Foundations of Machine Learning
Kernel Methods

Mehryar Mohri
Courant Institute and Google Research
mohri@cims.nyu.edu

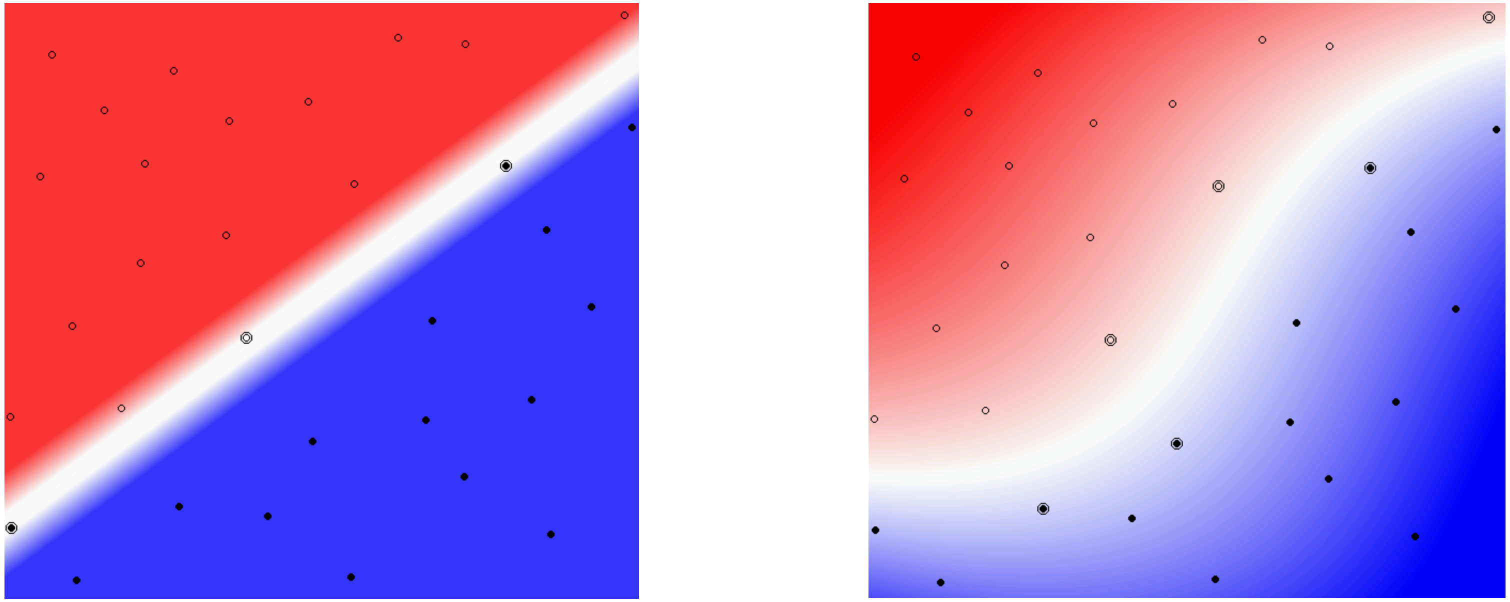
Motivation

- Efficient computation of inner products in high dimension.
- Non-linear decision boundary.
- Non-vectorial inputs.
- Flexible selection of more complex features.

This Lecture

- Kernels
- Kernel-based algorithms
- Closure properties
- Sequence Kernels
- Negative kernels

Non-Linear Separation



- Linear separation impossible in most problems.
- Non-linear mapping from input space to high-dimensional feature space: $\Phi\colon X \rightarrow F$.
- Generalization ability: independent of $\dim(F)$, depends only on margin and sample size.

Kernel Methods

- Idea:
  - Define $K: X \times X \rightarrow \mathbb{R}$, called kernel, such that:
  $$\Phi(x) \cdot \Phi(y) = K(x, y).$$
  - K often interpreted as a similarity measure.
- Benefits:
  - Efficiency: K is often more efficient to compute than Φ and the dot product.
  - Flexibility: K can be chosen arbitrarily so long as the existence of Φ is guaranteed (PDS condition or Mercer's condition).

PDS Condition

- Definition: a kernel $K\colon X\times X\to\mathbb{R}$ is positive definite symmetric (PDS) if for any $\{x_1, \ldots, x_m\}\subseteq X$, the matrix $\mathbf{K}=[K(x_i, x_j)]_{ij}\in\mathbb{R}^{m\times m}$ is symmetric positive semi-definite (SPSD).

- $\mathbf{K}$ SPSD if symmetric and one of the 2 equiv. cond.'s:
  - its eigenvalues are non-negative.
  - for any $\mathbf{c}\in\mathbb{R}^{m\times 1}$, $\mathbf{c}^\top\mathbf{K}\mathbf{c}=\sum_{i,j=1}^{m} c_i c_j K(x_i, x_j)\geq 0.$

- Terminology: PDS for kernels, SPSD for kernel matrices (see (Berg et al., 1984)).

Example - Polynomial Kernels

- Definition:

$$\forall x, y \in \mathbb{R}^N,\ K(x, y) = (x \cdot y + c)^d, \quad c > 0.$$

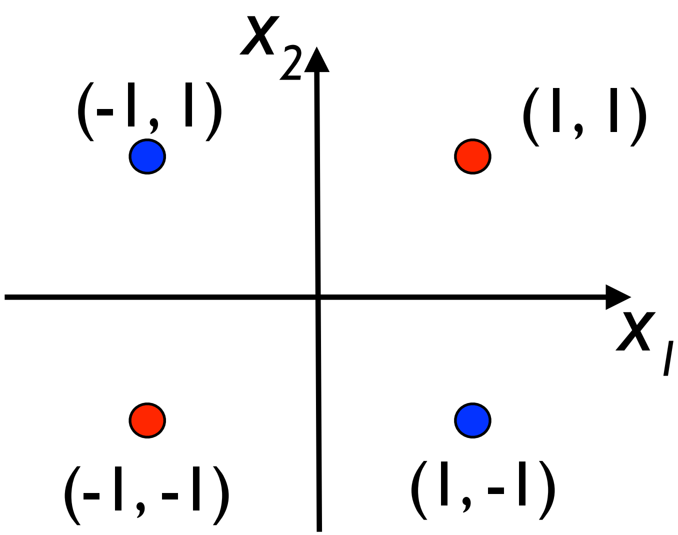
- Example: for $N = 2$ and $d = 2$,

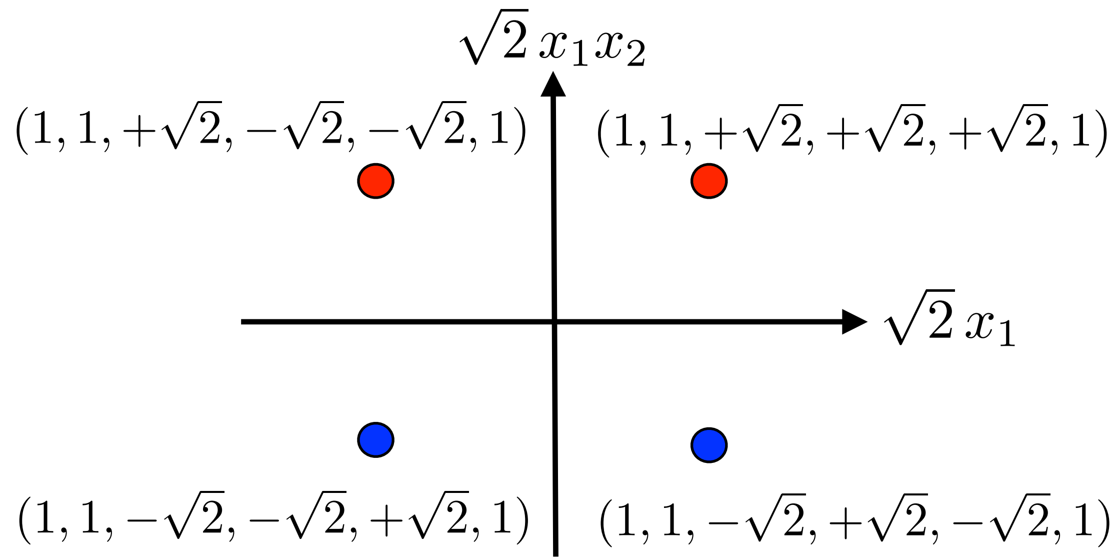
$$K(x, y) = (x_1 y_1 + x_2 y_2 + c)^2$$

$$= \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}\, x_1 x_2 \\ \sqrt{2c}\, x_1 \\ \sqrt{2c}\, x_2 \\ c \end{bmatrix} \cdot \begin{bmatrix} y_1^2 \\ y_2^2 \\ \sqrt{2}\, y_1 y_2 \\ \sqrt{2c}\, y_1 \\ \sqrt{2c}\, y_2 \\ c \end{bmatrix}.$$

XOR Problem

- Use second-degree polynomial kernel with $c = 1$:





Linearly non-separable

Linearly separable by $x_1 x_2 = 0.$

Normalized Kernels

- Definition: the normalized kernel K' associated to a kernel K is defined by

$$\forall x, x' \in \mathcal{X}, \;\; K'(x,x') = \begin{cases} 0 & \text{if } (K(x,x)=0) \vee (K(x',x')=0) \\ \frac{K(x,x')}{\sqrt{K(x,x)K(x',x')}} & \text{otherwise.} \end{cases}$$

  - If K is PDS, then K' is PDS:

$$\sum_{i,j=1}^{m} \frac{c_i c_j K(x_i,x_j)}{\sqrt{K(x_i,x_i)K(x_j,x_j)}} = \sum_{i,j=1}^{m} \frac{c_i c_j \langle \Phi(x_i), \Phi(x_j) \rangle}{\|\Phi(x_i)\|_H \, \|\Phi(x_j)\|_{\mathbb{H}}} = \left\| \sum_{i=1}^{m} \frac{c_i \Phi(x_i)}{\|\Phi(x_i)\|_H} \right\|_{\mathbb{H}}^2 \geq 0.$$

  - By definition, for all x with $K(x,x) \neq 0$,

$$K'(x,x) = 1.$$

Other Standard PDS Kernels

- Gaussian kernels:

$$K(x, y) = \exp\left(-\frac{||x - y||^2}{2\sigma^2}\right), \ \sigma \neq 0.$$

  - Normalized kernel of $(\mathbf{x}, \mathbf{x}') \mapsto \exp\left(\frac{\mathbf{x} \cdot \mathbf{x}'}{\sigma^2}\right)$.

- Sigmoid Kernels:

$$K(x, y) = \tanh(a(x \cdot y) + b), \ a, b \geq 0.$$

Reproducing Kernel Hilbert Space

(Aronszajn, 1950)

- Theorem: Let $K\colon X \times X \to \mathbb{R}$ be a PDS kernel. Then, there exists a Hilbert space H and a mapping Φ from X to H such that

$$\forall x, y \in X,\ \ K(x, y) = \Phi(x) \cdot \Phi(y).$$

- Proof: For any $x \in X$, define $\Phi(x) : X \to \mathbb{R}^X$ as follows:

$$\forall y \in X,\ \ \Phi(x)(y) = K(x, y).$$

  - Let $H_0 = \Big\{ \sum_{i \in I} a_i \Phi(x_i) \colon a_i \in \mathbb{R}, x_i \in X, \operatorname{card}(I) < \infty \Big\}.$
  - We are going to define an inner product $\langle \cdot, \cdot \rangle$ on H_0.

- Definition: for any $f = \sum_{i \in I} a_i \Phi(x_i)$, $g = \sum_{j \in J} b_j \Phi(y_j)$,

$$\langle f, g \rangle = \sum_{i \in I, j \in J} a_i b_j K(x_i, y_j) = \sum_{j \in J} b_j f(y_j) = \sum_{i \in I} a_i g(x_i).$$

- $\langle \cdot, \cdot \rangle$ does not depend on representations of f and g.
- $\langle \cdot, \cdot \rangle$ is bilinear and symmetric.
- $\langle \cdot, \cdot \rangle$ is positive semi-definite since K is PDS: for any f,

$$\langle f, f \rangle = \sum_{i, j \in I} a_i a_j K(x_i, x_j) \geq 0.$$

- note: for any $f_1, \ldots, f_m$ and $c_1, \ldots, c_m$,

$$\sum_{i,j=1}^{m} c_i c_j \langle f_i, f_j \rangle = \Big\langle \sum_{i=1}^{m} c_i f_i, \sum_{j=1}^{m} c_j f_j \Big\rangle \geq 0.$$

→ $\langle \cdot, \cdot \rangle$ is a PDS kernel on H_0.

- $\langle\cdot,\cdot\rangle$ is definite:
  - first, Cauchy-Schwarz inequality for PDS kernels. If K is PDS, $\mathbf{M} = \begin{pmatrix} K(x,x) & K(x,y) \\ K(y,x) & K(y,y) \end{pmatrix}$ is SPSD for all $x, y \in X$. In particular, the product of its eigenvalues, $\det(\mathbf{M})$ is non-negative:
$$\det(\mathbf{M}) = K(x,x)K(y,y) - K(x,y)^2 \geq 0.$$
  - since $\langle\cdot,\cdot\rangle$ is a PDS kernel, for any $f \in H_0$ and $x \in X$,
$$\langle f, \Phi(x)\rangle^2 \leq \langle f, f\rangle\langle\Phi(x), \Phi(x)\rangle.$$
  - observe the reproducing property of $\langle\cdot,\cdot\rangle$:
$$\forall f \in H_0, \forall x \in X, \;\; f(x) = \sum_{i \in I} a_i K(x_i, x) = \langle f, \Phi(x)\rangle.$$
  - Thus, $[f(x)]^2 \leq \langle f, f\rangle K(x,x)$ for all $x \in X$, which shows the definiteness of $\langle\cdot,\cdot\rangle$.

- Thus, $\langle \cdot, \cdot \rangle$ defines an inner product on H_0, which thereby becomes a pre-Hilbert space.
- H_0 can be completed to form a Hilbert space H in which it is dense.

Notes:

- H is called the reproducing kernel Hilbert space (RKHS) associated to K.
- A Hilbert space such that there exists $\Phi\colon X \to H$ with $K(x, y) = \Phi(x) \cdot \Phi(y)$ for all $x, y \in X$ is also called a feature space associated to K. Φ is called a feature mapping.
- Feature spaces associated to K are in general not unique.

This Lecture

- Kernels
- Kernel-based algorithms
- Closure properties
- Sequence Kernels
- Negative kernels

SVMs with PDS Kernels

(Boser, Guyon, and Vapnik, 1992)

- Constrained optimization:

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$

$$\text{subject to: } 0 \leq \alpha_i \leq C \wedge \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m].$$

- Solution:

$$h(x) = \operatorname{sgn}\Big(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b \Big),$$

with $b = y_i - \sum_{j=1}^{m} \alpha_j y_j K(x_j, x_i)$ for any x_i with $0 < \alpha_i < C$.

Rad. Complexity of Kernel-Based Hypotheses

- **Theorem**: Let $K\colon X \times X \to \mathbb{R}$ be a PDS kernel and let $\Phi\colon X \to \mathbb{H}$ be a feature mapping associated to K. Let $S \subseteq \{x\colon K(x,x) \leq R^2\}$ be a sample of size m, and let $H = \{\mathbf{x} \mapsto \mathbf{w} \cdot \Phi(x)\colon \|\mathbf{w}\|_{\mathbb{H}} \leq \Lambda\}$. Then,

$$\widehat{\mathfrak{R}}_S(H) \leq \frac{\Lambda\sqrt{\operatorname{Tr}[\mathbf{K}]}}{m} \leq \sqrt{\frac{R^2\Lambda^2}{m}}.$$

- **Proof**: $\widehat{\mathfrak{R}}_S(H) = \frac{1}{m}\operatorname*{E}_{\sigma}\Big[\sup_{\|\mathbf{w}\|\leq\Lambda} \mathbf{w} \cdot \sum_{i=1}^m \sigma_i \Phi(x_i)\Big] \leq \frac{\Lambda}{m}\operatorname*{E}_{\sigma}\Big[\Big\|\sum_{i=1}^m \sigma_i \Phi(x_i)\Big\|\Big]$

$$\text{(Jensen's ineq.)} \leq \frac{\Lambda}{m}\bigg[\operatorname*{E}_{\sigma}\Big[\Big\|\sum_{i=1}^m \sigma_i \Phi(x_i)\Big\|^2\Big]\bigg]^{1/2} \leq \frac{\Lambda}{m}\bigg[\operatorname*{E}_{\sigma}\Big[\sum_{i=1}^m \|\Phi(x_i)\|^2\Big]\bigg]^{1/2}$$

$$= \frac{\Lambda}{m}\bigg[\operatorname*{E}_{\sigma}\Big[\sum_{i=1}^m K(x_i,x_i)\Big]\bigg]^{1/2} = \frac{\Lambda\sqrt{\operatorname{Tr}[\mathbf{K}]}}{m} \leq \sqrt{\frac{R^2\Lambda^2}{m}}.$$

Generalization: Representer Theorem

(Kimeldorf and Wahba, 1971)

- **Theorem**: Let $K\colon X \times X \to \mathbb{R}$ be a PDS kernel with H the corresponding RKHS. Then, for any non-decreasing function $G\colon \mathbb{R} \to \mathbb{R}$ and any $L\colon \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ problem

$$\operatorname*{argmin}_{h \in H} F(h) = \operatorname*{argmin}_{h \in H} \; G(\|h\|_H) + L\big(h(x_1), \ldots, h(x_m)\big)$$

admits a solution of the form $h^* = \sum_{i=1}^m \alpha_i K(x_i, \cdot)$.

If G is further assumed to be increasing, then any solution has this form.

- **Proof**: let $H_1 = \text{span}(\{K(x_i, \cdot)\colon i \in [1, m]\})$. Any $h \in H$ admits the decomposition $h = h_1 + h^\perp$ according to $H = H_1 \oplus H_1^\perp$.
  - Since G is non-decreasing,
  $$G(\|h_1\|_H) \leq G\Big(\sqrt{\|h_1\|_H^2 + \|h^\perp\|_H^2}\Big) = G(\|h\|_H).$$
  - By the reproducing property, for all $i \in [1, m]$,
  $$h(x_i) = \langle h, K(x_i, \cdot)\rangle = \langle h_1, K(x_i, \cdot)\rangle = h_1(x_i).$$
  - Thus, $L\big(h(x_1), \ldots, h(x_m)\big) = L\big(h_1(x_1), \ldots, h_1(x_m)\big)$ and $F(h_1) \leq F(h)$.
  - If G is increasing, then $F(h_1) < F(h)$ when $h^\perp \neq 0$ and any solution of the optimization problem must be in H_1.

Kernel-Based Algorithms

- PDS kernels used to extend a variety of algorithms in classification and other areas:
  - regression.
  - ranking.
  - dimensionality reduction.
  - clustering.
- But, how do we define PDS kernels?

This Lecture

- Kernels
- Kernel-based algorithms
- Closure properties
- Sequence Kernels
- Negative kernels

Closure Properties of PDS Kernels

- Theorem: Positive definite symmetric (PDS) kernels are closed under:
  - sum,
  - product,
  - tensor product,
  - pointwise limit,
  - composition with a power series with non-negative coefficients.

Closure Properties - Proof

- Proof: closure under sum:

$$\mathbf{c}^\top \mathbf{K} \mathbf{c} \geq 0 \wedge \mathbf{c}^\top \mathbf{K}' \mathbf{c} \geq 0 \Rightarrow \mathbf{c}^\top (\mathbf{K} + \mathbf{K}') \mathbf{c} \geq 0.$$

- closure under product: $\mathbf{K} = \mathbf{M}\mathbf{M}^\top$,

$$\sum_{i,j=1}^{m} c_i c_j (\mathbf{K}_{ij} \mathbf{K}'_{ij}) = \sum_{i,j=1}^{m} c_i c_j \Big(\Big[\sum_{k=1}^{m} \mathbf{M}_{ik} \mathbf{M}_{jk} \Big] \mathbf{K}'_{ij} \Big)$$

$$= \sum_{k=1}^{m} \Big[\sum_{i,j=1}^{m} c_i c_j \mathbf{M}_{ik} \mathbf{M}_{jk} \mathbf{K}'_{ij} \Big]$$

$$= \sum_{k=1}^{m} \begin{bmatrix} c_1 \mathbf{M}_{1k} \\ \cdots \\ c_m \mathbf{M}_{mk} \end{bmatrix}^\top \mathbf{K}' \begin{bmatrix} c_1 \mathbf{M}_{1k} \\ \cdots \\ c_m \mathbf{M}_{mk} \end{bmatrix} \geq 0.$$

- Closure under tensor product:
  - definition: for all $x_1, x_2, y_1, y_2 \in X$,

$$(K_1 \otimes K_2)(x_1, y_1, x_2, y_2) = K_1(x_1, x_2) K_2(y_1, y_2).$$

  - thus, PDS kernel as product of the kernels

$$(x_1, y_1, x_2, y_2) \to K_1(x_1, x_2) \quad (x_1, y_1, x_2, y_2) \to K_2(y_1, y_2).$$

- Closure under pointwise limit: if for all $x, y \in X$,

$$\lim_{n \to \infty} K_n(x, y) = K(x, y),$$

Then, $(\forall n, \mathbf{c}^\top \mathbf{K}_n \mathbf{c} \geq 0) \Rightarrow \lim_{n \to \infty} \mathbf{c}^\top \mathbf{K}_n \mathbf{c} = \mathbf{c}^\top \mathbf{K} \mathbf{c} \geq 0.$

- Closure under composition with power series:
  - assumptions: K PDS kernel with $|K(x,y)| < \rho$ for all $x, y \in X$ and $f(x) = \sum_{n=0}^{\infty} a_n x^n, a_n \geq 0$ power series with radius of convergence ρ.
  - $f \circ K$ is a PDS kernel since K^n is PDS by closure under product, $\sum_{n=0}^{N} a_n K^n$ is PDS by closure under sum, and closure under pointwise limit.
- Example: for any PDS kernel K, $\exp(K)$ is PDS.

This Lecture

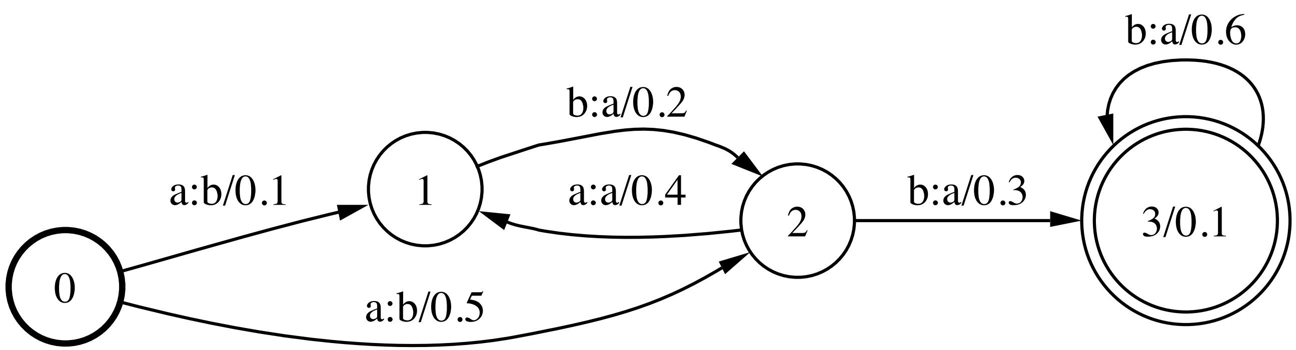
- Kernels
- Kernel-based algorithms
- Closure properties
- Sequence Kernels
- Negative kernels

Sequence Kernels

- Definition: Kernels defined over pairs of strings.
  - Motivation: computational biology, text and speech classification.
  - Idea: two sequences are related when they share some common substrings or subsequences.
  - Example: bigram kernel;

$$K(x, y) = \sum_{\text{bigram } u} \text{count}_x(u) \times \text{count}_y(u).$$

Weighted Transducers



$T(x, y) =$ Sum of the weights of all accepting paths with input x and output y.

$$T(abb, baa) = .1 \times .2 \times .3 \times .1 + .5 \times .3 \times .6 \times .1$$

Rational Kernels over Strings

(Cortes et al., 2004)

- Definition: a kernel $K: \Sigma^* \times \Sigma^* \to \mathbb{R}$ is rational if $K = T$ for some weighted transducer T.

- Definition: let $T_1: \Sigma^* \times \Delta^* \to \mathbb{R}$ and $T_2: \Delta^* \times \Omega^* \to \mathbb{R}$ be two weighted transducers. Then, the composition of T_1 and T_2 is defined for all $x \in \Sigma^*, y \in \Omega^*$ by

$$(T_1 \circ T_2)(x, y) = \sum_{z \in \Delta^*} T_1(x, z)\; T_2(z, y).$$

- Definition: the inverse of a transducer $T: \Sigma^* \times \Delta^* \to \mathbb{R}$ is the transducer $T^{-1}: \Delta^* \times \Sigma^* \to \mathbb{R}$ obtained from T by swapping input and output labels.

PDS Rational Kernels
General Construction

- **Theorem**: for any weighted transducer $T\colon \Sigma^* \times \Sigma^* \to \mathbb{R}$, the function $K = T \circ T^{-1}$ is a PDS rational kernel.

- **Proof**: by definition, for all $x, y \in \Sigma^*$,

$$K(x, y) = \sum_{z \in \Delta^*} T(x, z)\, T(y, z).$$

  - K is pointwise limit of $(K_n)_{n \geq 0}$ defined by

$$\forall x, y \in \Sigma^*, \;\; K_n(x, y) = \sum_{|z| \leq n} T(x, z)\, T(y, z).$$

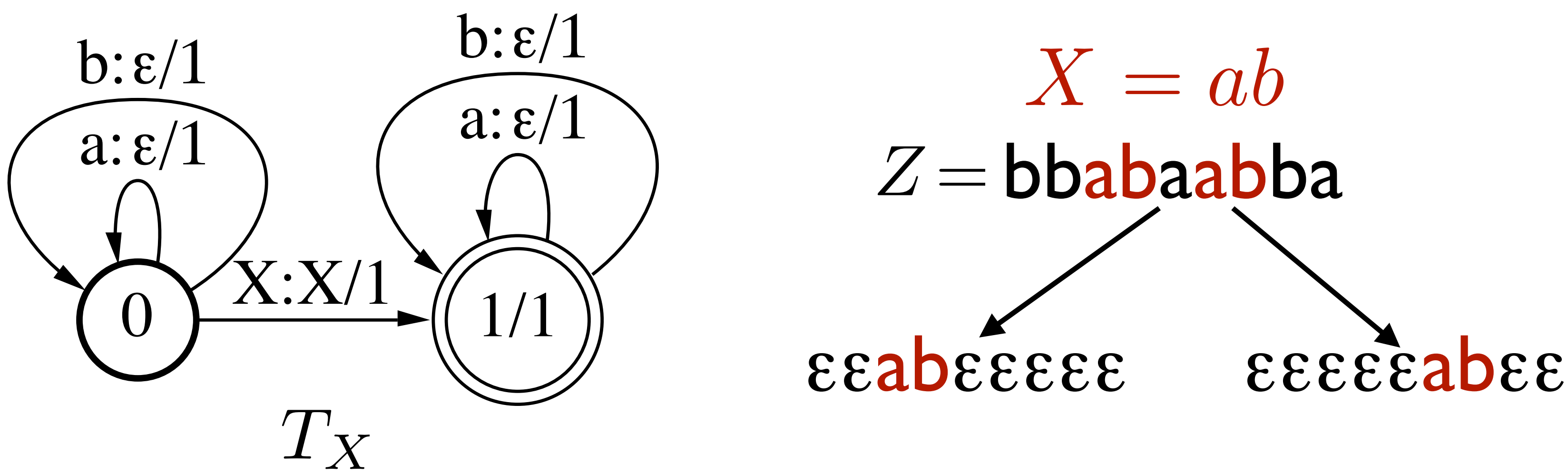
  - K_n is PDS since for any sample $(x_1, \ldots, x_m)$,

$$\mathbf{K}_n = \mathbf{A}\mathbf{A}^\top \text{ with } \mathbf{A} = (K_n(x_i, z_j))_{\substack{i \in [1, m] \\ j \in [1, N]}}.$$

PDS Sequence Kernels

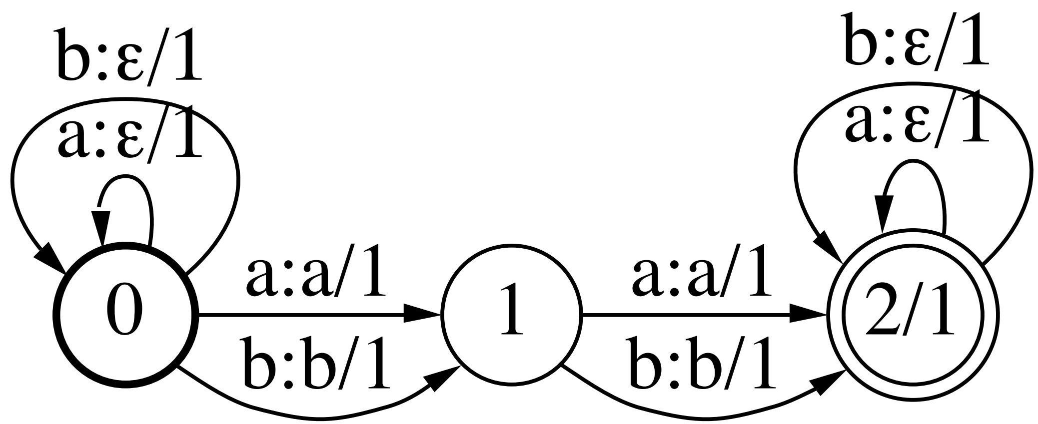
- PDS sequences kernels in computational biology, text classification, other applications:
  - special instances of PDS rational kernels.
  - PDS rational kernels easy to define and modify.
  - single general algorithm for their computation: composition + shortest-distance computation.
  - no need for a specific 'dynamic-programming' algorithm and proof for each kernel instance.
  - general sub-family: based on counting transducers.

Counting Transducers



- X may be a string or an automaton representing a regular expression.
- Counts of Z in X: sum of the weights of accepting paths of $Z \circ T_X$.

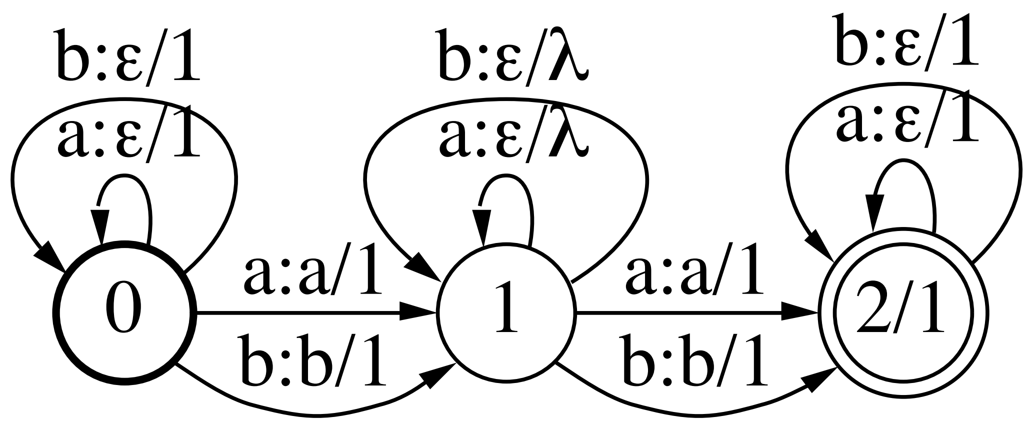
Transducer Counting Bigrams



T_{bigram}

Counts of Z given by $Z \circ T_{\text{bigram}} \circ ab$.

Transducer Counting Gappy Bigrams



$T_{\text{gappy bigram}}$

Counts of Z given by $Z \circ T_{\text{gappy bigram}} \circ ab$,
gap penalty $\lambda \in (0, 1)$.

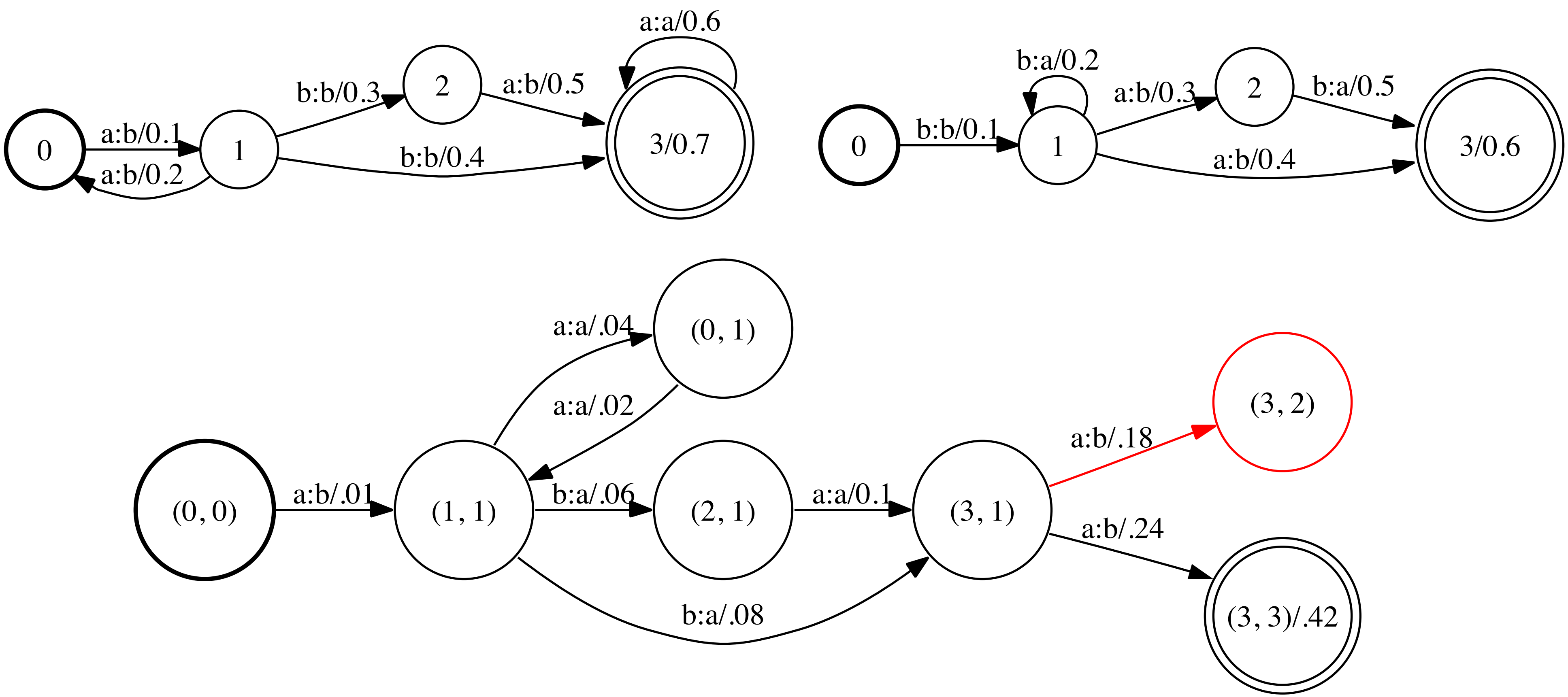
Composition

- Theorem: the composition of two weighted transducer is also a weighted transducer.
- Proof: constructive proof based on composition algorithm.
  - states identified with pairs.
  - ϵ-free case: transitions defined by

$$E = \biguplus_{\substack{(q_1,a,b,w_1,q_2)\in E_1\\(q_1',b,c,w_2,q_2')\in E_2}} \Big\{\Big((q_1,q_1'),a,c,w_1\times w_2,(q_2,q_2')\Big)\Big\}.$$

  - general case: use of intermediate ϵ-filter.

Composition Algorithm
ε-Free Case

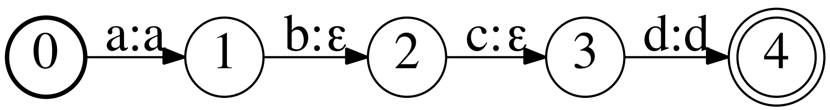


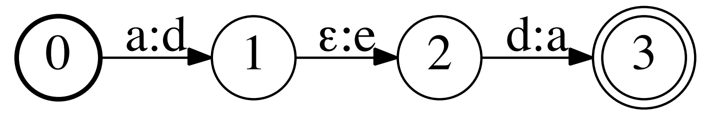
Complexity: $O(|T_1|\,|T_2|)$ in general, linear in some cases.

Redundant ε-Paths Problem

(MM, Pereira, and Riley, 1996; Pereira and Riley, 1997)

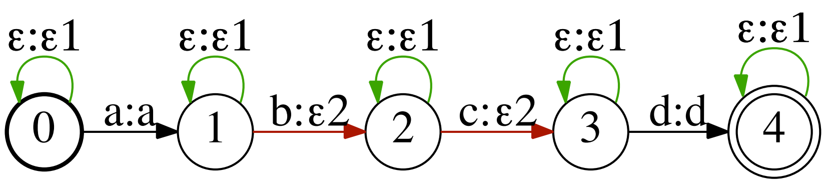
T_1

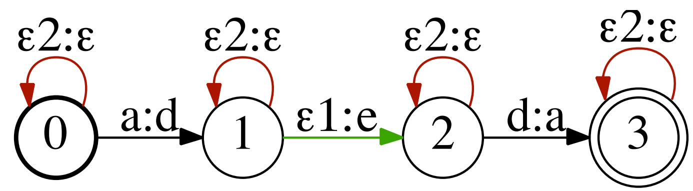




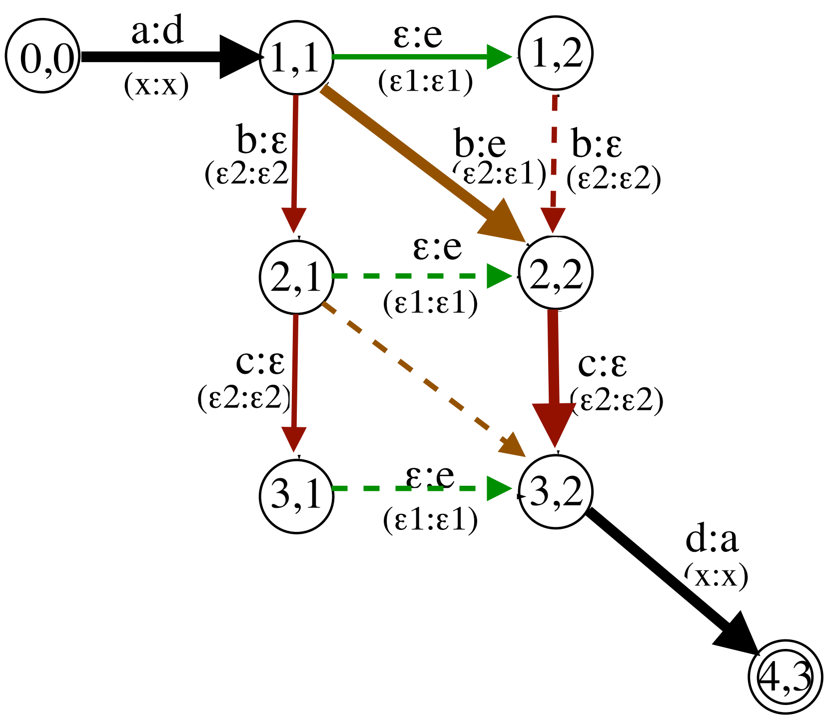
T_2

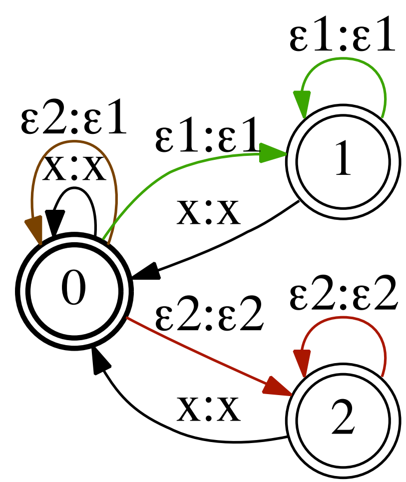
$\widetilde{T}_1$





$\widetilde{T}_2$





F

$$T = \widetilde{T}_1 \circ F \circ \widetilde{T}_2.$$

Kernels for Other Discrete Structures

- Similarly, PDS kernels can be defined on other discrete structures:
  - Images,
  - graphs,
  - parse trees,
  - automata,
  - weighted automata.

This Lecture

- Kernels
- Kernel-based algorithms
- Closure properties
- Sequence Kernels
- Negative kernels

Questions

- Gaussian kernels have the form $\exp(-d^2)$ where d is a metric.
  - for what other functions d does $\exp(-d^2)$ define a PDS kernel?
  - what other PDS kernels can we construct from a metric in a Hilbert space?

Negative Definite Kernels

(Schoenberg, 1938)

- Definition: A function $K \colon X \times X \to \mathbb{R}$ is said to be a negative definite symmetric (NDS) kernel if it is symmetric and if for all $\{x_1, \ldots, x_m\} \subseteq X$ and $\mathbf{c} \in \mathbb{R}^{m \times 1}$ with $\mathbf{1}^\top \mathbf{c} = 0$,

$$\mathbf{c}^\top \mathbf{K} \mathbf{c} \leq 0.$$

- Clearly, if K is PDS, then $-K$ is NDS, but the converse does not hold in general.

Examples

- The squared distance $||x - y||^2$ in a Hilbert space H defines an NDS kernel. If $\sum_{i=1}^m c_i = 0$,

$$\begin{aligned}
\sum_{i,j=1}^m c_i c_j ||\mathbf{x}_i - \mathbf{x}_j||^2 &= \sum_{i,j=1}^m c_i c_j (\mathbf{x}_i - \mathbf{x}_j) \cdot (\mathbf{x}_i - \mathbf{x}_j) \\
&= \sum_{i,j=1}^m c_i c_j (\|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2 - 2\mathbf{x}_i \cdot \mathbf{x}_j) \\
&= \sum_{i,j=1}^m c_i c_j (\|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2) - 2 \sum_{i=1}^m c_i \mathbf{x}_i \cdot \sum_{j=1}^m c_j \mathbf{x}_j \\
&\leq \sum_{i,j=1}^m c_i c_j (\|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2) \\
&= \sum_{j=1}^m c_j \Big(\sum_{i=1}^m c_i (\|\mathbf{x}_i\|^2 \Big) + \sum_{i=1}^m c_i \Big(\sum_{j=1}^m c_j \|\mathbf{x}_j\|^2 \Big) = 0.
\end{aligned}$$

NDS Kernels - Property

(Schoenberg, 1938)

- **Theorem**: Let $K\colon X \times X \to \mathbb{R}$ be an NDS kernel such that for all $x, y \in X$, $K(x, y) = 0$ iff $x = y$. Then, there exists a Hilbert space H and a mapping $\Phi\colon X \to H$ such that

$$\forall x, y \in X, \;\; K(x, y) = \|\Phi(x) - \Phi(y)\|^2.$$

Thus, under the hypothesis of the theorem, $\sqrt{K}$ defines a metric.

PDS and NDS Kernels

(Schoenberg, 1938)

- Theorem: let $K\colon X \times X \to \mathbb{R}$ be a symmetric kernel, then:
  - K is NDS iff $\exp(-tK)$ is a PDS kernel for all $t > 0$.
  - Let K' be defined for any x_0 by

    $$K'(x, y) = K(x, x_0) + K(y, x_0) - K(x, y) - K(x_0, x_0)$$

    for all $x, y \in X$. Then, K is NDS iff K' is PDS.

Example

- The kernel defined by $K(x, y) = \exp(-t||x - y||^2)$ is PDS for all $t > 0$ since $||x - y||^2$ is NDS.

- The kernel $\exp(-|x - y|^p)$ is not PDS for $p > 2$. Otherwise, for any $t > 0$, $\{x_1, \ldots, x_m\} \subseteq X$ and $\mathbf{c} \in \mathbb{R}^{m \times 1}$

$$\sum_{i,j=1}^{m} c_i c_j e^{-t|x_i - x_j|^p} = \sum_{i,j=1}^{m} c_i c_j e^{-|t^{1/p} x_i - t^{1/p} x_j|^p} \geq 0.$$

- This would imply that $|x - y|^p$ is NDS for $p > 2$, but that cannot be (see past homework assignments).

Conclusion

- PDS kernels:
  - rich mathematical theory and foundation.
  - general idea for extending many linear algorithms to non-linear prediction.
  - flexible method: any PDS kernel can be used.
  - widely used in modern algorithms and applications.
  - can we further learn a PDS kernel and a hypothesis based on that kernel from labeled data? (see tutorial: http://www.cs.nyu.edu/~mohri/icml2011-tutorial/).

References

- N. Aronszajn, Theory of Reproducing Kernels, *Trans. Amer. Math. Soc.*, 68, 337-404, 1950.
- Peter Bartlett and John Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In *Advances in kernel methods: support vector learning*, pages 43–54. MIT Press, Cambridge, MA, USA, 1999.
- Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic Analysis on Semigroups*. Springer-Verlag: Berlin-New York, 1984.
- Bernhard Boser, Isabelle M. Guyon, and Vladimir Vapnik. *A training algorithm for optimal margin classifiers*. In proceedings of COLT 1992, pages 144-152, Pittsburgh, PA, 1992.
- Corinna Cortes, Patrick Haffner, and Mehryar Mohri. Rational Kernels: Theory and Algorithms. *Journal of Machine Learning Research (JMLR)*, 5:1035-1062, 2004.
- Corinna Cortes and Vladimir Vapnik, Support-Vector Networks, *Machine Learning*, 20, 1995.
- Kimeldorf, G. and Wahba, G. *Some results on Tchebycheffian Spline Functions*, J. Mathematical Analysis and Applications, 33, 1 (1971) 82-95.

References

- James Mercer. Functions of Positive and Negative Type, and Their Connection with the Theory of Integral Equations. In *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, Vol. 83, No. 559, pp. 69-70, 1909.

- Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley. *Weighted Automata in Text and Speech Processing*, In *Proceedings of the 12th biennial European Conference on Artificial Intelligence (ECAI-96), Workshop on Extended finite state models of language*. Budapest, Hungary, 1996.

- Fernando C. N. Pereira and Michael D. Riley. Speech Recognition by Composition of Weighted Finite Automata. In Finite-State Language Processing, pages 431-453. MIT Press, 1997.

- I. J. Schoenberg, Metric Spaces and Positive Definite Functions. *Transactions of the American Mathematical Society*, Vol. 44, No. 3, pp. 522-536, 1938.

- Vladimir N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer, Basederlin, 1982.

- Vladimir N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1995.

- Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.

Appendix

Mercer's Condition

(Mercer, 1909)

- **Theorem**: Let $X \times X$ be a compact subset of $\mathbb{R}^N$ and let $K : X \times X \to \mathbb{R}$ be in $L_\infty(X \times X)$ and symmetric. Then, K admits a uniformly convergent expansion

$$K(x, y) = \sum_{n=0}^{\infty} a_n \phi_n(x) \phi_n(y), \text{ with } a_n > 0,$$

iff for any function c in $L_2(X)$,

$$\int\int_{X \times X} c(x) c(y) K(x, y) dx dy \geq 0.$$

SVMs with PDS Kernels

- Constrained optimization:

Hadamard product

$$\max_{\boldsymbol{\alpha}} 2\, \mathbf{1}^\top \boldsymbol{\alpha} - (\boldsymbol{\alpha} \circ \mathbf{y})^\top \mathbf{K} (\boldsymbol{\alpha} \circ \mathbf{y})$$

$$\text{subject to: } \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} \wedge \boldsymbol{\alpha}^\top \mathbf{y} = 0.$$

- Solution:

$$h = \operatorname{sgn}\Big(\sum_{i=1}^{m} \alpha_i y_i K(x_i, \cdot) + b\Big),$$

with $b = y_i - (\boldsymbol{\alpha} \circ \mathbf{y})^\top \mathbf{K} \mathbf{e}_i$ for any x_i with $0 < \alpha_i < C.$