Positive Definite Rational Kernels

Corinna Cortes, Patrick Haffner, and Mehryar Mohri

AT&T Labs – Research
180 Park Avenue, Florham Park, NJ 07932, USA
{corinna, haffner, mohri}@research.att.com

Abstract. Kernel methods are widely used in statistical learning techniques. We recently introduced a general kernel framework based on weighted transducers or rational relations, *rational kernels*, to extend kernel methods to the analysis of variable-length sequences or more generally weighted automata. These kernels are efficient to compute and have been successfully used in applications such as spoken-dialog classification. Not all rational kernels are *positive definite and symmetric* (PDS) however, a sufficient property for guaranteeing the convergence of discriminant classification algorithms such as Support Vector Machines. We present several theoretical results related to PDS rational kernels. We show in particular that under some conditions these kernels are closed under sum, product, or Kleene-closure and give a general method for constructing a PDS rational kernel from an arbitrary transducer defined on some non-idempotent semirings. We also show that some commonly used string kernels or similarity measures such as the edit-distance, the convolution kernels of Haussler, and some string kernels used in the context of computational biology are specific instances of rational kernels. Our results include the proof that the edit-distance over a non-trivial alphabet is not *negative definite*, which, to the best of our knowledge, was never stated or proved before.

1 Motivation

Many classification algorithms were originally designed for fixed-length vectors. Recent applications in text and speech processing and computational biology require however the analysis of variable-length sequences and even more generally weighted automata. Indeed, the output of a large-vocabulary speech recognizer for a particular input speech utterance, or that of a complex information extraction system combining several information sources for a specific input query, is typically a weighted automaton compactly representing a large set of alternative sequences. The weights assigned by the system to each sequence are used to rank different alternatives according to the models the system is based on. The error rate of such complex systems is still too high in many tasks to rely only on their one-best output, thus it is preferable instead to use the full output weighted automata which contain the correct result in most cases.

Kernel methods [13] are widely used in statistical learning techniques such as Support Vector Machines (SVMs) [2, 4, 14] due to their computational efficiency in high-dimensional feature spaces. Recently, a general kernel framework

Semiring	Set	$\oplus$	$\otimes$	$\overline{0}$	$\overline{1}$
Boolean	$\{0,1\}$	$\vee$	$\wedge$	0	1
Probability	$\mathbb{R}_+$	$+$	$\times$	0	1
Log	$\mathbb{R} \cup \{-\infty, +\infty\}$	$\oplus_{\log}$	$+$	$+\infty$	0
Tropical	$\mathbb{R} \cup \{-\infty, +\infty\}$	min	$+$	$+\infty$	0

Table 1. *Semiring examples.* $\oplus_{\log}$ *is defined by:* $x \oplus_{\log} y = -\log(e^{-x} + e^{-y})$.

based on weighted transducers or rational relations, *rational kernels*, was introduced to extend kernel methods to the analysis of variable-length sequences or more generally weighted automata [3]. It was shown that there are general and efficient algorithms for computing rational kernels. Rational kernels have been successfully used for applications such as spoken-dialog classification.

Not all rational kernels are *positive definite and symmetric* (PDS), or equivalently verify the Mercer condition [1], a condition that guarantees the convergence of discriminant classification algorithms such as SVMs. This motivates the study undertaken in this paper. We present several theoretical results related to PDS rational kernels. In particular, we show that under some conditions these kernels are closed under sum, product, or Kleene-closure and give a general method for constructing a PDS rational kernel from an arbitrary transducer defined on some non-idempotent semirings. We also study the relationship between rational kernels and some commonly used string kernels or similarity measures such as the edit-distance, the convolution kernels of Haussler [6], and some string kernels used in the context of computational biology [8]. We show that these kernels are all specific instances of rational kernels. In each case, we explicitly describe the corresponding weighted transducer. These transducers are often simple and efficient for computing kernels. Their diagram often provides more insight into the definition of kernels and can guide the design of new kernels. Our results also include the proof of the fact that the edit-distance over a non-trivial alphabet is not *negative definite*, which, to the best of our knowledge, was never stated or proved before.

2 Preliminaries

In this section, we present the algebraic definitions and notation necessary to introduce rational kernels.

Definition 1 ([7]). *A system* $(\mathbb{K}, \oplus, \otimes, \overline{0}, \overline{1})$ *is a* semiring *if:* $(\mathbb{K}, \oplus, \overline{0})$ *is a commutative monoid with identity element* $\overline{0}$*;* $(\mathbb{K}, \otimes, \overline{1})$ *is a monoid with identity element* $\overline{1}$*;* $\otimes$ *distributes over* $\oplus$*; and* $\overline{0}$ *is an annihilator for* $\otimes$*: for all* $a \in \mathbb{K}, a \otimes \overline{0} = \overline{0} \otimes a = \overline{0}$.

Thus, a semiring is a ring that may lack negation. Table 1 lists some familiar semirings.

Definition 2. *A* weighted finite-state transducer T *over a semiring* $\mathbb{K}$ *is an 8-tuple* $T = (\Sigma, \Delta, Q, I, F, E, \lambda, \rho)$ *where:* Σ *is the finite input alphabet of the* *transducer; Δ is the finite output alphabet; Q is a finite set of states; $I \subseteq Q$ the set of initial states; $F \subseteq Q$ the set of final states; $E \subseteq Q \times (\Sigma \cup \{\epsilon\}) \times (\Delta \cup \{\epsilon\}) \times \mathbb{K} \times Q$ a finite set of transitions; $\lambda : I \rightarrow \mathbb{K}$ the initial weight function; and $\rho : F \rightarrow \mathbb{K}$ the final weight function mapping F to $\mathbb{K}$.*

Weighted automata can be formally defined in a similar way by simply omitting the input or output labels.

Given a transition $e \in E$, we denote by $p[e]$ its origin or previous state and $n[e]$ its destination state or next state, and $w[e]$ its weight. A *path* $\pi = e_1 \cdots e_k$ is an element of E^* with consecutive transitions: $n[e_{i-1}] = p[e_i]$, $i = 2, \ldots, k$. We extend n and p to paths by setting: $n[\pi] = n[e_k]$ and $p[\pi] = p[e_1]$. The weight function w can also be extended to paths by defining the weight of a path as the $\otimes$-product of the weights of its constituent transitions: $w[\pi] = w[e_1] \otimes \cdots \otimes w[e_k]$. We denote by $P(q, q')$ the set of paths from q to q' and by $P(q, x, y, q')$ the set of paths from q to q' with input label $x \in \Sigma^*$ and output label y. These definitions can be extended to subsets $R, R' \subseteq Q$, by: $P(R, x, y, R') = \cup_{q \in R,\, q' \in R'} P(q, x, y, q')$. A transducer T is *regulated* if the output weight associated by T to any pair of input-output string (x, y) by:

$$[\![T]\!](x, y) = \bigoplus_{\pi \in P(I, x, y, F)} \lambda(p[\pi]) \otimes w[\pi] \otimes \rho[n[\pi]] \tag{1}$$

is well-defined and in $\mathbb{K}$. $[\![T]\!](x, y) = \overline{0}$ when $P(I, x, y, F) = \emptyset$. If for all $q \in Q$ $\bigoplus_{\pi \in P(q, \epsilon, \epsilon, q)} w[\pi] \in \mathbb{K}$, then T is regulated. In particular, when T does not have any ϵ-cycle, it is regulated. In the following, we will assume that all the transducers considered are regulated. Regulated weighted transducers are closed under $\oplus$, $\otimes$ and Kleene-closure. For any transducer T, we denote by T^{-1} its *inverse*, that is the transducer obtained from T by transposing the input and output labels of each transition. The *composition* of two weighted transducers T_1 and T_2 is a weighted transducer denoted by $T_1 \circ T_2$ when the sum:

$$[\![T_1 \circ T_2]\!](x, y) = \bigoplus_{z \in \Sigma^*} [\![T_1]\!](x, z) \otimes [\![T_2]\!](z, y) \tag{2}$$

is well-defined and in $\mathbb{K}$ for all $x \in \Sigma^*$ and $y \in \Delta^*$ [7].

3 Rational Kernels - Definition

Definition 3. *A kernel K is said to be* rational *if there exist a weighted transducer $T = (\Sigma, \Delta, Q, I, F, E, \lambda, \rho)$ over the semiring $\mathbb{K}$ and a function $\psi : \mathbb{K} \rightarrow \mathbb{R}$ such that for all $x \in \Sigma^*$ and $y \in \Delta^*$:*

$$K(x, y) = \psi([\![T]\!](x, y)) \tag{3}$$

This definition and many of the results presented in this paper can be generalized by replacing the free monoids Σ^* and Δ^* with arbitrary monoids M_1 and M_2.

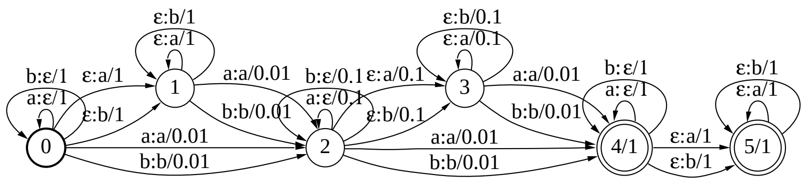


Fig. 1. Gappy bigram rational kernel with decay factor $\lambda = .1$. Bold face circles represent initial states and double circles indicate final states.

Also, note that we are not making any particular assumption about the function ψ in this definition. In general, it is an arbitrary function mapping $\mathbb{K}$ to $\mathbb{R}$.

Figure 1 shows an example of a transducer over the probability semiring corresponding to the gappy n-gram kernel with decay factor λ as defined by [10]. Such gappy n-gram kernels are rational kernels [3].

Rational kernels can be naturally extended to kernels over weighted automata. Let A be a weighted automaton defined over the semiring $\mathbb{K}$ and the alphabet Σ and B a weighted automaton defined over the semiring $\mathbb{K}$ and the alphabet Δ, $K(A, B)$ is defined by:

$$K(A,B) = \psi\left(\bigoplus_{(x,y)\in\Sigma^*\times\Delta^*} [\![A]\!](x)\otimes[\![T]\!](x,y)\otimes[\![B]\!](y)\right) \tag{4}$$

for all weighted automata A and B such that the $\oplus$-sum:

$$\bigoplus_{(x,y)\in\Sigma^*\times\Delta^*} [\![A]\!](x)\otimes[\![T]\!](x,y)\otimes[\![B]\!](y)$$

is well-defined and in $\mathbb{K}$. This sum is always defined and in $\mathbb{K}$ when A and B are acyclic weighted automata since the sum then runs over a finite set. It is defined for all weighted automata in all *closed semirings* [7] such as the tropical semiring. In the probability semiring, the sum is well-defined for all A, B, and T representing probability distributions. When $K(A, B)$ is defined, Equation 4 can be equivalently written as:

$$K(A,B) = \psi\left(\bigoplus_{(x,y)\in\Sigma^*\times\Delta^*} [\![A\circ T\circ B]\!](x,y)\right) \tag{5}$$

A general algorithm for computing rational kernels efficiently was given in [3]. It is based on the composition of weighted transducers and a general shortest-distance algorithm in a semiring $\mathbb{K}$.

In learning techniques such as those based on SVMs, we are particularly interested in kernels that are *positive definite symmetric* (PDS), or, equivalently, kernels verifying Mercer's condition, which guarantee the existence of a Hilbert space and a dot product associated to the kernel considered. Thus, in what follows, we will focus on theoretical results related to the construction of rational kernels that are PDS. Due to the symmetry condition, the input and output alphabets Σ and Δ will coincide in the rest of the paper.

4 Theoretical Results on Positive Definite Rational Kernels

This section reviews a number of results related to PDS kernels and extends them to *PDS rational kernels*, that is the class of rational kernels that have the Mercer property [1]. These results can be used to combine PDS rational kernels to design new PDS rational kernels or to determine if a rational kernel is PDS.

Definition 4. *Let X be a non-empty set. A function $K : X \times X \to \mathbb{R}$ is said to be a PDS kernel if it is symmetric ($K(x,y) = K(y,x)$ for all $x, y \in X$) and*

$$\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) \geq 0 \tag{6}$$

for all $n \geq 1$, $\{x_1, \ldots, x_n\} \subseteq X$ and $\{c_1, \ldots, c_n\} \subseteq \mathbb{R}$.

It is clear from classical results of linear algebra that K is a PDS kernel iff the matrix $K(x_i, x_j)_{i,j \leq n}$ for all $n \geq 1$ and all $\{x_1, \ldots, x_n\} \subseteq X$ is symmetric and all its eigenvalues are non-negative.

PDS kernels can be used to construct other families of kernels that also meet these conditions [13]. *Polynomial kernels* of degree p are formed from the expression $(K + a)^p$, and *Gaussian kernels* can be formed as $\exp(-d^2/\sigma^2)$ with $d^2(x,y) = K(x,x) + K(y,y) - 2K(x,y)$. The following sections will provide other ways of constructing PDS rational kernels.

4.1 General Closure Properties of PDS Kernels

The following theorem summarizes general closure properties of PDS kernels [1].

Theorem 1. *Let X and Y be two non-empty sets.*

1. Closure under sum*: Let $K_1, K_2 : X \times X \to \mathbb{R}$ be PDS kernels, then $K_1 + K_2 : X \times X \to \mathbb{R}$ is a PDS kernel.*
2. Closure under product*: Let $K_1, K_2 : X \times X \to \mathbb{R}$ be PDS kernels, then $K_1 \cdot K_2 : X \times X \to \mathbb{R}$ is a PDS kernel.*
3. Closure under tensor product*: Let $K_1 : X \times X \to \mathbb{R}$ and $K_2 : Y \times Y \to \mathbb{R}$ be PDS kernels, then their tensor product $K_1 \odot K_2 : (X \times Y) \times (X \times Y) \to \mathbb{R}$, where $K_1 \odot K_2((x_1, y_1), (x_2, y_2)) = K_1(x_1, x_2) \cdot K_2(y_1, y_2)$ is a PDS kernel.*
4. Closure under pointwise limit*: Let $K_n : X \times X \to \mathbb{R}$ be a PDS kernel for all $n \in \mathbb{N}$ and assume that $\lim_{n\to\infty} K_n(x,y)$ exists for all $x, y \in X$, then K defined by $K(x,y) = \lim_{n\to\infty} K_n(x,y)$ is a PDS kernel.*
5. Closure under composition with a power series*: Let $K : X \times X \to \mathbb{R}$ be a PDS kernel such that $|K(x,y)| < \rho$ for all $(x,y) \in X \times X$. Then if the radius of convergence of the power series $S = \sum_{n=0}^{\infty} a_n x^n$ is ρ and $a_n \geq 0$ for all $n \geq 0$, the composed kernel $S \circ K$ is a PDS kernel. In particular, if $K : X \times X \to \mathbb{R}$ is a PDS kernel, then so is $\exp(K)$.*

Clearly, these closure properties all apply to PDS rational kernels as well. In the next section, we present other closure properties more specific to the class of PDS rational kernels.

4.2 Closure Properties of PDS Rational Kernels

By definition, weighted transducers are closed under rational operations. The rational operations (sum, product, and closure operations) are defined as follows for all transducers T_1 and T_2 and $(x, y) \in \Sigma^* \times \Sigma^*$:

$$[\![T_1 \oplus T_2]\!](x,y) = [\![T_1]\!](x,y) \oplus [\![T_2]\!](x,y) \tag{7}$$

$$[\![T_1 \otimes T_2]\!](x,y) = \bigoplus_{x=x_1x_2, y=y_1y_2} [\![T_1]\!](x_1,y_1) \otimes [\![T_2]\!](x_2,y_2)$$

$$[\![T^*]\!](x,y) = \bigoplus_{n=0}^{\infty} T^n(x,y)$$

In this section, we assume that a fixed function ψ is used in the definition of all the rational kernels mentioned. We denote by K_T the rational kernel corresponding to the transducer T and defined for all $x, y \in \Sigma^*$ by $K_T(x,y) = \psi([\![T]\!](x,y))$.

Theorem 2. *Let Σ be a non-empty alphabet. The following closure properties hold for PDS rational kernels.*

1. Closure under $\oplus$-sum*: Assume that $\psi : (\mathbb{K}, \oplus, \overline{0}) \to (\mathbb{R}, +, 0)$ is a monoid morphism. Let $K_{T_1}, K_{T_2} : \Sigma^* \times \Sigma^* \to \mathbb{R}$ be PDS rational kernels, then $K_{T_1 \oplus T_2} : \Sigma^* \times \Sigma^* \to \mathbb{R}$ is a PDS rational kernel and $K_{T_1 \oplus T_2} = K_{T_1} + K_{T_2}$.*
2. Closure under $\otimes$-product*: Assume that $\psi : (\mathbb{K}, \oplus, \otimes, \overline{0}, \overline{1}) \to (\mathbb{R}, +, \times, 0, 1)$ is a semiring morphism. Let $K_{T_1}, K_{T_2} : \Sigma^* \times \Sigma^* \to \mathbb{R}$ be PDS rational kernels, then $K_{T_1 \otimes T_2} : \Sigma^* \times \Sigma^* \to \mathbb{R}$ is a PDS rational kernel.*
3. Closure under Kleene-closure*: Assume that $\psi : (\mathbb{K}, \oplus, \otimes, \overline{0}, \overline{1}) \to (\mathbb{R}, +, \times, 0, 1)$ is a continuous semiring morphism. Let $K_T : \Sigma^* \times \Sigma^* \to \mathbb{R}$ be a PDS rational kernel, then $K_{T^*} : \Sigma^* \times \Sigma^* \to \mathbb{R}$ is a PDS rational kernel.*

Proof. The closure under $\oplus$-sum follows directly Theorem 1 and the fact that for all $x, y \in \Sigma^*$:

$$\psi([\![T_1]\!](x,y) \oplus [\![T_2]\!](x,y)) = \psi([\![T_1]\!](x,y)) + \psi([\![T_2]\!](x,y))$$

when $\psi : (\mathbb{K}, \oplus, \overline{0}) \to (\mathbb{R}, +, 0)$ is a monoid morphism. For the closure under $\otimes$-product, when ψ is a semiring morphism, for all $x, y \in \Sigma^*$:

$$\begin{aligned} \psi([\![T_1 \otimes T_2]\!](x,y)) &= \sum_{x_1x_2=x, y_1y_2=y} \psi([\![T_1]\!](x_1,y_1)) \cdot \psi([\![T_2]\!](x_2,y_2)) \\ &= \sum_{x_1x_2=x, y_1y_2=y} K_{T_1} \odot K_{T_2}((x_1,x_2),(y_1,y_2)) \end{aligned} \tag{8}$$

By Theorem 1, since K_{T_1} and K_{T_2} are PDS kernels, their tensor product $K_{T_1} \odot K_{T_2}$ is a PDS kernel and there exists a Hilbert space $H \subseteq \mathbb{R}^{\Sigma^*}$ and a mapping $u \to \phi_u$ such that $K_{T_1} \odot K_{T_2}(u,v) = \langle \phi_u, \phi_v \rangle$ [1]. Thus

$$\begin{aligned} \psi([\![T_1 \otimes T_2]\!](x,y)) &= \sum_{x_1x_2=x, y_1y_2=y} \langle \phi_{(x_1,x_2)}, \phi_{(y_1,y_2)} \rangle \\ &= \left\langle \sum_{x_1x_2=x} \phi_{(x_1,x_2)}, \sum_{y_1y_2=y} \phi_{(y_1,y_2)} \right\rangle \end{aligned} \tag{9}$$

Since a dot product is positive definite, this equality implies that $K_{T_1 \otimes T_2}$ is a PDS kernel. The closure under Kleene-closure is a direct consequence of the closure under $\oplus$-sum and $\otimes$-product of PDS rational kernels and the closure under pointwise limit of PDS kernels (Theorem 1). $\square$

Theorem 2 provides a general method for constructing complex PDS rational kernels from simpler ones. PDS rational kernels defined to model specific prior knowledge sources can be combined to create a more general PDS kernel. In contrast to Theorem 2, PDS rational kernels are not closed under composition. This is clear since the ordinary matrix multiplication does not preserve positive definiteness in general.[1] The next section studies a general construction of PDS rational kernels using composition.

4.3 A General Construction of PDS Rational Kernels

In this section, we assume that $\psi : (\mathbb{K}, \oplus, \otimes, \overline{0}, \overline{1}) \to (\mathbb{R}, +, \times, 0, 1)$ is a continuous semiring morphism.[2] We show that there exists a general way of constructing a PDS rational kernel from any transducer T. The construction is based on composing T with its inverse T^{-1}. The composition of two weighted transducers T_1 and T_2 is a weighted transducer denoted by $T_1 \circ T_2$ and defined by:

$$[\![T_1 \circ T_2]\!](x, y) = \bigoplus_{z \in \Sigma^*} [\![T_1]\!](x, z) \otimes [\![T_2]\!](z, y) \tag{10}$$

Proposition 1. *Let $T = (\Sigma, \Delta, Q, I, F, E, \lambda, \rho)$ be a weighted transducer defined over $(\mathbb{K}, \oplus, \otimes, \overline{0}, \overline{1})$. Assume that the weighted transducer $T \circ T^{-1}$ is regulated, then $T \circ T^{-1}$ defines a PDS rational kernel over $\Sigma^* \times \Sigma^*$.*

Proof. Denote by S the composed transducer $T \circ T^{-1}$. Let K be the rational kernel defined by S. By definition of composition

$$K(x, y) = \psi([\![S]\!](x, y)) = \psi\left(\bigoplus_{z \in \Delta^*} [\![T]\!](x, z) \otimes [\![T]\!](y, z)\right) \tag{11}$$

for all $x, y \in \Sigma^*$. Since ψ is a continuous semiring morphism, for all $x, y \in \Sigma^*$

$$K(x, y) = \psi([\![S]\!](x, y)) = \sum_{z \in \Delta^*} \psi([\![T]\!](x, z)) \cdot \psi([\![T]\!](y, z)) \tag{12}$$

For all $n \in \mathbb{N}$ and $x, y \in \Sigma^*$, define $K_n(x, y)$ by:

$$K_n(x, y) = \sum_{|z| \leq n} \psi([\![T]\!](x, z)) \cdot \psi([\![T]\!](y, z)) \tag{13}$$

[1] It is not difficult to prove however that the composition of two PDS transducers T_1 and T_2 is a PDS transducer when $T_1 \circ T_2 = T_2 \circ T_1$.

[2] In some cases such a morphism may not exist. Its existence implies among other properties that $\mathbb{K}$ is commutative and that $\mathbb{K}$ is non-idempotent. Indeed, if $\mathbb{K}$ is idempotent, for any $x \in \mathbb{K}$, $\psi(x) = \psi(x \oplus x) = \psi(x) + \psi(x) = 2\psi(x)$, which implies that $\psi(x) = 0$ for all x.

where the sum runs over all strings $z \in \Delta^*$ of length less than or equal to n. Clearly, K_n defines a symmetric kernel. For any $l \geq 1$ and any $x_1, \ldots, x_l \in \Sigma^*$, define the matrix M_n by: $M_n = (K_n(x_i, x_j))_{i \leq l, j \leq l}$. Let $z_1, z_2, \ldots, z_m$ be an arbitrary ordering of the strings of length less than or equal to n. Define the matrix A by:

$$A = (\psi(\llbracket T \rrbracket(x_i, z_j)))_{i \leq l, j \leq m} \tag{14}$$

By definition of K_n, $M_n = AA^t$. Thus, the eigenvalues of M_n are all non-negative, which implies that K_n is a PDS kernel. Since K is a pointwise limit of K_n, $K(x, y) = \lim_{n \to \infty} K_n(x, y)$, by Theorem 1, K is a PDS kernel. This ends the proof of the proposition. □

The next propositions provide results related to the converse of Proposition 1.

Proposition 2. *Let $S = (\Sigma, \Sigma, Q, I, F, E, \lambda, \rho)$ be an acyclic weighted transducer over $(\mathbb{K}, \oplus, \otimes, \overline{0}, \overline{1})$ defining a PDS rational kernel over $\Sigma^* \times \Sigma^*$, then there exists a weighted transducer T such that $S = T \circ T^{-1}$.*

Proof. The proof is based on the classical result of linear algebra that any positive definite (finite) matrix M can be written as $M = AA^t$ for some matrix A. The full proof of the proposition is reserved to a longer version of the paper. □

Assume that the same continuous semiring morphism ψ is used in the definition of all the rational kernels.

Proposition 3. *Let Θ be the subset of weighted transducers over $(\mathbb{K}, \oplus, \otimes, \overline{0}, \overline{1})$ defining a PDS rational kernel such that for any $S \in \Theta$ there exists a weighted transducer T such that $S = T \circ T^{-1}$. Then Θ is closed under $\oplus$-sum, $\otimes$-product, and Kleene-closure.*

Proof. The proof is based on various technical arguments related to the composition of weighted transducers and is left to a longer version of the paper. □

Proposition 1 leads to a natural question: under the same assumptions, are all weighted transducers S defining a PDS rational kernel of the form $S = T \circ T^{-1}$? We conjecture that this is the case and that this property provides a characterization of the weighted transducers defining PDS rational kernels under the assumptions made in Proposition 1. Indeed, we have not (yet) found a counter-example contradicting this statement and have proved a number of results in support of it, including the two propositions above.

4.4 Negative Definite Kernels

As mentioned before, given a set X and a distance or dis-similarity measure $d : X \times X \to \mathbb{R}_+$, a common method used to define a kernel K is the following. For all $x, y \in X$,

$$K(x, y) = \exp(-t d^2(x, y)) \tag{15}$$

where $t > 0$ is some constant typically used for normalization. Gaussian kernels are defined in this way. However, such kernels K are not necessarily positive definite, e.g., for $X = \mathbb{R}$, $d(x, y) = |x - y|^p$, $p > 1$ and $t = 1$, K is not positive definite. The positive definiteness of K depends on t and the properties of the function d. The classical results presented in this section exactly address such questions [1]. They include a characterization of PDS kernels based on *negative definite kernels* which may be viewed as distances with some specific properties.[3]

The results we are presenting are general, but we are particularly interested in the case where d can be represented by a rational kernel. We will use these results later when dealing with the case of the edit-distance.

Definition 5. *Let X be a non-empty set. A function $K : X \times X \to \mathbb{R}$ is said to be a* negative definite symmetric kernel *(NDS kernel) if it is symmetric ($K(x, y) = K(y, x)$ for all $x, y \in X$) and*

$$\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) \leq 0 \tag{16}$$

for all $n \geq 1$, $\{x_1, \ldots, x_n\} \subseteq X$ and $\{c_1, \ldots, c_n\} \subseteq \mathbb{R}$ with $\sum_{i=1}^{n} c_i = 0$.

Clearly, if K is a PDS kernel then $-K$ is a NDS kernel, however the converse does not hold in general. Negative definite kernels often correspond to distances, e.g., $K(x, y) = (x - y)^\alpha$, with $0 < \alpha \leq 2$ is a negative definite kernel.

The next theorem summarizes general closure properties of NDS kernels [1].

Theorem 3. *Let X be a non-empty set.*

1. Closure under sum*: Let $K_1, K_2 : X \times X \to \mathbb{R}$ be NDS kernels, then $K_1 + K_2 : X \times X \to \mathbb{R}$ is a NDS kernel.*
2. Closure under log and exponentiation*: Let $K : X \times X \to \mathbb{R}$ be a NDS kernel with $K \geq 0$, and α a real number with $0 < \alpha < 1$, then $\log(1 + K), K^\alpha : X \times X \to \mathbb{R}$ are NDS kernels.*
3. Closure under pointwise limit*: Let $K_n : X \times X \to \mathbb{R}$ be a NDS kernel for all $n \in \mathbb{N}$, then K defined by $K(x, y) = \lim_{n\to\infty} K_n(x, y)$ is a NDS kernel.*

The following theorem clarifies the relation between NDS and PDS kernels and provides in particular a way of constructing PDS kernels from NDS ones [1].

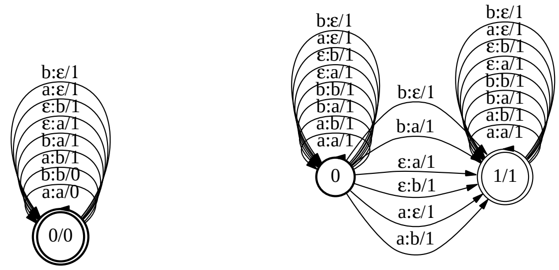
Theorem 4. *Let X be a non-empty set, $x_o \in X$, and let $K : X \times X \to \mathbb{R}$ be a symmetric kernel.*

1. *K is negative definite iff $\exp(-tK)$ is positive definite for all $t > 0$.*
2. *Let K' be the function defined by:*

$$K'(x, y) = K(x, x_0) + K(y, x_0) - K(x, y) - K(x_0, x_0) \tag{17}$$

Then K is negative definite iff K' is positive definite.

[3] Many of the results described by [1] are also included in [12] with the terminology of *conditionally positive definite* instead of *negative definite kernels*. We adopt the original terminology used by [1].



(a) (b)

Fig. 2. (a) Weighted transducer over the tropical semiring representing the edit-distance over the alphabet $\Sigma = \{a, b\}$. (b) Weighted transducer over the probability semiring computing the cost of alignments over the alphabet $\Sigma = \{a, b\}$.

The theorem gives two ways of constructing a positive definite kernel using a negative definite kernel. The first construction is similar to the way Gaussian kernels are defined. The second construction has been put forward by [12].

5 Relationship with some commonly used kernels or similarity measures

This section studies the relationships between several families of kernels or similarities measures and rational kernels.

5.1 Edit-Distance

A common similarity measure in many applications is that of the *edit-distance* [9]. We denote by $d_e(x, y)$ the edit-distance between two strings x and y over the alphabet Σ with cost 1 assigned to all edit operations.

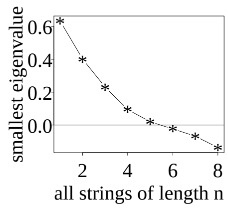
Proposition 4. *Let Σ be a non-empty finite alphabet and let d_e be the edit-distance over Σ, then d_e is a symmetric rational kernel. Furthermore, (1): d_e is not a PDS kernel, and (2): d_e is a NDS kernel iff $|\Sigma| = 1$.*

Proof. The edit-distance between two strings, or weighted automata, can be represented by a simple weighted transducer over the tropical semiring [11]. Since the edit-distance is symmetric, this shows that d_e is a symmetric rational kernel. Figure 2(a) shows the corresponding transducer when the alphabet is $\Sigma = \{a, b\}$. The cost of the alignment between two sequences can also be computed by a weighted transducer over the probability semiring [11], see Figure 2(b).

Let $a \in \Sigma$, then the matrix $(d_e(x_i, x_j))_{1 \leq i,j \leq 2}$ with $x_1 = \epsilon$ and $x_2 = a$ has a negative eigenvalue (-1), thus d_e is not a PDS kernel.

When $|\Sigma| = 1$, the edit-distance simply measures the absolute value of the difference of length between two strings. A string $x \in \Sigma^*$ can then be viewed as a vector of the Hilbert space $\mathbb{R}^\infty$. Denote by $\|\cdot\|$ the corresponding norm. For all $x, y \in \Sigma^*$:

$$d_e(x, y) = \|x - y\|$$



(a)

i	1	2	3	4	5
x_i	abc	bad	dab	adc	bcd
c_i	1	1	$-\frac{2}{3}$	$-\frac{2}{3}$	$-\frac{2}{3}$

(b)

Fig. 3. (a) Smallest eigenvalue of the matrix $M_n = (\exp(-d_e(x_i, x_j)))_{1 \leq i,j, \leq 2^n}$ as a function of n. (b) Example demonstrating that the edit-distance is not negative definite.

The square distance $\| \cdot \|^2$ is negative definite, thus by Theorem 3, $d_e = (\| \cdot \|^2)^{1/2}$ is also negative definite.

Assume now that $|\Sigma| > 1$. We show that $\exp(-d_e)$ is not PDS. By theorem 4, this implies that d_e is not negative definite. Let $x_1, \cdots, x_{2^n}$ be any ordering of the strings of length n over the alphabet $\{a, b\}$. Define the matrix M_n by:

$$M_n = (\exp(-d_e(x_i, x_j)))_{1 \leq i,j, \leq 2^n} \tag{18}$$

Figure 3(a) shows the smallest eigenvalue α_n of M_n as a function of n. Clearly, there are values of n for which $\alpha_n < 0$, thus the edit-distance is not negative definite. Table 3(b) provides a simple example with five strings of length 3 over the alphabet $\Sigma = \{a, b, c, d\}$ showing directly that the edit-distance is not negative definite. Indeed, it is easy to verify that: $\sum_{i=1}^{5} \sum_{j=1}^{5} c_i c_j K(x_i, x_j) = \frac{2}{3} > 0$. $\square$

To our knowledge, this is the first statement and proof of the fact that d_e is not NDS for $|\Sigma| > 1$. This result has a direct consequence on the design of kernels in computational biology, often based on the edit-distance or other related similarity measures. When $|\Sigma| > 1$, Proposition 4 shows that d_e is not NDS. Thus, there exists $t > 0$ for which $\exp(-td_e)$ is not PDS. Similarly, d_e^2 is not NDS since otherwise by Theorem 3, $d_e = (d_e^2)^{1/2}$ would be NDS.

5.2 Haussler's Convolution Kernels for Strings

D. Haussler describes a class of kernels for strings built by applying iteratively *convolution kernels* [6]. We show that these convolution kernels for strings are specific instances of rational kernels. To define these kernels, Haussler introduces for $0 \leq \gamma < 1$ the γ-infinite iteration of a mapping $H : \Sigma^* \times \Sigma^* \to \mathbb{R}$ by:

$$H_\gamma^* = (1 - \gamma) \sum_{n=1}^{\infty} \gamma^{n-1} H^{(n)} \tag{19}$$

where $H^{(n)} = H \star H^{(n-1)}$ is the result of the convolution of H with itself $n - 1$ times. Note that $H_\gamma^* = 0$ for $\gamma = 0$.

Lemma 1. *For $0 < \gamma < 1$, the γ-infinite iteration of a rational transduction $H : \Sigma^* \times \Sigma^* \to \mathbb{R}$ can be defined in the following way with respect to the Kleene $\dagger$-operator:*

$$H_\gamma^* = \frac{1 - \gamma}{\gamma} (\gamma H)^\dagger \tag{20}$$

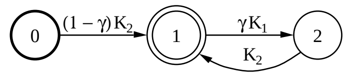


Fig. 4. Haussler's convolution kernels K_H for strings: specific instances of rational kernels. K_1, (K_2), corresponds to a specific weighted transducer over the probability semiring and modeling substitutions (resp. insertions).

Proof. Haussler's convolution simply corresponds to the Cauchy product or concatenation in the case of rational transductions. Thus, for $0 < \gamma < 1$, by definition of the †-operator:

$$(\gamma H)^\dagger = \sum_{n=1}^{\infty} (\gamma H)^n = \sum_{n=1}^{\infty} \gamma^n H^n = \frac{\gamma}{1-\gamma} \sum_{n=1}^{\infty} (1-\gamma)\gamma^{n-1} H^n = \frac{\gamma}{1-\gamma} H^*_\gamma \qquad \square$$

Given a probability distribution p over all symbols of Σ, Haussler's convolution kernels for strings are defined by:

$$K_H(x,y) = \gamma K_2 \star (K_1 \star K_2)^\star_\gamma + (1-\gamma) K_2$$

where K_1 is the specific polynomial PDS rational transduction over the probability semiring defined by: $K_1(x,y) = \sum_{a \in \Sigma} p(x|a)p(y|a)p(a)$ and models substitutions, and K_2 another specific PDS rational transduction over the probability semiring modeling insertions.

Proposition 5. *For any $0 \leq \gamma < 1$, Haussler's convolution kernels K_H coincide with the following special cases of rational kernels:*

$$K_H = (1-\gamma)[K_2(\gamma K_1 K_2)^*] \tag{21}$$

Proof. As mentioned above, Haussler's convolution simply corresponds to concatenation in this context. When $\gamma = 0$, by definition, K_H is reduced to K_2 which is a rational transducer and the proposition's formula above is satisfied. Assume now that $\gamma \neq 0$. By lemma 1, K_H can be re-written as:

$$\begin{aligned} K_H &= \gamma K_2 (K_1 K_2)^\star_\gamma + (1-\gamma) K_2 = \gamma K_2 \frac{1-\gamma}{\gamma} (\gamma K_1 K_2)^\dagger + (1-\gamma) K_2 \qquad (22) \\ &= (1-\gamma)[K_2(\gamma K_1 K_2)^\dagger + K_2] = (1-\gamma)[K_2(\gamma K_1 K_2)^*] \end{aligned}$$

Since rational transductions are closed under rational operations, K_H also defines a rational transduction. Since K_1 and K_2 are PDS kernels, by theorem 2, K_H defines a PDS kernel. $\square$

The transducer of Figure 4 illustrates the convolution kernels for strings proposed by Haussler. They correspond to special cases of rational kernels whose mechanism is clarified by the figure: the kernel corresponds to a substitution with weight $(1-\gamma)$ modeled by K_2 followed by any number of sequences of insertions modeled by K_1 and substitutions modeled by K_2 with weight γ. Clearly, there are many other ways of defining kernels based on weighted transducers with more complex definitions and perhaps more data-driven definitions.

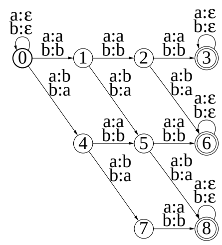


Fig. 5. Mismatch kernel $K_{(k,m)} = T_{k,m} \circ T_{k,m}^{-1}$ [8] with $k = 3$ and $m = 2$ and with $\Sigma = \{a, b\}$. The transducer $T_{3,2}$ defined over the probability semiring is shown. All transition weights and final weights are equal to one. Note that states 3, 6, and 8 of the transducer are equivalent and thus can be merged and similarly that states 2 and 5 can then be merged as well.

5.3 Other Kernels Used in Computational Biology

In this section we show the relationship between rational kernels and another class of kernels used in computational biology.

A family of kernels, *mismatch string kernels*, was introduced by [8] for protein classification using SVMs. Let Σ be a finite alphabet, typically that of amino acids for protein sequences. For any two sequences $z_1, z_2 \in \Sigma^*$ of same length ($|z_1| = |z_2|$), we denote by $d(z_1, z_2)$ the total number of mismatching symbols between these sequences. For all $m \in \mathbb{N}$, we define the bounded distance d_m between two sequences of same length by:

$$d_m(z_1, z_2) = \begin{cases} 1 & \text{if } (d(z_1, z_2) \leq m) \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

and for all $k \in \mathbb{N}$, we denote by $F_k(x)$ the set of all factors of x of length k:

$$F_k(x) = \{z : x \in \Sigma^* z \Sigma^*, |z| = k\}$$

For any $k, m \in \mathbb{N}$ with $m \leq k$, a (k, m)-mismatch kernel $K_{(k,m)} : \Sigma^* \times \Sigma^* \to \mathbb{R}$ is the kernel defined over protein sequences $x, y \in \Sigma^*$ by:

$$K_{(k,m)}(x, y) = \sum_{z_1 \in F_k(x),\, z_2 \in F_k(y),\, z \in \Sigma^k} d_m(z_1, z)\, d_m(z, z_2) \tag{24}$$

Proposition 6. *For any $k, m \in \mathbb{N}$ with $m \leq k$, the (k, m)-mismatch kernel $K_{(k,m)} : \Sigma^* \times \Sigma^* \to \mathbb{R}$ is a PDS rational kernel.*

Proof. Let M, S, and D be the weighted transducers over the probability semiring defined by:

$$M = \sum_{a \in \Sigma} (a, a) \qquad S = \sum_{a \neq b} (a, b) \qquad D = \sum_{a \in \Sigma} (a, \epsilon) \tag{25}$$

M associates weight 1 to each pair of identical symbols of the alphabet Σ, S associates 1 to each pair of distinct or mismatching symbols, and D associates 1 to all pairs with second element ϵ.

For $i, k \in \mathbb{N}$ with $0 \leq i \leq k$, Define the *shuffle* of S^i and M^{k-i}, denoted by $S^i \sqcup\!\sqcup M^{k-i}$, as the the sum over all products made of factors S and M with exactly i factors S and $k-i$ factors M. As a finite sum of products of S and M, $S^i \sqcup\!\sqcup M^{k-i}$ is rational. Since weighted transducers are closed under rational operations the following defines a weighted transducer T over the probability semiring for any $k, m \in \mathbb{N}$ with $m \leq k$: $T_{k,m} = D^* R D^*$ with $R = \sum_{i=0}^{m} S^i \sqcup\!\sqcup M^{k-i}$. Consider two sequences z_1, z_2 such that $|z_1| = |z_2| = k$. By definition of M and S and the shuffle product, for any i, with $0 \leq i \leq m$,

$$[\![S^i \sqcup\!\sqcup M^{k-i}]\!](z_1, z_2) = \begin{cases} 1 & \text{if } (d(z_1, z_2) = i) \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

Thus,
$$\begin{aligned}[\![R]\!](z_1, z_2) &= \sum_{i=0}^{m} S^i \sqcup\!\sqcup M^{k-i}(z_1, z_2) = \begin{cases} 1 & \text{if } (d(z_1, z_2) \leq m) \\ 0 & \text{otherwise} \end{cases} \\ &= d_m(z_1, z_2)\end{aligned}$$

By definition of the product of weighted transducers, for any $x \in \Sigma^*$ and $z \in \Sigma^k$,

$$\begin{aligned} T_{k,m}(x, z) &= \sum_{x=uvw, z=u'v'w'} [\![D^*]\!](u, u')\ [\![R]\!](v, v')\ [\![D^*]\!](w, w') \\ &= \sum_{v \in F_k(x), z=v'} [\![R]\!](v, v') = \sum_{v \in F_k(x)} d_m(v, z) \end{aligned} \tag{27}$$

It is clear from the definition of $T_{k,m}$ that $T_{k,m}(x, z) = 0$ for all $x, z \in \Sigma^*$ with $|z| > k$. Thus, by definition of the composition of weighted transducer, for all $x, y \in \Sigma^*$

$$\begin{aligned} [\![T_{k,m} \circ {T_{k,m}}^{-1}]\!](x, y) &= \sum_{z_1 \in F_k(x),\, z_2 \in F_k(y),\, z \in \Sigma^*} d_m(z_1, z)\ d_m(z, z_2) \\ &= \sum_{z_1 \in F_k(x),\, z_2 \in F_k(y),\, z \in \Sigma^k} d_m(z_1, z)\ d_m(z, z_2) = K_{(k,m)}(x, y) \end{aligned} \tag{28}$$

By proposition 1, this proves that $K_{(k,m)}$ is a PDS rational kernel. □

Figure 5 shows $T_{3,2}$, a simple weighted transducer over the probability semiring that can be used to compute the mismatch kernel $K_{(3,2)} = T_{3,2} \circ {T_{3,2}}^{-1}$. Such transducers provide a compact representation of the kernel and are very efficient to use with the composition algorithm already described in [3]. The transitions of these transducers can be defined implicitly and expanded on-demand as needed for the particular input strings or weighted automata. This substantially reduces the space needed for their representation, e.g., a single transition with labels $x : y$, $x \neq y$ can be used to represent all transitions with similar labels $((a : b),\ a, b \in \Sigma$, with $a \neq b)$. Similarly, composition can also be performed on-the-fly. Furthermore, the transducer of Figure 5 can be made more compact since it admits several states that are equivalent.

6 Conclusion

In general, the transducer representation provides a very compact representation benefiting from existing and well-studied optimizations and leads to an efficient computation of rational kernels. It further avoids the design of special-purpose algorithms for the computation of the kernels covered by the framework of rational kernels. We gave the proof of several new and general properties related to PDS rational kernels. These results can be used to design a PDS rational kernel from simpler ones or from an arbitrary weighted transducer over an appropriate semiring, or from negative definite kernels. Rational kernels provide a unified framework for the design of computationally efficient kernels for strings or weighted automata. The framework includes in particular pair-HMM string kernels [5, 15], Haussler's convolution kernels for strings and other classes of string kernels introduced for computational biology.

References

1. Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic Analysis on Semigroups*. Springer-Verlag: Berlin-New York, 1984.
2. B. E. Boser, I. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop of Computational Learning Theory*, volume 5, pages 144–152, Pittsburg, 1992. ACM.
3. Corinna Cortes, Patrick Haffner, and Mehryar Mohri. Rational Kernels. In *NIPS 2002*, Vancouver, Canada, March 2003. MIT Press.
4. Corinna Cortes and Vladimir N. Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.
5. R. Durbin, S.R. Eddy, A. Krogh, and G.J. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge UK, 1998.
6. David Haussler. Convolution Kernels on Discrete Structures. Technical Report UCSC-CRL-99-10, University of California at Santa Cruz, 1999.
7. Werner Kuich and Arto Salomaa. *Semirings, Automata, Languages*. Number 5 in EATCS Monographs on Theoretical Computer Science. Springer-Verlag, 1986.
8. Christina Leslie, Eleazar Eskin, Jason Weston, and William Stafford Noble. Mismatch String Kernels for SVM Protein Classification. In *NIPS 2002*, Vancouver, Canada, March 2003. MIT Press.
9. Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics - Doklady*, 10:707–710, 1966.
10. Huma Lodhi, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *NIPS 2000*, pages 563–569. MIT Press, 2001.
11. Mehryar Mohri. Edit-Distance of Weighted Automata: General Definitions and Algorithms. *International Journal of Foundations of Computer Science*, 2003.
12. Bernhard Schölkopf. The Kernel Trick for Distances. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *NIPS 2001*, pages 301–307. MIT Press, 2001.
13. Bernhard Schölkopf and Alex Smola. *Learning with Kernels*. MIT Press: Cambridge, MA, 2002.
14. Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.
15. Chris Watkins. Dynamic alignment kernels. Technical Report CSD-TR-98-11, Royal Holloway, University of London, 1999.