2(x,y)$ in such a radial basis kernel by a positive "width" parameter σ^2, to obtain

$$\tilde{K}(x,y) = e^{-d^2(x,y)/\sigma^2} = \tilde{K}^{1/\sigma^2}(x,y). \tag{20}$$

If the width is too large, then $\tilde{K}(x,y)$ is nearly 1 for all x, y, and if the width is too small, $\tilde{K}(x,y) \approx \delta(x,y)$. The kernel is not useful in either case. The width parameter is often set by cross-validation to optimize a given performance measure. If K is a P-kernel, then we might get away by keeping $\sigma = 1$, and adjusting parameters internal to the P-kernel to make it either more or less peaked on the diagonal, which would have an effect similar to that of adjusting the width parameter. For example, we can adjust the parameter γ, and the definition of the kernels K_1 and K_2 modeling substitution and insertion, to control how much the regular string kernel K from (12) is peaked on the diagonal. However, it would be more convenient to simply adjust the width parameter σ, and truer to the underlying generative probability model.

Unfortunately, if $1/\sigma^2$ is positive but not a positive integer, then we have no reason to believe that the kernel $e^{-d^2(x,y)/\sigma^2}$ defined above is positive definite. Thus we are lead to another general question:

Question 3 *If $K(x,y)$ is a positive kernel, and $d^2(x,y) = \frac{1}{2}(\log K(x,x) + \log K(y,y)) - \log K(x,y)$, when is $\tilde{K}(x,y) = e^{-d^2(x,y)/\sigma^2} = \tilde{K}^{1/\sigma^2}(x,y)$ also a kernel for all $\sigma > 0$?*

It turns out that the answer to Questions 2 and 3 are the same, and that these properties hold if and only if K is *infinitely divisible*. We briefly review the relevant theory in the following section.

6 Infinitely Divisible Kernels

6.1 Definition of infinitely divisible kernels

Let $K(x,y)$ be a positive kernel on a set $X \times X$. The kernel K is called *infinitely divisible* if for each positive integer n there is a kernel K_n such that $K = K_n^n$. Such kernels are related to the family of infinitely divisible probability distributions, which are the limits of sums of independent variables (see [7]). To describe the properties of infinitely divisible kernels, we need some additional definitions.

6.2 Negative definite kernels

A function $N(x, y)$ is *negative definite* if it is symmetric and for all $x_1, \ldots, x_n$ in X and real $c_1, \ldots, c_n$ such that $\sum_{i=1}^n c_i = 0$,

$$\sum_{ij} c_i c_j N(x_i, x_j) \leq 0.$$

Note the extra condition that $\sum_{i=1}^n c_i = 0$. This means that if K is positive definite, then $-K$ is negative definite, but the converse does not generally hold. Clearly the class of negative definite functions also forms a closed convex cone, as does the class of kernels.

Negative definite kernels have one useful closure property that positive definite kernels do not.

Lemma 2 *Let N be a function on $X \times X$ and $f : X \to \Re$. Then $N(x, y)$ is negative definite iff $N(x, y) + f(x) + f(y)$ is negative definite.*

Proof: Clearly N is symmetric iff $N(x, y)+f(x)+f(y)$ is. For any $x_1, \ldots, x_n$ in X and real $c_1, \ldots, c_n$ such that $\sum_{i=1}^n c_i = 0$,

$$\begin{aligned} & \sum_{ij} c_i c_j (N(x_i, x_j) + f(x_i) + f(x_j)) \\ = & \sum_{ij} c_i c_j N(x_i, x_j) + (\sum_j c_j)(\sum_i c_i f(x_i)) + (\sum_i c_i)(\sum_j c_j f(x_j)) \\ = & \sum_{ij} c_i c_j N(x_i, x_j) \end{aligned}$$

Hence the former is negative for any $c_1, \ldots, c_n$ such that $\sum_{i=1}^n c_i = 0$ iff the latter is. □

We say that two negative definite functions are *equivalent* if they differ by $f(x) + f(y)$ for some function f. We say that a function N on $X \times X$ *has a zero diagonal* if $N(x, x) = 0$ for all $x \in X$. Note that if N is negative definite, then $N'(x, y) = N(x, y) - (1/2)(N(x, x) + N(y, y))$ is equivalent to N and has a zero diagonal, so up to equivalence, we may assume that all negative definite functions have a zero diagonal.

There is a close relationship between positive and negative definite functions.

Lemma 3 *Let $N(x, y)$ be a symmetric function on $X \times X$ with zero diagonal, and z be any element of X. Let $K(x, y) = N(x, z) + N(y, z) - N(x, y)$. Then K is positive definite iff N is negative definite.*

Proof: Like the previous lemma, this follows easily from the definitions. See, e.g., [1], Lemma 2.1, page 74 for a proof. □

As we will see below, there is also a close relationship between negative definite functions and squared distances. Here is some intuition about this. Assume $N(x, y)$ is a symmetric function with a zero diagonal. We say that $N(x, y)$ satisfies the *2n-gonal inequalities* if for every positive integer n and every $x_1, \ldots, x_n, y_1, \ldots, y_n$ in X,

$$\sum_{i<j} N(x_i, x_j) + \sum_{i<j} N(y_i, y_j) \leq \sum_{ij} N(x_i, y_j).$$

The 2-gonal inequality says that N is positive, and the 4-gonal inequality says (basically) that the sum of squared lengths of the two diagonals of a quadrilateral is always less than or equal to the sum of the squares of the side lengths, which is true of any Euclidean distance.

Here is a simple result that is well-known.

Theorem 3 *Let $N(x, y)$ be a symmetric function with a zero diagonal. Then $N(x, y)$ is negative definite iff it satisfies the 2n-gonal inequalities for every positive integer n.*

Proof: First note that in the condition $\sum_{ij} c_i c_j N(x_i, x_j) \leq 0$ in the definition of a negative definite function, we can restrict to integer $c_1, \ldots, c_n$. This follows easily from the fact that negative definite kernels are closed under pointwise limits and multiplication by a positive constant. Furthermore, by duplicating elements in $x_1, \ldots, x_n$ as needed, we can even stipulate that $c_i \in \{\pm 1\}$. Now the condition $\sum_i c_i = 0$ implies that half the c_i are $+1$ and the other half are -1. Rename the x_i associated with the negative c_i as y_i, and assume we end up with $x_1, \ldots, x_n$ and $y_i, \ldots, y_n$. If you now write out the condition $\sum_{ij} c_i c_j N(x_i, x_j) \leq 0$ you get the $2n$-gonal inequality. □

6.3 Main result on infinitely divisible kernels

For convenience, it is useful to allow negative definite functions to have the value $N(x, y) = \infty$ for some $x \neq y$. This way, for any positive kernel K, $-\log K(x, y)$ is defined, and equals ∞ when $K(x, y)$ is zero. Assume N is a symmetric function with some off-diagonal values of infinity, $x_1, \ldots, x_n \in X$, $I = \{(i, j) : N(x_i, x_j) = \infty\}$, and $J = \{(i, j) : N(x_i, x_j) < \infty\}$. Then for any real $c_1, \ldots, c_n$ we define $\sum_{i,j} c_i c_j N(x_i, x_j)$ to be

- ∞ if $\sum_{i,j\in I} c_i c_j > 0$,
- $-\infty$ if $\sum_{i,j\in I} c_i c_j < 0$, and
- $\sum_{i,j\in J} c_i c_j N(x_i, x_j)$ if $\sum_{i,j\in I} c_i c_j = 0$.

We say that N is negative definite if $\sum_{i,j} c_i c_j N(x_i, x_j) \leq 0$ for all $c_1, \ldots, c_n$ such that $\sum_{i=1}^n c_i = 0$.

We define the relation $x \equiv_N y$ iff $N(x, y) < \infty$. It is easily verified that N is negative definite iff $\equiv_N$ is an equivalence relation and N is negative definite in the usual sense (as defined at the beginning of this section) on each equivalence class of this equivalence relation. Furthermore, these extended negative definite kernels remain closed under pointwise limits.

A characterization of infinitely divisible kernels is the following, primarily obtained by Schönberg around 1940.

Theorem 4 *Let K be a positive kernel, $N = -\log K$, and $d^2(x, y) = \frac{1}{2}(\log K(x, x) + \log K(y, y)) - \log K(x, y)$. Then the following are equivalent*

1. *K is infinitely divisible*
2. *$K^t(x, y)$ is a kernel for every $t > 0$*
3. *N is negative definite*
4. *d^2 is negative definite*
5. *$\equiv_{d^2}$ is an equivalence relation on X and (S, d) can be isometrically embedded[5] in l_2 for every equivalence class S in this equivalence relation.*

Proof: Our treatment follows [1], Proposition 2.7, page 77. To see that 1 implies 3, first note that if K is infinitely divisible then $K^{1/n} = e^{-N/n}$ is positive definite for all positive integers n. Hence $1 - e^{-N/n}$ is negative definite for all positive integers n. It is clear that $N = \lim_{n\to\infty} n\left(1 - e^{-N/n}\right)$. Thus, since the negative definite kernels are closed under pointwise limits, N is negative definite.

[5]Recall that we are assuming that K is a continuous function and X is a separable metric space X. Without this assumption, this result holds with l_2 replaced by an arbitrary Hilbert space.

To see that 3 implies 2, note that if N is negative definite, then so is tN for any $t > 0$. Define $N'(x, y) = tN(x, y) - (1/2)t(N(x, x) + N(y, y))$, which is also negative definite by Lemma (2), and has zero diagonal. Assume $z \in X$ and define

$$K'(x, y) = N'(x, z)+N'(y, z)-N'(x, y) = t(N(x, z)+N(y, z)-N(x, y)-N(z, z)).$$

K' is positive definite by Lemma 3. Note that

$$e^{-tN(x,y)} = \left(e^{K'(x,y)}\right)\left(e^{-tN(x,z)}e^{-tN(y,z)}\right)\left(e^{tN(z,z)}\right).$$

The last two terms are clearly positive definite functions of x and y, and the first is positive definite since by Taylor expansion it is a limit of sums of powers of K' with positive coefficients, which are all positive definite by the closure under product. Thus $e^{-tN(x,y)}$ is positive definite by closure under product. Since $K^t = e^{-tN(x,y)}$, the result follows.

Finally, 2 clearly implies 1 by taking $t = 1/2, 1/3, 1/4, \ldots$. Hence 1, 2 and 3 are all equivalent.

The equivalence of 3 and 4 follows from Lemma 2. The equivalence of 4 and 5 is somewhat more involved, and is postponed until Section 7. □

It is obvious from this that Questions 2 and 3 from the previous section are equivalent to each other, and to the infinite divisibility of K.

Not every positive kernel is infinitely divisible, but it is not immediate to produce an example that shows this. Fitzgerald and Horn give a nice example in their paper, the main theorem of which is that if a real symmetric $n \times n$ matrix K is positive and positive definite, then the fractional Schur power $K^t = \{K_{ij}^t\}$ is positive definite for all $t \geq n - 2$ [8]. (This result was rediscovered 19 years later [25].) Let $\mathbf{1}$ denote the all 1s vector and $\mathbf{n}$ denote $(1, 2, \ldots, n)^T$. The example Fitzgerald and Horn supply to show this bound is tight is a matrix of the form $M_\delta = \mathbf{1}\mathbf{1}^T + \delta\mathbf{n}\mathbf{n}^T$. They show that for $n \geq 3$ and sufficiently small δ, M_δ^t is not positive definite for any non-integer $t < n - 2$, and hence M_δ is not infinitely divisible. (This can't work for $n = 2$: it is easily verified that for any positive 2 by 2 matrix M, M is positive definite iff M is infinitely divisible.)

It is obvious that any independent positive kernel is infinitely divisible. Hence the example of Fitzgerald and Horn shows that the class of infinitely divisible kernels is not closed under sum. Since we showed in Section 2.4 that sum can be represented as a convolution, it follows that this class is not

closed under convolution either. It is closed under product and fractional positive powers, however.

We have made some progress in answering Questions 2 and 3 above, but we still lack a useful operational way to test for infinite divisibility in practice. Assuming the kernels $K_1, \ldots, K_D$ are infinitely divisible, when is their convolution infinitely divisible? Given a positive kernel that is not infinitely divisible, how can we modify it in the least manner so that it becomes infinitely divisible? It can be shown that this can be accomplished for finite X by multiplying the diagonal elements of the kernel by a large enough constant. This is analogous to adding to the diagonal of a (not necessarily positive) symmetric function on a finite set to make it positive definite, a trick that is often used in practice. However, this does not appear to be appropriate for kernels on countably infinite sets. It would also be particularly pleasing if a rich subclass of convolution kernels could be proven to be infinitely divisible, or if simple methods could be found for modifying them so that they are infinitely divisible.

7 Kernels and reproducing kernel Hilbert spaces

Here we summarize some aspects of Reproducing Kernel Hilbert Spaces (RKHSs) as they relate to the results above. This allows us to characterize a kernel K as $K(x,y) = \sum_n \phi_n(x)\phi_n(y)$, as promised in Section 2.1 and finally provide the proof to Lemma 1, and the final step of the proof of Theorem 4. Our treatment follows [22, 1].

Let X be a set with a metric $d(x,y)$. We refer to X as a *metric space* in this case, when we wish to leave the specific metric d unspecified. We say X is *separable* if there exists a countable set $X_0 \subseteq X$ that is *dense* in X, i.e. for all $x \in X$ and all $\epsilon > 0$ there exists $y \in X_0$ with $d(x,y) \leq \epsilon$. If X is countable, then we assume that X is endowed with the discrete metric $d(x,y) = \delta(x,y)$, thus rendering it a separable metric space. A function $f : X^n \to \Re$ for some $n \geq 1$ is *continuous* if for all $\epsilon > 0$ there exists a $\delta > 0$ such that whenever $d(x_i, y_i) \leq \delta$ for all $1 \leq i \leq n$, then $|f(x_1, \ldots, x_n) - f(y_1, \ldots, y_n)| \leq \epsilon$. It is clear that if X is endowed with the discrete metric then any function $f : X^n \to \Re$ is continuous. Here, and throughout this paper, we assume that X is a separable metric space, and $K : X \times X \to \Re$ is a continuous kernel.

By H_0 we denote the linear space of real-valued functions on X generated

by the functions $\{K_x : x \in X\}$, where

$$K_x(y) = K(x, y).$$

Let $f = \sum_{i=1}^{n} c_i K_{x_i}$ and $g = \sum_{j=1}^{m} d_j K_{y_j}$ be elements of H_0, where $c_i, d_j \in \Re$ and $x_i, y_j \in X$ for $1 \leq i \leq n$ and $1 \leq j \leq m$. Define

$$\langle f, g \rangle = \sum_{j=1}^{m} d_j f(y_j) = \sum_{i,j} c_i d_j K(x_i, y_j) = \sum_{i=1}^{n} c_i f(x_i). \tag{21}$$

Note that the function $\langle \cdot, \cdot \rangle$ does not depend on the chosen representations of f and g. It is clearly bilinear, i.e. $\langle cf + g, h \rangle = c \langle f, h \rangle + \langle g, h \rangle$ for any $c \in \Re$ and $f, g, h \in H_0$, and symmetric, i.e. $\langle f, g \rangle = \langle g, f \rangle$ for any $f, g \in H_0$, and because K is positive definite, for any f as above,

$$\langle f, f \rangle = \sum_{i,j} c_i c_j K(x_i, x_j) \geq 0.$$

Hence it is a semi-inner product on H_0 (see, e.g., [3], page 123).

It follows from Equation (21) that this semi-inner product has the *reproducing property*

$$\langle f, K_x \rangle = f(x) \tag{22}$$

for all $f \in H_0$ and $x \in X$. This implies that

$$\langle K_x, K_y \rangle = K(x, y). \tag{23}$$

Any semi-inner product $\langle \cdot, \cdot \rangle$ is a kernel, because for any $a_1, \ldots, a_N \in \Re$ and $f_1, \ldots, f_N$ in the space that the semi-inner product is defined on,

$$\begin{aligned} \sum_{i,j=1}^{N} a_i a_j \langle f_i, f_j \rangle &= \sum_{j=1}^{N} a_j \sum_{i=1}^{N} a_i \langle f_i, f_j \rangle \\ &= \sum_{j=1}^{N} a_j \left\langle \sum_{i=1}^{N} a_i f_i, f_j \right\rangle \\ &= \left\langle \sum_{i=1}^{N} a_i f_i, \sum_{j=1}^{N} a_j f_j \right\rangle \\ &= \left\langle \sum_{i=1}^{N} a_i f_i, \sum_{i=1}^{N} a_i f_i \right\rangle \geq 0. \end{aligned} \tag{24}$$

Hence $\langle \cdot, \cdot \rangle$ is a kernel on $H_0 \times H_0$. Thus by Inequality (16) and Equation (22),

$$f^2(x) \leq \langle f, f \rangle K(x, x), \tag{25}$$

which implies that $\langle f, f \rangle = 0$ only if for all $x \in X$, $f(x) = 0$. Hence $\langle \cdot, \cdot \rangle$ is not just a semi-inner product, but an inner product on H_0. The norm for the inner product space $(H_0, \langle \cdot, \cdot \rangle)$, as for any inner product space, is defined by $||f|| = \sqrt{\langle f, f \rangle}$, and the distance between f and g by $||f - g||$.

A *Hilbert space* is an inner product space that is *complete*, in the sense that every Cauchy sequence $\{f_n\}_{n \geq 1}$, i.e. every sequence such that $\sup_{m \geq n} ||f_n - f_m|| \to 0$ as $n \to \infty$, converges to a g that is in the space. If $\{f_n\}_{n \geq 1}$ is a Cauchy sequence in H_0, then $(f_n(x) - f_m(x))^2 \leq ||f_n - f_m||^2 K(x, x)$ by Inequality (25), and hence $f_n(x) \to f(x)$ for some real-valued function f on X. It is possible to *complete* H_0 by adding the limits of Cauchy sequences to it, extending it and its inner product to larger class H of real-valued functions H that includes the functions in H_0, and is a Hilbert space. Moreover, the space H_0 is dense in H. The space H is called the *Reproducing Kernel Hilbert Space* associated with the kernel K [22].

It is easy to see that H also has the reproducing property (22). Assume $f \in H$ and $x \in X$. Let $\{f_n\}_{n \geq 1}$ be a Cauchy sequence in H_0 such that $||f_n - f|| \to 0$. As discussed above, this implies that $f_n(x) \to f(x)$. By (22), $f_n(x) = \langle f_n, K_x \rangle$ for all n. Thus

$$\langle f, K_x \rangle = \lim_{n \to \infty} \langle f_n, K_x \rangle = \lim_{n \to \infty} f_n(x) = f(x).$$

We now show that since X is a separable metric space and K is continuous, H is separable. Let X_0 be a countable dense subset of X. Let H_0' be the set of all functions of the form $f = \sum_{i=1}^n c_i' K_{x_i'}$ for rational numbers $c_1', \ldots, c_n'$ and $x_1', \ldots, x_n' \in X_0$. Clearly H_0' is countable. Let $g = \sum_{i=1}^n c_i K_{x_i}$ be a member of H_0. Then by (21)

$$||f - g|| = \sum_{i,j=1}^n \left(c_i' c_j' K(x_i', x_j') + c_i c_j K(x_i, x_j) - c_i' c_j K(x_i', x_j) - c_i c_j' K(x_i, x_j') \right).$$

For a given $c_1, \ldots, c_n$ and $x_1, \ldots, x_n$, this can be made as small as one likes by appropriate choice of $c_1', \ldots, c_n'$ and $x_1', \ldots, x_n'$, since K is continuous. Hence H_0' is dense in H_0, and thus H_0 is separable. Since H_0 is dense in H, it follows that H is separable.

Every separable Hilbert space H has a countable orthonormal basis, that is, a set $B = \{\phi_n\}_{n\geq 1} \subset H$ such that $\langle \phi_n, \phi_m \rangle = \delta(n,m)$ and for all $f \in H$ there exist unique real $\{c_n\}_{n\geq 1}$ such that $f = \sum_n c_n \phi_n$. Moreover, $c_n = \langle f, \phi_n \rangle$. If B is infinite, we say H has infinite dimension. In that case, H is isometric to the space l^2 of all infinite real sequences $\{c_n\}_{n\geq 1}$ with $\sum_n c_n^2 < \infty$. Otherwise, the dimension of H is the size of B, and H is isometric with a Euclidean space of this dimension.

We can now state the main result.

Theorem 5 *For any continuous kernel K on $X \times X$, where X is a separable metric space, the associated RKHS H is separable and is thus isometric with either $\Re^N$ for some finite N or with l^2. In either case, H has an orthonormal basis $\{\phi_n\}_{n\geq 1}$ such that for every $x, y \in X$,*

$$K(x,y) = \sum_n \phi_n(x)\phi_n(y).$$

Proof: The fact that H is separable and isometric with either $\Re^N$ of l^2, and that it has an orthonormal basis $B = \{\phi_n\}_{n\geq 1}$ follows from the above discussion. By definition $K(x,y) = K_x(y)$, where $K_x \in H_0$. Since B is an orthonormal basis for H, which contains H_0,

$$K_x(y) = \sum_n \langle K_x, \phi_n \rangle \phi_n(y). \tag{26}$$

It follows from the reproducing property of H that

$$\langle K_x, \phi_n \rangle = \phi_n(x). \tag{27}$$

This establishes the result. □

We are now in a position to prove that parts 4 and 5 of Theorem 4 in Section 6 are equivalent. First, suppose $d^2(x,y)$ is finite, negative definite, and zero on the diagonal. Let $K'(x,y) = (1/2)(d^2(x,z) + d^2(y,z) - d^2(x,y))$ for some z. K' is a kernel by Lemma 3. Let H be the RKHS associated with K' and $K'_x(y) = K'(x,y)$. Then

$$||K'_x - K'_y||^2 = K'(x,x) + K'(y,y) - 2K'(x,y) = d^2(x,y).$$

Hence $\phi(x) = K'_x$ is an isometric embedding of (X,d) into H. Since there is an isometric embedding of H into l_2 by the above result, it follows that (X,d)

can be isometrically embedded in l_2. If $d^2(x,y)$ is negative definite and is infinite for some $x \neq y$, then as discussed in Section 6, X can be decomposed into equivalence classes such that $d^2(x,y)$ is finite on each equivalence class and infinite between equivalence classes. It follows that 4 implies 5.

To see that 5 implies 4, first note that, as remarked in Section 6, whenever X can be decomposed into equivalence classes such that $d^2(x,y)$ is finite on each equivalence class and infinite between equivalence classes, and such that $d^2(x,y)$ is negative definite on each equivalence class, then $d^2(x,y)$ is negative definite on all of X. It is trivial to verify that if (S,d) can be isometrically embedded in l_2, then d^2 is a negative definite function on $S \times S$. The result follows. $\Box$

Finally, the reproducing kernel Hilbert space H associated with the kernel K also gives us a simple **Proof of Lemma** 1: Let K be a kernel on a set $U \times U$ and for all finite, nonempty $A, B \subseteq U$ define $K'(A,B) = \sum_{u \in A, v \in B} K(u,v)$. We must show that K' is a kernel on the product of the set of all finite, nonempty subsets of U with itself. For any nonempty finite subset $A \subseteq U$, let $f_A = \sum_{u \in A} K_u \in H_0$, where H_0 is the pre-Hilbert space associated with K'. If for nonempty finite $A, B \subseteq U$, we define $K'(A,B) = \sum_{u \in A, v \in B} K(u,v)$, then by Equation (21), $K'(A,B) = \langle f_A, f_B \rangle$. Since an inner product is a kernel, it follows that K' is a kernel on the product of the set of all nonempty finite subsets of U with itself. $\Box$

This construction suggests a broad generalization of the notion of an R-convolution as well. Let $U = X_1 \times \cdots \times X_D$. For kernels K_d on $X_d \times X_d$, let $\tilde{K} = K_1 \otimes \cdots \otimes K_D$ be a kernel on $U \times U$. Let $W : U \times X \to \Re$. We say W *represents the relation* R if $W(u,x) = 1$ when $R(u,x)$ and $W(u,x) = 0$ else. If $X_1, \ldots, X_D$ are countable, then we can define the W-convolution by

$$K(x,y) = \sum_{u,v \in U} W(u,x) W(v,y) \tilde{K}(u,v) = \langle f_x, f_y \rangle ,$$

where $f_x(v) = \sum_{u \in U} W(u,x) \tilde{K}_u(v)$ and $\tilde{K}_u(v) = \tilde{K}(u,v)$. This is a well-defined kernel if f_x is in the RKHS associated with the kernel $\tilde{K}$ for all $x \in X$. It is clear that if W represents the finite relation R, then f_x is in the RKHS associated with the kernel $\tilde{K}$, and the W-convolution is the same as the R-convolution. If $X_1, \ldots, X_D$ and X are uncountable, we can in some cases use an integral representation of the element f_x of the RKHS in place of the sum, and modify the definition of the W-convolution accordingly. This

should lead to other interesting connections between this theory and some of the more classical uses of kernels.

Acknowledgments

I'd like to thank Tommi Jaakkola for providing the main inspiration for this research, and for his valuable ideas in the early stages. Manfred Opper and Richard Montgomery gave helpful advice on the material treated in Sections 5 and 7, and Andrzej Ehrenfeucht, Vincent Mirelli and Bill Grundy also provided other good suggestions. This research was supported partly by a grant from the Army Research Office, and partly by grants from the National Science Foundation and the Department of Energy.

References

[1] C. Berg, J. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions.* Springer, 1984.

[2] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[3] D. Dudley. *Real Analysis and Probability.* Wiley, 1989.

[4] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, 1998.

[5] R. Elliott, L. Aggoun, and J. Moore. *Hidden Markov Models, Estimation and Control.* Springer Verlag, 1995.

[6] William Feller. *An Introduction to Probability Theory and its Applications*, volume 1. John Wiley, 1971.

[7] William Feller. *An Introduction to Probability Theory and its Applications*, volume 2. John Wiley, 1971.

[8] C. Fitzgerald and R. Horn. On fractional hadamard powers of positive definite matrices. *Journal of Mathematical Analysis and Applications*, 61:633–642, 1977.

[9] K. S. Fu. *Syntactic pattern recognition and applications*. Prentice-Hall, Englewood Cliffs, NJ, 1982.

[10] King-Sun Fu and Taylor L. Booth. Grammatical inference: Introduction and survey – part i. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-5(1):95–111, January 1975.

[11] King-Sun Fu and Taylor L. Booth. Grammatical inference: Introduction and survey – part ii. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-5(4):409–423, July 1975.

[12] S Geman and D Geman. Stochastic relaxations, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6:721–742, 1984.

[13] F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6):1455–1480, 1998.

[14] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, Reading, Massachussetts, 1979.

[15] T. Jaakkola, M. Diekhans, and D. Haussler. Using the Fisher kernel method to detect remote protein homologies. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, Aug 1999.

[16] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*, San Mateo, CA, 1998. Morgan Kauffmann Publishers. To appear.

[17] T. Jaakkola and D. Haussler. Probabilistic kernel regression models. In *Proc. of the Seventh Int. Workshop on AI and Statistics*, 1998. To appear.

[18] S. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[19] M. Linial, N. Linial, N. Tishby, and G. Yona. Global self-organization of all known protein sequences reveals inherent biological signatures. *jmb*, 268(2):539–556, 1997.

[20] N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995.

[21] D. J. C. MacKay. Introduction to gaussian processes, 1997. Available from `http://wol.ra.phy.cam.ac.uk/mackay/`.

[22] L. Máté. *Hilbert Space Methods in Science and Engineering.* Adam Hilger, 1989.

[23] T. Poggio and F. Girosi. A sparse representation for function approximation. *Neural Computation*, 10(6):1445–1454, 1998.

[24] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–286, February 1989.

[25] L. Rosen. Positive powers of positive positive definite matrices. *Canadian Journal of Mathematics*, 48:196–209, 1996.

[26] David Sankoff and Joseph B. Kruskal. *Time warps, string edits, and macromolecules : the theory and practice of sequence comparison.* Addison-Wesley, 1983.

[27] B. Schölkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing*, 45(11):2758–2765, November 1997.

[28] C. Scholkopf, J.C. Burges, and A.J. Smola. *Advances in Kernel Methods: Support Vector Learning.* MIT Press, 1999.

[29] Gilbert Strang. *Linear Algebra and its Applications.* Harcourt Brace Jovanovich, 1986.

[30] C. Suanders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML98)*, 1998. available at: http://lara.enm.bris.ac.uk/cig/pubs_nf.htm.

[31] V. Vapnik. *Statistical Learning Theory.* Wiley, 1998.

[32] G. Wahba. *Spline models for observational data.* CBMS-NSF Regional Conference Series in Applied Mathematics, 1990.

[33] G. Wahba. Support Vector Machines, Reproducing Kernel Hilbert Spaces and the Randomized GACV. Technical Report TR984rr, University of Wisconsin at Madison, 1997.