

KERNELS

Evgeny Burnaev

Skoltech, Moscow, Russia

OUTLINE

- 1 MOTIVATION
- 2 KERNELS
- 3 SVMs WITH KERNELS
- 4 CLOSURE PROPERTIES
- 5 NEGATIVE KERNELS

1 MOTIVATION

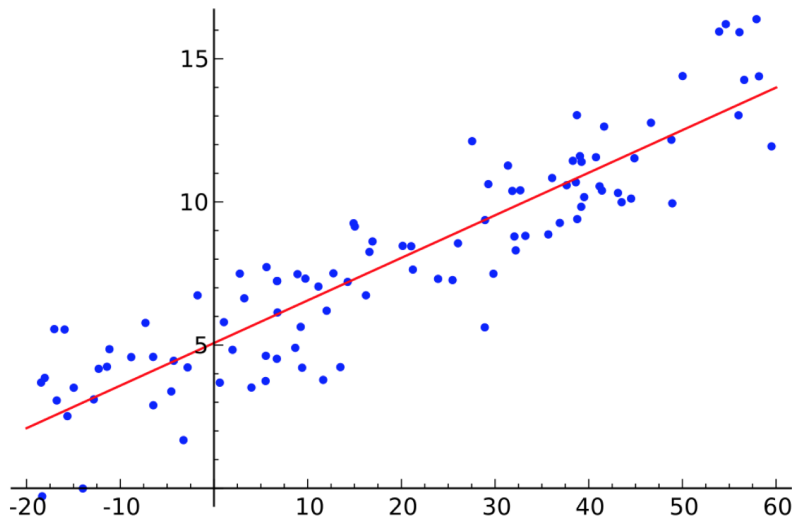
2 KERNELS

3 SVMs WITH KERNELS

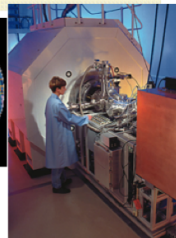
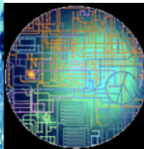
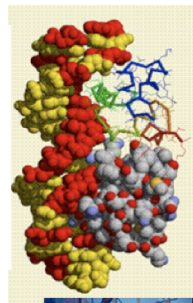
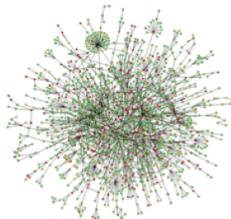
4 CLOSURE PROPERTIES

5 NEGATIVE KERNELS

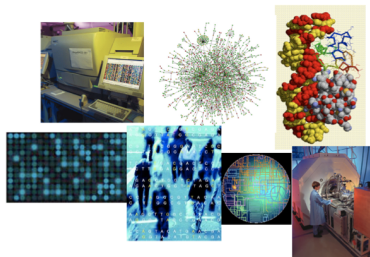
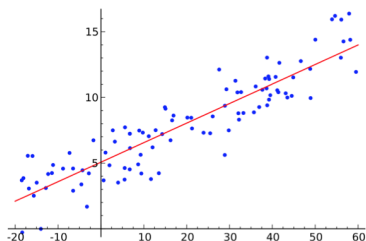
WHAT WE KNOW HOW TO SOLVE



BUT REAL DATA ARE OFTEN MORE COMPLICATED ...



MAIN GOAL OF THIS LECTURE



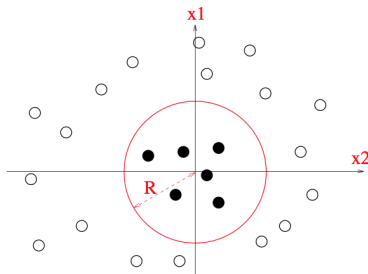
Show some classical examples how to extend well-understood, linear statistical learning techniques to real-world, complicated, structured, high-dimensional data (texts, time series, graphs, distributions, permutations, ...)

MOTIVATION

- Efficient computation of inner products in high dimension
- Non-linear decision boundary
- Learning with non-vectorial inputs
- More informative features
- Kernels allow to perform pairwise comparisons

- 1 MOTIVATION
- 2 KERNELS**
- 3 SVMs WITH KERNELS
- 4 CLOSURE PROPERTIES
- 5 NEGATIVE KERNELS

NON-LINEAR SEPARATION



- Linear separation impossible in most problems
- Non-linear mapping $\Phi : X \rightarrow \mathbb{H}$ from input space to high-dimensional feature space
- Generalization ability: independent of $\dim(\mathbb{H})$, depends only on ρ and m

Kernel Methods

• Idea:

- Define $K : X \times X \rightarrow \mathbb{R}$ called kernel, such that

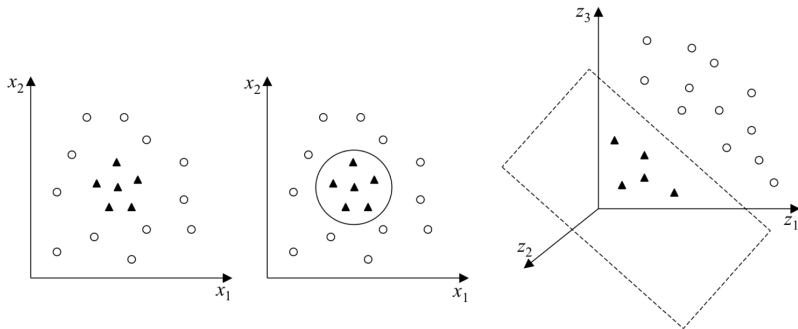
$$\Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}') = K(\mathbf{x}, \mathbf{x}')$$

- K is 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)

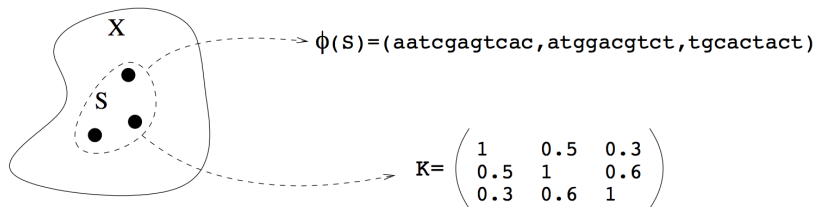
EXAMPLE: POLYNOMIAL KERNEL



For $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$, let $\Phi(\mathbf{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2) \in \mathbb{R}^3$. Then

$$\begin{aligned} K(\mathbf{x}, \mathbf{x}') &= x_1^2(x_1')^2 + 2x_1x_2x_1'x_2' + x_2^2(x_2')^2 \\ &= (x_1x_1' + x_2x_2')^2 = (\mathbf{x} \cdot \mathbf{x}')^2 \end{aligned}$$

REPRESENTATION BY PAIRWISE COMPARISONS

**Idea:**

- Define a “comparison function”: $K : X \times X \rightarrow \mathbb{R}$
- Represent a set of m data points $S = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ by the $m \times m$ matrix

$$[K]_{i,j} := K(\mathbf{x}_i, \mathbf{x}_j)$$

PDS CONDITION

- **Definition:** a kernel $K : X \times X \rightarrow \mathbb{R}$ is *positive definite symmetric* (PDS) if for any $\{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subseteq X$ the matrix $K = [K(\mathbf{x}_i, \mathbf{x}_j)]_{ij} \in \mathbb{R}^{m \times m}$ is symmetric positive semi-definite (SPSD)
- K SPD if symmetric and one of the 2 equiv. cond.'s:
 - its eigenvalues are non-negative
 - for any $\mathbf{c} \in \mathbb{R}^{m \times 1}$, $\mathbf{c}^T K \mathbf{c} = \sum_{i,j=1}^m c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$
- **Terminology:** PDS for kernels, SPDS for kernel matrices

EXAMPLE: POLYNOMIAL KERNELS

- Definition:

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

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

$$K(\mathbf{x}, \mathbf{x}') = (x_1x'_1 + x_2x'_2 + c)^2$$

$$= \begin{bmatrix} \mathbf{x}_1^2 \\ \mathbf{x}_2^2 \\ \sqrt{2}\mathbf{x}_1\mathbf{x}_2 \\ \sqrt{2c}\mathbf{x}_1 \\ \sqrt{2c}\mathbf{x}_2 \\ c \end{bmatrix} \cdot \begin{bmatrix} \mathbf{x}'_1^2 \\ \mathbf{x}'_2^2 \\ \sqrt{2}\mathbf{x}'_1\mathbf{x}'_2 \\ \sqrt{2c}\mathbf{x}'_1 \\ \sqrt{2c}\mathbf{x}'_2 \\ c \end{bmatrix}$$

NORMALIZED KERNELS

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

$$\forall \mathbf{x}, \mathbf{x}' \in X, K'(\mathbf{x}, \mathbf{x}') = \begin{cases} 0, & \text{if } K(\mathbf{x}, \mathbf{x}) = 0 \text{ or } K(\mathbf{x}', \mathbf{x}') = 0 \\ \frac{K(\mathbf{x}, \mathbf{x}')}{\sqrt{K(\mathbf{x}, \mathbf{x})K(\mathbf{x}', \mathbf{x}')}} & \end{cases}$$

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

$$K'(\mathbf{x}, \mathbf{x}) = 1$$

NORMALIZED KERNELS

- If K is PDS, then K' is PDS

$$\begin{aligned}\sum_{i,j=1}^m \frac{c_i c_j K(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{K(\mathbf{x}_i, \mathbf{x}_i) K(\mathbf{x}_j, \mathbf{x}_j)}} &= \sum_{i,j=1}^m \frac{c_i c_j \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle}{\|\Phi(\mathbf{x}_i)\|_{\mathbb{H}} \|\Phi(\mathbf{x}_j)\|_{\mathbb{H}}} \\ &= \left\| \sum_{i=1}^m \frac{c_i \Phi(\mathbf{x}_i)}{\|\Phi(\mathbf{x}_i)\|_{\mathbb{H}}} \right\|_{\mathbb{H}}^2 \geq 0\end{aligned}$$

STANDARD PDS KERNELS

- **Gaussian kernels:**

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

Gaussian kernel is a normalized kernel of

$$(\mathbf{x}, \mathbf{x}') \rightarrow \exp \left(\frac{\mathbf{x} \cdot \mathbf{x}'}{\sigma^2} \right) = \sum_{n=0}^{\infty} \frac{(\mathbf{x} \cdot \mathbf{x}')^n}{\sigma^n n!}$$

- **Sigmoid kernels:**

$$K(\mathbf{x}, \mathbf{x}') = \tanh(a(\mathbf{x} \cdot \mathbf{x}') + b), \quad a, b \geq 0$$

REPR. KERNEL HILBERT SPACE I (ARONSZAJN, 1950)

- Let $\mathbf{x}, \mathbf{z} \in \mathbb{R}^N$. We consider the space of functions \mathbb{H} generated by the linear span of $\{K(\cdot, \mathbf{z}), \mathbf{z} \in \mathbb{R}^N\}$; i.e. arbitrary linear combinations of the form

$$f(\mathbf{x}) = \sum_m a_m K(\mathbf{x}, \mathbf{z}_m),$$

where each kernel term is viewed as a function of the first argument, and indexed by the second

- Suppose K has an eigen-expansion

$$K(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{\infty} a_i \phi_i(\mathbf{x}) \phi_i(\mathbf{z})$$

with $a_i > 0$, $\sum_{i=1}^{\infty} a_i^2 < \infty$

REPR. KERNEL HILBERT SPACE II

- Elements of \mathbb{H} have an expansion in terms of these eigen-functions

$$f(\mathbf{x}) = \sum_{i=1}^{\infty} c_i \phi_i(\mathbf{x}),$$

with the constraint that

$$\|f\|_{\mathbb{H}}^2 := \sum_{i=1}^{\infty} \frac{c_i^2}{a_i} < \infty$$

- For $f \in \mathbb{H}$ it can be easily seen that

$$\langle K(\cdot, \mathbf{x}_i), f \rangle = f(\mathbf{x}_i), \quad \langle K(\cdot, \mathbf{x}_i), K(\cdot, \mathbf{x}_j) \rangle = K(\mathbf{x}_i, \mathbf{x}_j)$$

- Thus for $f(\mathbf{x}) = \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}_i)$ we get that

$$\|f\|_{\mathbb{H}}^2 = \sum_{i,j=1}^m \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

MERCER'S CONDITION

There is no need to explicitly define or compute a mapping Φ

- **Theorem:** Let $X \subset \mathbb{R}^N$ be a compact set and $K : X \times X \rightarrow \mathbb{R}$ be a continuous and symmetric. Then, K admits a uniformly convergent expansion

$$K(\mathbf{x}, \mathbf{x}') = \sum_{n=0}^{\infty} a_n \phi_n(\mathbf{x}) \phi_n(\mathbf{x}'),$$

with $a_n > 0$ iff for any square integrable function c ($c \in L_2(X)$), the following condition holds

$$\int \int_{X \times X} c(\mathbf{x}) c(\mathbf{x}') K(\mathbf{x}, \mathbf{x}') d\mathbf{x} d\mathbf{x}' \geq 0$$

- This condition is important to guarantee the convexity of the optimization problem for algorithms such as SVMs
- However, this construction is valid only for $X \subset \mathbb{R}^N$. The next theorem provides construction in a general case

REPR. KERNEL HILBERT SPACE III

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

$$\forall \mathbf{x}, \mathbf{x}' \in X, K(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}')$$

- **Proof:** for any $\mathbf{x} \in X$, define $\Phi(\mathbf{x}) : X \rightarrow \mathbb{R}^X$ as follows:

$$\forall \mathbf{z} \in X, \Phi(\mathbf{x})(\mathbf{z}) = K(\mathbf{x}, \mathbf{z})$$

— Let

$$\mathbb{H}_0 = \left\{ \sum_{i \in I} a_i \Phi(\mathbf{x}_i) : a_i \in \mathbb{R}, \mathbf{x}_i \in X, \text{card}(I) < \infty \right\}$$

— We are going to define an inner product $\langle \cdot, \cdot \rangle$ on \mathbb{H}_0

REPRODUCING KERNEL HILBERT SPACE IV

— **Definition:** for any $f = \sum_{i \in I} a_i \Phi(\mathbf{x}_i)$, $g = \sum_{j \in J} b_j \Phi(\mathbf{z}_j)$

$$\langle f, g \rangle = \sum_{i \in I, j \in J} a_i b_j K(\mathbf{x}_i, \mathbf{z}_j) = \sum_{j \in J} b_j f(\mathbf{z}_j) = \sum_{i \in I} a_i g(\mathbf{x}_i)$$

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

$$\text{for any } f, \langle f, f \rangle = \sum_{i, j \in I} a_i a_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$$

for any f_1, \dots, f_m and c_1, \dots, c_m

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

$\Rightarrow \langle \cdot, \cdot \rangle$ is a PDS kernel on \mathbb{H}_0

REPRODUCING KERNEL HILBERT SPACE V

- $\langle \cdot, \cdot \rangle$ is well-defined:
 - Let us consider **Cauchy-Schwarz** inequality for PDS kernels. If K is PDS, then

$$\mathbf{M} = \begin{pmatrix} K(\mathbf{x}, \mathbf{x}) & K(\mathbf{x}, \mathbf{z}) \\ K(\mathbf{z}, \mathbf{x}) & K(\mathbf{z}, \mathbf{z}) \end{pmatrix}$$

is SPSD for all $\mathbf{x}, \mathbf{z} \in X$. In particular, the product of its eigenvalues, $\det(\mathbf{M})$ is non-negative:

$$\det(\mathbf{M}) = K(\mathbf{x}, \mathbf{x})K(\mathbf{z}, \mathbf{z}) - K(\mathbf{x}, \mathbf{z})^2 \geq 0$$

- Since $\langle \cdot, \cdot \rangle$ is a PDS kernel, for any $f \in \mathbb{H}_0$ and $\mathbf{x} \in X$

$$\langle f, \Phi(\mathbf{x}) \rangle^2 \leq \langle f, f \rangle \cdot \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}) \rangle$$

REPR. KERNEL HILBERT SPACE VI (PROOF CONT.'D)

- Observe the **reproducing property** of $\langle \cdot, \cdot \rangle$:

$$\forall f \in \mathbb{H}_0, \forall \mathbf{x} \in X, f(\mathbf{x}) = \sum_{i \in I} a_i K(\mathbf{x}_i, \mathbf{x}) = \langle f, \Phi(\mathbf{x}) \rangle$$

- Thus, $[f(\mathbf{x})]^2 \leq \langle f, f \rangle K(\mathbf{x}, \mathbf{x})$ for all $\mathbf{x} \in X$, which shows the definiteness of $\langle \cdot, \cdot \rangle$
- Thus $\langle \cdot, \cdot \rangle$ defines an inner product on \mathbb{H}_0 , which thereby becomes a pre-Hilbert space
- \mathbb{H}_0 can be completed to form a Hilbert space \mathbb{H} in which it is dense
- By the Cauchy-Schwarz inequality, for any $\mathbf{x} \in X$, $f \rightarrow \langle f, \Phi(\mathbf{x}) \rangle$ is Lipschitz, therefore continuous. Thus since \mathbb{H}_0 is dense in \mathbb{H} , the reproducing property also holds over \mathbb{H}

COMMENTS

- \mathbb{H} is called the Reproducing Kernel Hilbert Space (RKHS), associated to K
- A Hilbert space such that there exists $\Phi : X \rightarrow \mathbb{H}$ with $K(\mathbf{x}, \mathbf{z}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{z})$ for all $\mathbf{x}, \mathbf{z} \in X$ 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*

- 1 MOTIVATION
- 2 KERNELS
- 3 SVMs WITH KERNELS**
- 4 CLOSURE PROPERTIES
- 5 NEGATIVE KERNELS

SVMs WITH PDS KERNELS

- Constrained Optimization:

$$\max_{\alpha} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j}^m \alpha_i \alpha_j y_i y_j \underbrace{K(\mathbf{x}_i, \mathbf{x}_j)}_{\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)}$$

$$\text{s.t. } 0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^m \alpha_i y_i = 0, i \in [1, m]$$

- Solution

$$h(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^m \alpha_i y_i \underbrace{K(\mathbf{x}_i, \mathbf{x})}_{\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})} + b \right),$$

with $b = y_i - \sum_{j=1}^m \alpha_j y_j \underbrace{K(\mathbf{x}_j, \mathbf{x}_i)}_{\Phi(\mathbf{x}_j) \cdot \Phi(\mathbf{x}_i)}$ for any SV \mathbf{x}_i with

$$0 < \alpha_i < C$$

GENERAL REGULARIZATION PROBLEM STATEMENT I

- A general class of regularization problems has the form

$$\min_{f \in \mathbb{H}} \left[\sum_{i=1}^m L(y_i, f(\mathbf{x}_i)) + \lambda J(f) \right]$$

where $L(y, f(\mathbf{x}))$ is a loss function, $J(f)$ is a penalty functional, \mathbb{H} is a space of functions

- In case of RKHS \mathbb{H}_K , induced by the kernel K we use $J(f) = \|f\|_{\mathbb{H}_K}^2$ and get

$$\min_{f \in \mathbb{H}_K} \left[\sum_{i=1}^m L(y_i, f(\mathbf{x}_i)) + \lambda \|f\|_{\mathbb{H}_K}^2 \right]$$

GENERAL REGULARIZATION PROBLEM STATEMENT II

- Using RKHS basis representation we get equivalent problem formulation

$$\min_{\{c_j\}_{j=1}^{\infty}} \left[\sum_{i=1}^m L \left(y_i, \sum_{j=1}^{\infty} c_j \phi_j(\mathbf{x}_i) \right) + \lambda \sum_{j=1}^{\infty} \frac{c_j^2}{a_j} \right]$$

- In the next theorem it is shown that the solution is finite-dimensional, and has the form

$$f(\mathbf{x}) = \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}_i)$$

GENERALIZATION: REPRESENTER THEOREM

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

$$\arg \min_{h \in \mathbb{H}} F(h) = \arg \min_{h \in \mathbb{H}} G(\|h\|_{\mathbb{H}}) + L(h(\mathbf{x}_1), \dots, h(\mathbf{x}_m))$$

admits a solution of the form

$$h^* = \sum_{i=1}^m \alpha_i K(\mathbf{x}_i, \cdot)$$

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

GENERALIZATION: REPRESENTER THEOREM

- **Proof:** let $\mathbb{H}_1 = \text{span}(\{K(\mathbf{x}_i, \cdot) : i \in [1, m]\})$. Any $h \in \mathbb{H}$ admits the decomposition $h = h_1 + h^\perp$ according to $\mathbb{H} = \mathbb{H}_1 \oplus \mathbb{H}_1^\perp$
 - Since G is non-decreasing,

$$G(\|h_1\|_{\mathbb{H}}) \leq G\left(\sqrt{\|h_1\|_{\mathbb{H}}^2 + \|h^\perp\|_{\mathbb{H}}^2}\right) = G(\|h\|_{\mathbb{H}})$$

- By the reproducing property, for all $i \in [1, m]$

$$h(\mathbf{x}_i) = \langle h, K(\mathbf{x}_i, \cdot) \rangle = \langle h_1, K(\mathbf{x}_i, \cdot) \rangle = h_1(\mathbf{x}_i)$$

- Thus, $L(h(\mathbf{x}_1), \dots, h(\mathbf{x}_m)) = L(h_1(\mathbf{x}_1), \dots, h_1(\mathbf{x}_m))$ 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 \mathbb{H}_1

KERNEL-BASED ALGORITHMS

- PDS kernels are used to extend a variety of algorithms in classification and other areas
 - regression
 - ranking
 - dimensionality reduction
 - clustering
- How to define PDS kernels?

- 1 MOTIVATION
- 2 KERNELS
- 3 SVMs WITH KERNELS
- 4 CLOSURE PROPERTIES**
- 5 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

PROOF OF CLOSURE PROPERTIES

- **Proof:**

- closure under *sum*

$$\mathbf{c}^T \mathbf{K} \mathbf{c} \geq 0 \text{ and } \mathbf{c}^T \mathbf{K}' \mathbf{c} \geq 0 \Rightarrow \mathbf{c}^T (\mathbf{K} + \mathbf{K}') \mathbf{c} \geq 0$$

- closure under *product*: $\mathbf{K} = \mathbf{M} \mathbf{M}^T$

$$\begin{aligned} \sum_{i,j=1}^m c_i c_j (\mathbf{K}_{ij} \mathbf{K}'_{ij}) &= \sum_{i,j=1}^m c_i c_j \left(\left[\sum_{k=1}^m \mathbf{M}_{ik} \mathbf{M}_{jk} \right] \mathbf{K}'_{ij} \right) \\ &= \sum_{k=1}^m \left[\sum_{i,j=1}^m c_i c_j \mathbf{M}_{ik} \mathbf{M}_{jk} \mathbf{K}'_{ij} \right] = \sum_{k=1}^m \mathbf{z}_k^T \mathbf{K}' \mathbf{z}_k \geq 0 \end{aligned}$$

$$\text{with } \mathbf{z}_k = \begin{bmatrix} c_1 \mathbf{M}_{1k} \\ \vdots \\ c_m \mathbf{M}_{mk} \end{bmatrix}$$

PROOF OF CLOSURE PROPERTIES

- Closure under *tensor product*

- definition: for all $\mathbf{x}_1, \mathbf{z}_1, \mathbf{x}_2, \mathbf{z}_2 \in X$

$$(K_1 \oplus K_2)(\mathbf{x}_1, \mathbf{z}_1, \mathbf{x}_2, \mathbf{z}_2) = K_1(\mathbf{x}_1, \mathbf{x}_2)K_2(\mathbf{z}_1, \mathbf{z}_2)$$

- thus PDS kernel as a product of the kernels

$$(\mathbf{x}_1, \mathbf{z}_1, \mathbf{x}_2, \mathbf{z}_2) \rightarrow K_1(\mathbf{x}_1, \mathbf{x}_2), (\mathbf{x}_1, \mathbf{z}_1, \mathbf{x}_2, \mathbf{z}_2) \rightarrow K_2(\mathbf{z}_1, \mathbf{z}_2)$$

- closure under *pointwise limit*: if for all $\mathbf{x}, \mathbf{z} \in X$

$$\lim_{n \rightarrow \infty} K_n(\mathbf{x}, \mathbf{z}) = K(\mathbf{x}, \mathbf{z}),$$

Then,

$$(\forall n, \mathbf{c}^T K_n \mathbf{c}) \Rightarrow \lim_{n \rightarrow \infty} \mathbf{c}^T K_n \mathbf{c} = \mathbf{c}^T K \mathbf{c} \geq 0$$

PROOF OF CLOSURE PROPERTIES

- Closure under *composition with power series*
 - assumption: K is a PDS kernel with $|K(\mathbf{x}, \mathbf{z})| < \rho$ for all $\mathbf{x}, \mathbf{z} \in X$ and $f(\mathbf{x}) = \sum_{n=0}^{\infty} a_n x^n$, $a_n \geq 0$ is a power series with radius of convergence ρ
 - $f \circ K$ is a PDS kernel since K^n is a 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

- 1 MOTIVATION
- 2 KERNELS
- 3 SVMs WITH KERNELS
- 4 CLOSURE PROPERTIES
- 5 NEGATIVE KERNELS**

MOTIVATION

- 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

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

$$\mathbf{c}^T \mathbf{K} \mathbf{c} \leq 0$$

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

EXAMPLES

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

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

EXAMPLES — CONT.'D

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

NDS KERNELS — MAIN PROPERTY

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

$$\forall \mathbf{x}, \mathbf{z} \in X, K(\mathbf{x}, \mathbf{z}) = \|\Phi(\mathbf{x}) - \Phi(\mathbf{z})\|^2$$

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

PDS AND NDS KERNELS

- **Theorem:** Let $K : X \times X \rightarrow \mathbb{R}$ 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 \mathbf{x}_0 by
$$K'(\mathbf{x}, \mathbf{z}) = K(\mathbf{x}, \mathbf{x}_0) + K(\mathbf{z}, \mathbf{x}_0) - K(\mathbf{x}, \mathbf{z}) - K(\mathbf{x}_0, \mathbf{x}_0)$$
for all $\mathbf{x}, \mathbf{z} \in X$. Then K is NDS iff K' is PDS

EXAMPLE

- The kernel defined by $K(\mathbf{x}, \mathbf{z}) = \exp(-t\|\mathbf{x} - \mathbf{z}\|^2)$ is PDS for all $t > 0$ since $\|\mathbf{x} - \mathbf{z}\|^2$ is NDS
- The kernel $\exp(-|x - z|^p)$ is not PDS for $p > 2$. Otherwise, for any $t > 0$, $\{x_1, \dots, x_m\} \subseteq X$ 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 $|x - z|^p$ is NDS for $p > 2$, but that is not true (prove!!!)