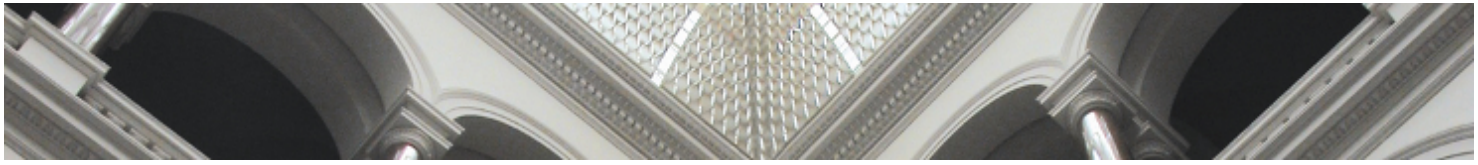
Technische Universität Berlin

Kernels I

Nico Görnitz
Technische Universität Berlin - Machine Learning Group
Beginner's Workshop Machine Learning 2018



Technische Universität Berlin

Agenda

Feature Maps

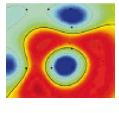
- Input Space vs. Feature Space
- Example I: The XOR Problem
- Example II: Ordinary Least Squares

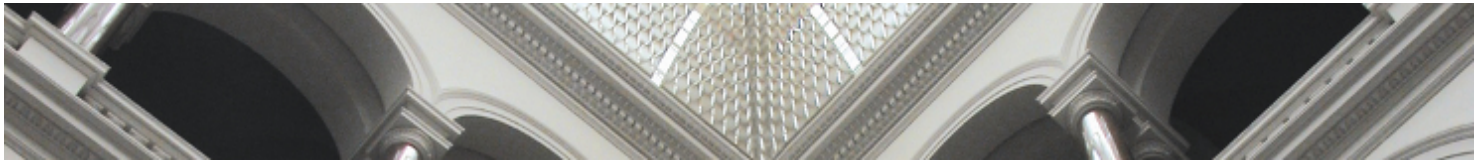
Kernels

- From Feature Space to Kernels
- Examples of Kernels
- Mercer Kernels and PSD Kernels
- Calculation Rules
- Representer Theorem

Reproducing Kernel Hilbert Spaces

- Hilbert Space and RKHS
- Feature Maps, Kernels, and RKHS: The connection





Technische Universität Berlin

Agenda

Feature Maps

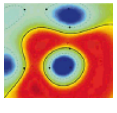
- Input Space vs. Feature Space
- Example I: The XOR Problem
- Example II: Ordinary Least Squares

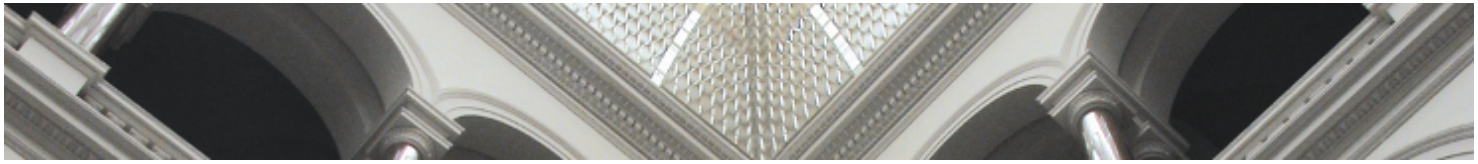
Kernels

- From Feature Space to Kernels
- Examples of Kernels
- Mercer Kernels and PSD Kernels
- Calculation Rules
- Representer Theorem

Reproducing Kernel Hilbert Spaces

- Hilbert Space and RKHS
- Feature Maps, Kernels, and RKHS: The connection





Technische Universität Berlin

Agenda

Feature Maps

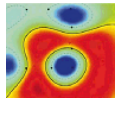
- Input Space vs. Feature Space
- Example I: The XOR Problem
- Example II: Ordinary Least Squares

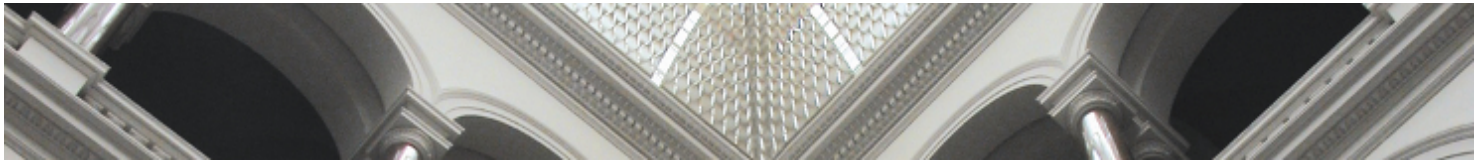
Kernels

- From Feature Space to Kernels
- Examples of Kernels
- Mercer Kernels and PSD Kernels
- Calculation Rules
- Representer Theorem

Reproducing Kernel Hilbert Spaces

- Hilbert Space and RKHS
- Feature Maps, Kernels, and RKHS: The connection

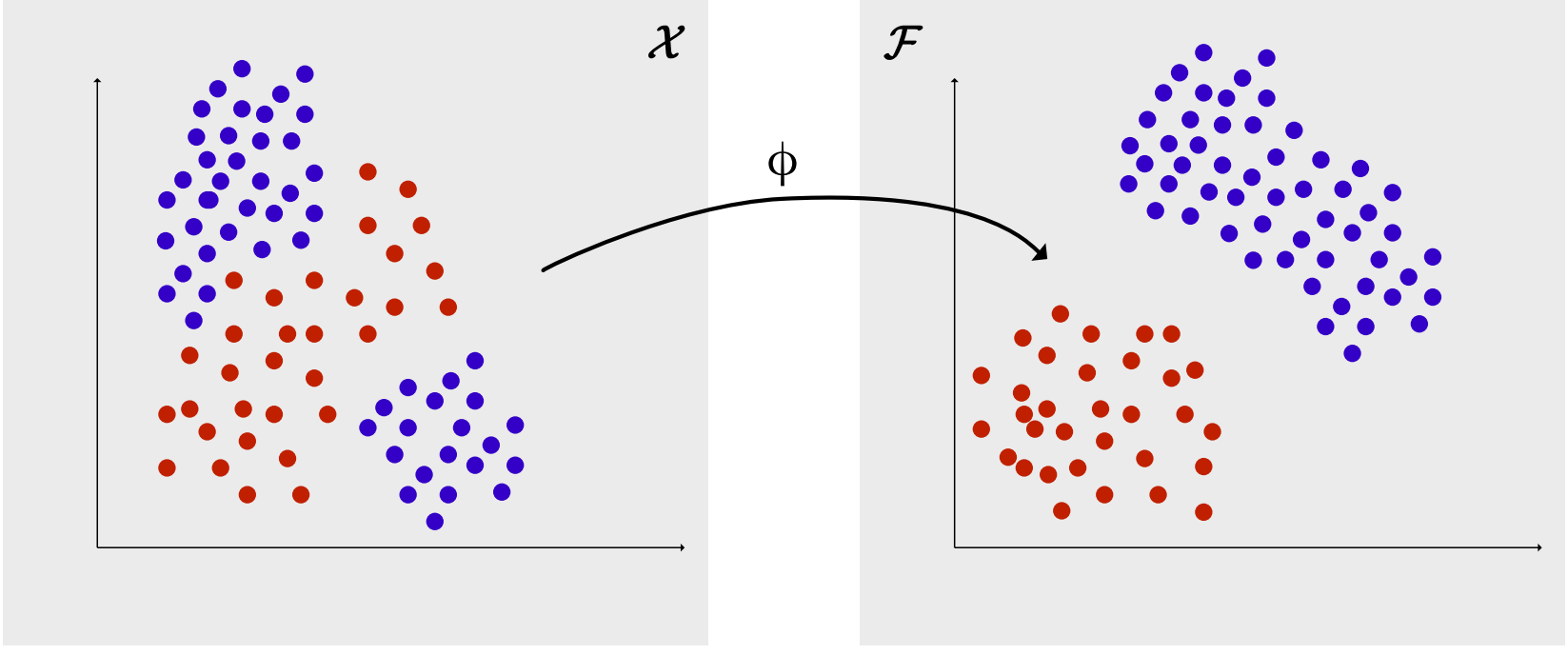


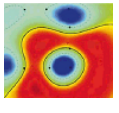


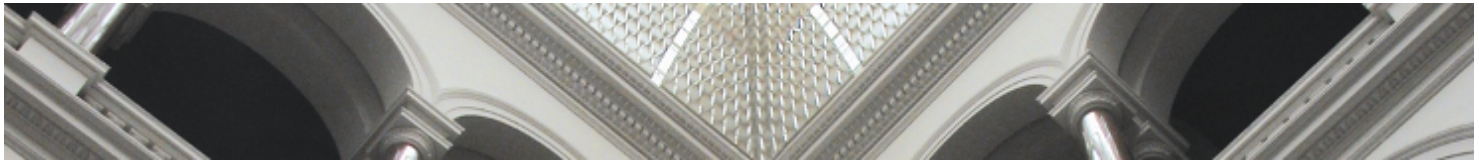
Technische Universität Berlin

Input Space vs. Feature Space

We want to transform $\phi : \mathcal{X} \rightarrow \mathcal{F}$ our data from our input space $\mathcal{X}$ to some other 'feature' space $\mathcal{F}$ such that our problem (classification, regression, etc) becomes easier.







Technische Universität Berlin

Agenda

Feature Maps

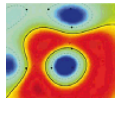
- Input Space vs. Feature Space
- Example I: The XOR Problem
- Example II: Ordinary Least Squares

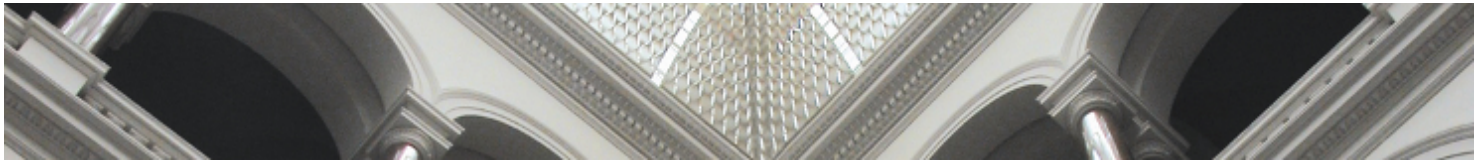
Kernels

- From Feature Space to Kernels
- Examples of Kernels
- Mercer Kernels and PSD Kernels
- Calculation Rules
- Representer Theorem

Reproducing Kernel Hilbert Spaces

- Hilbert Space and RKHS
- Feature Maps, Kernels, and RKHS: The connection

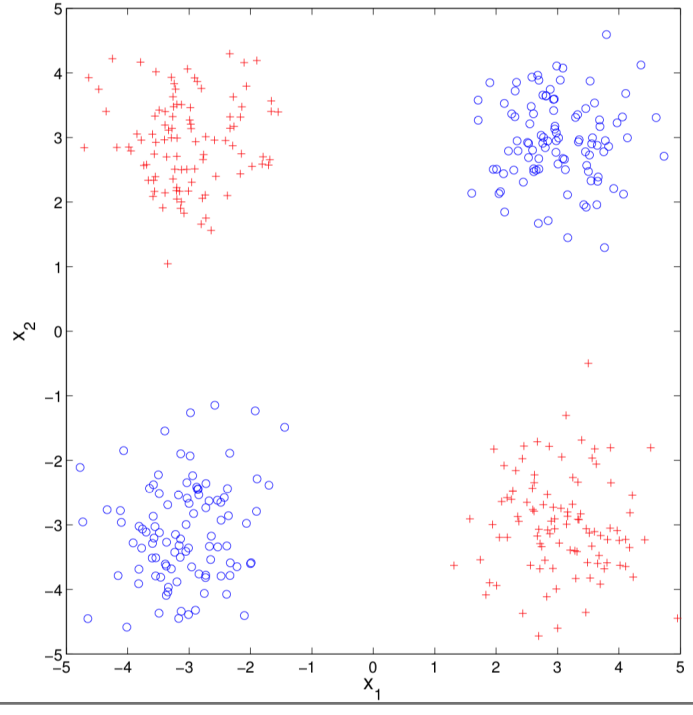


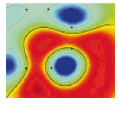


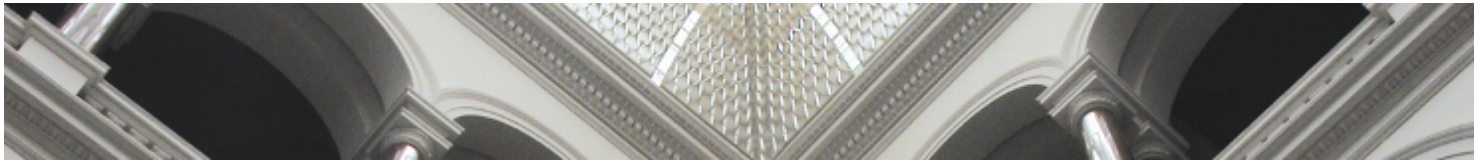
Technische Universität Berlin

Example: The XOR Problem

We would like to do binary classification using a linear model and with the data looking like below (hence, our input space is $\mathcal{X} = \mathbb{R}^2$). This is not linear separable! However, using the following feature map $\phi(x) = [x_1, x_2, x_1 x_2]$ the problem becomes linear separable in the 3 dimensional feature space $\mathcal{F} = \mathbb{R}^3$.

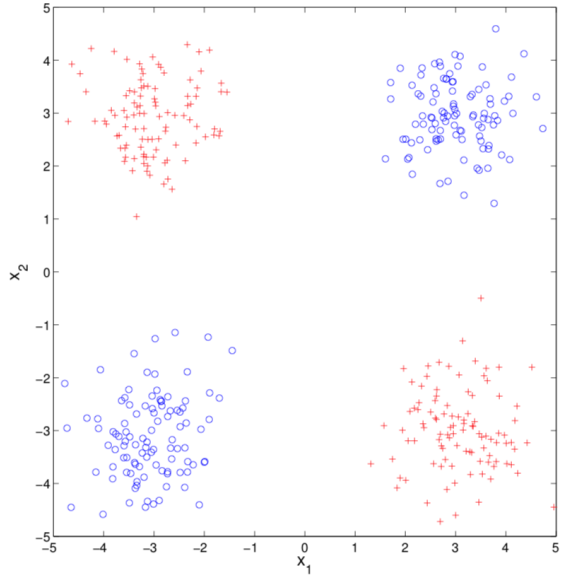


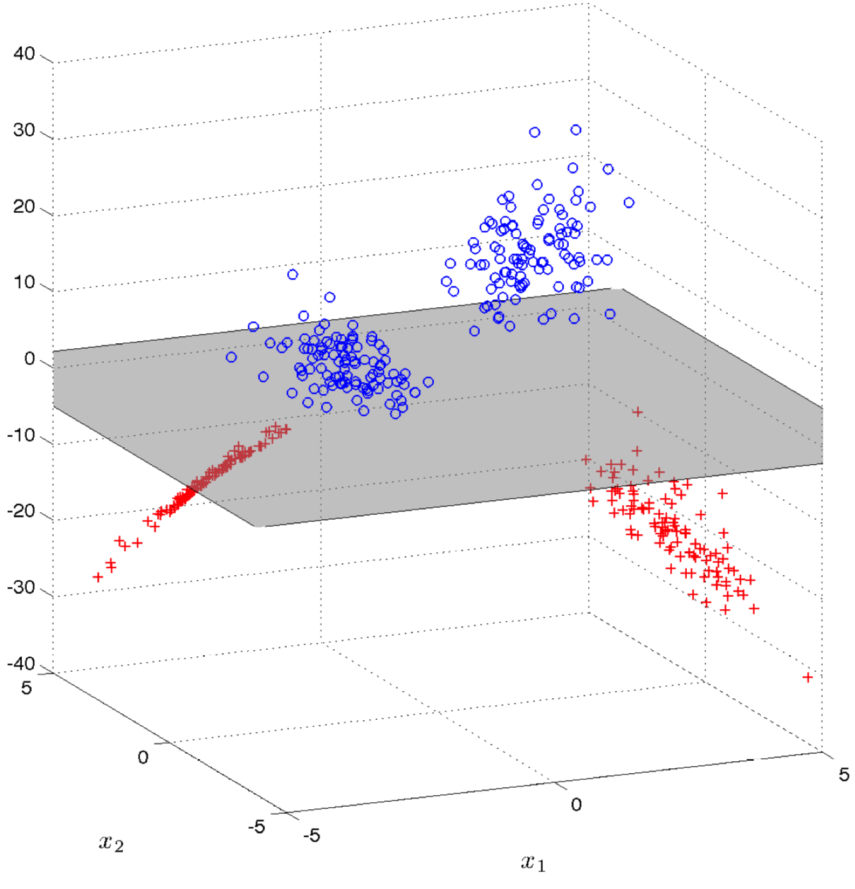


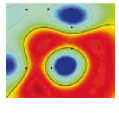


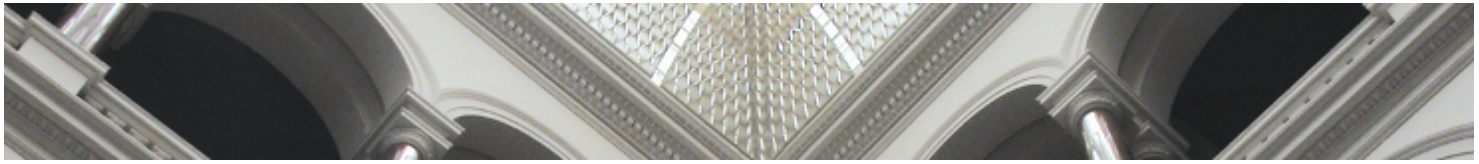
Technische Universität Berlin

Example I: The XOR Problem









Technische Universität Berlin

Agenda

Feature Maps

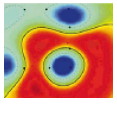
- Input Space vs. Feature Space
- Example I: The XOR Problem
- Example II: Ordinary Least Squares

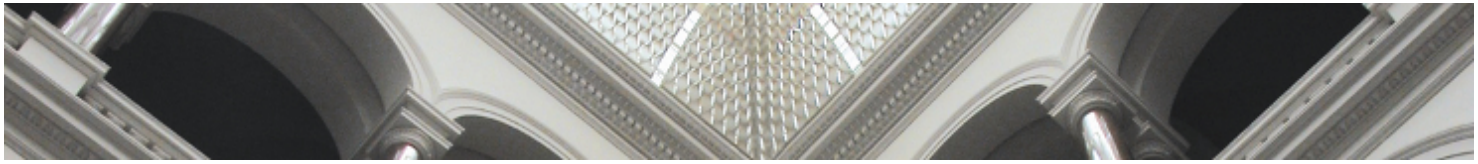
Kernels

- From Feature Space to Kernels
- Examples of Kernels
- Mercer Kernels and PSD Kernels
- Calculation Rules
- Representer Theorem

Reproducing Kernel Hilbert Spaces

- Hilbert Space and RKHS
- Feature Maps, Kernels, and RKHS: The connection



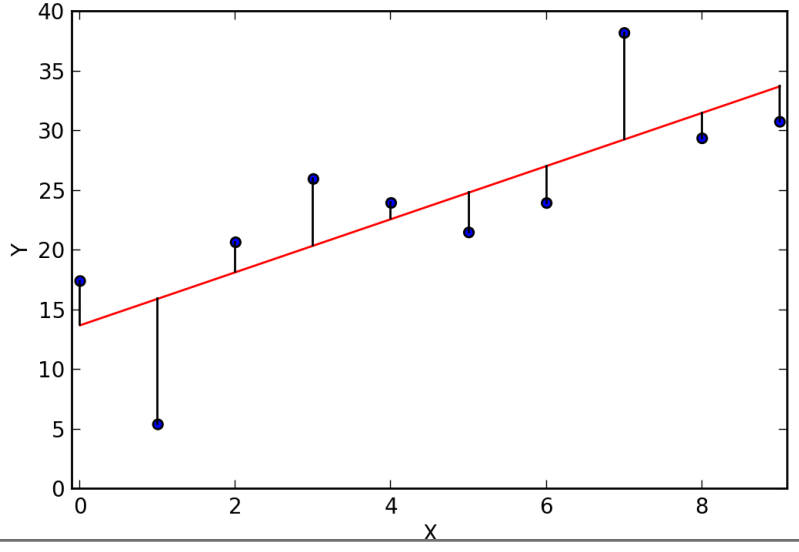


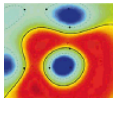
Technische Universität Berlin

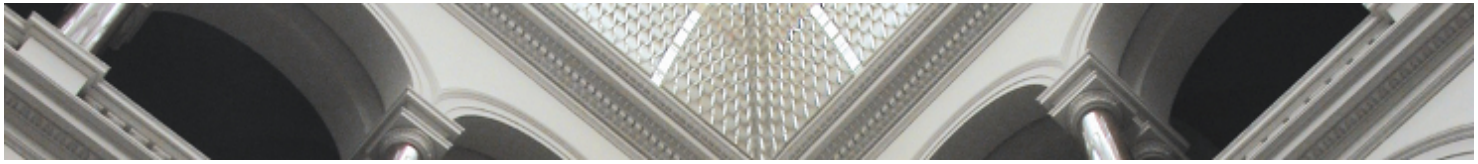
Ordinary Least Squares (OLS)

Consider a simple linear model $f(x) = \langle w, x \rangle$ with x and $w \in \mathbb{R}^d$ (d-dimensional). Further, assume that we are given a sample of size $i = 1, \ldots, n$ of i.i.d. data points $x_i \in \mathbb{R}^d$ and corresponding labels $y_i \in \mathbb{R}$ which we will use to fit our parameter vector w to produce the least squared error on that given sample,

$$w^* = \arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \ell(w, x_i, y_i) \quad \text{with} \quad \ell(w, x, y) := \frac{1}{2}(y - \langle w, x \rangle)^2 .$$







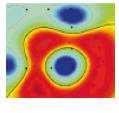
Technische Universität Berlin

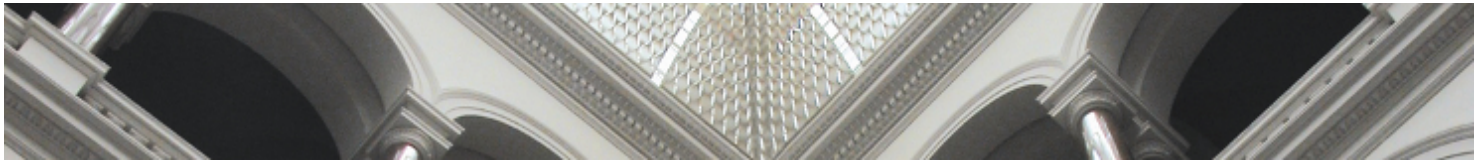
Ordinary Least Squares (OLS)

For the sake of simplicity, we employ stochastic gradient descent as an optimization technique. Therefore, at time step t we pick a data point x_t and the corresponding label y_t from our training sample and update the parameter vector according to the following formula (with $w_0 = 0$):

$$\begin{aligned} w_{t+1} &= w_t - \eta \frac{\partial \ell(w_t, x_t, y_t)}{\partial w} \\ &= w_t + \underbrace{\eta\big(y_t - \langle w_t, x_t \rangle\big)}_{\alpha_t \in \mathbb{R}} x_t \\ &= w_t + \alpha_t x_t \,. \end{aligned}$$

Here, $\eta > 0$ is a constant (”step size”).





Technische Universität Berlin

Ordinary Least Squares (OLS)

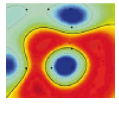
Rather surprisingly, it will attain the optimal value while not leaving the span of the data. Therefore, the parameter vector w^T (after T iterations) can be expressed as a weighted sum of feature vectors:

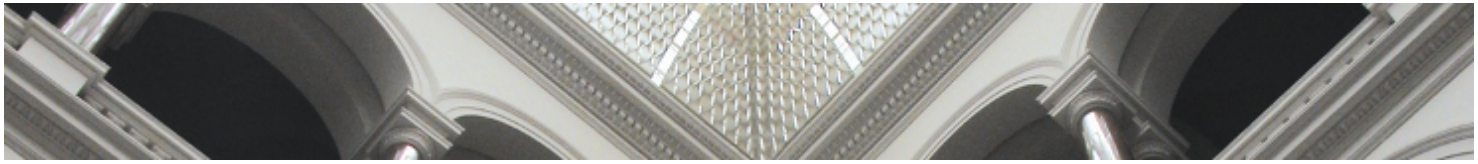
$$w^T = \sum_{t=1}^{T} \alpha_t x_t \, .$$

Given this expansion, we could also re-place our linear model function with

$$f(x) = \langle w, x \rangle = \sum_{t} \alpha_t \langle x_t, x \rangle \, .$$

This gives an alternative view on the above optimization problem, where the key is to find weightings of similarities, as encoded with inner products, between data points.





Technische Universität Berlin

Agenda

Feature Maps

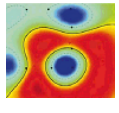
- Input Space vs. Feature Space
- Example I: The XOR Problem
- Example II: Ordinary Least Squares

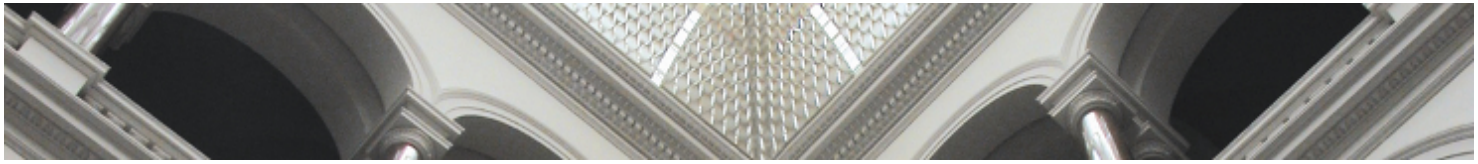
Kernels

- From Feature Space to Kernels
- Examples of Kernels
- Mercer Kernels and PSD Kernels
- Calculation Rules
- Representer Theorem

Reproducing Kernel Hilbert Spaces

- Hilbert Space and RKHS
- Feature Maps, Kernels, and RKHS: The connection





Technische Universität Berlin

Agenda

Feature Maps

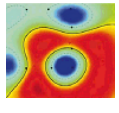
- Input Space vs. Feature Space
- Example I: The XOR Problem
- Example II: Ordinary Least Squares

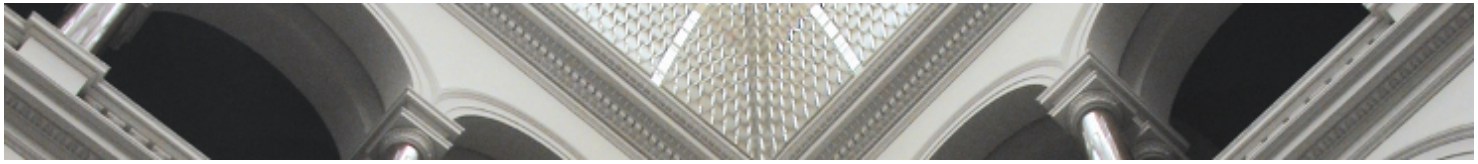
Kernels

- From Feature Space to Kernels
- Examples of Kernels
- Mercer Kernels and PSD Kernels
- Calculation Rules
- Representer Theorem

Reproducing Kernel Hilbert Spaces

- Hilbert Space and RKHS
- Feature Maps, Kernels, and RKHS: The connection





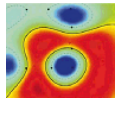
Ordinary Least Squares (OLS) with Kernels

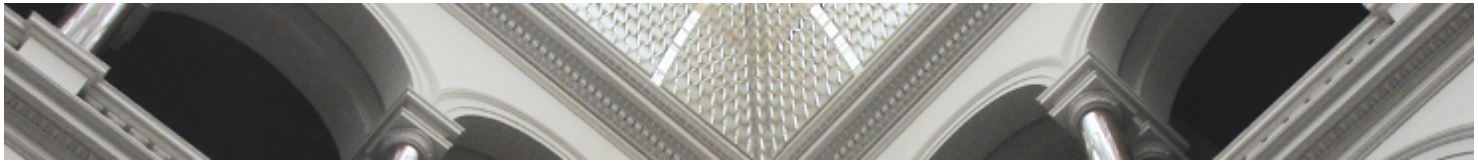
Let us consider, again, the linear model $f(x) = \langle w, x \rangle$ but now, we transform x using the feature map $\phi : \mathbb{R}^d \to \mathcal{F}$, i.e. $f(x) = \langle w, \phi(x) \rangle$ with x and $w \in \mathcal{F}$ (possibly very high-dimensional).

$$w^* = \arg\min_{w \in \mathcal{F}} \sum_{i=1}^{n} \ell(w, x_i, y_i) \quad \text{with} \quad \ell(w, x, y) := \frac{1}{2}(y - \langle w, \phi(x) \rangle)^2 .$$

Nothing essentially changed and the above result still holds:

$$f(x) = \langle w, \phi(x) \rangle = \sum_t \alpha_t \underbrace{\langle \phi(x_t), \phi(x) \rangle}_{=:K(x_t, x)} = \sum_t \alpha_t K(x_t, x) .$$





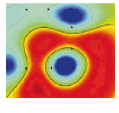
Technische Universität Berlin

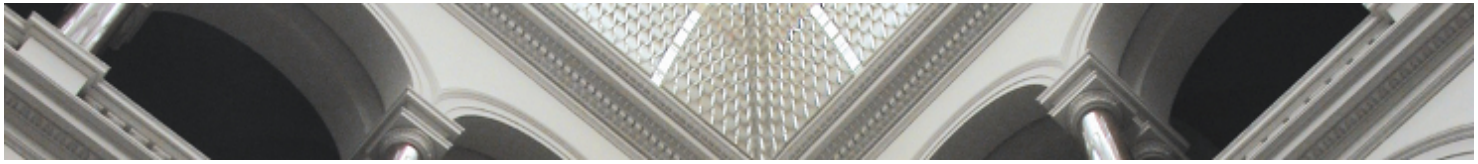
Ordinary Least Squares (OLS) with Kernels

So, two things just happened:

$$f(x) = \langle w, \phi(x) \rangle = \sum_t \alpha_t K(x_t, x) \,.$$

1. Instead of designing and accessing an appropriate (high-dimensional) feature space directly, we only need to measure similarities between examples using an lookup table K. This is called the **kernel trick** and K is referred to as the kernel.
2. We have seen that the optimal solution of the ordinary least squares problem can be expressed as a weighted sum of training samples. This result can be generalized to a very large class of problems (**Representer Theorem)**.





Technische Universität Berlin

Agenda

Feature Maps

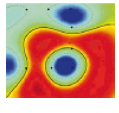
- Input Space vs. Feature Space
- Example I: The XOR Problem
- Example II: Ordinary Least Squares

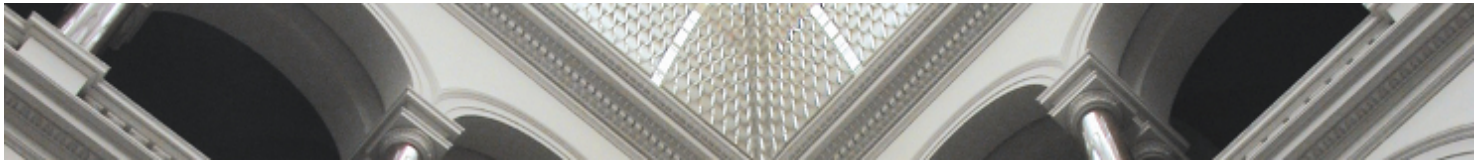
Kernels

- From Feature Space to Kernels
- Examples of Kernels
- Mercer Kernels and PSD Kernels
- Calculation Rules
- Representer Theorem

Reproducing Kernel Hilbert Spaces

- Hilbert Space and RKHS
- Feature Maps, Kernels, and RKHS: The connection





Technische Universität Berlin

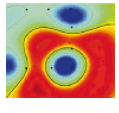
Standard Kernels

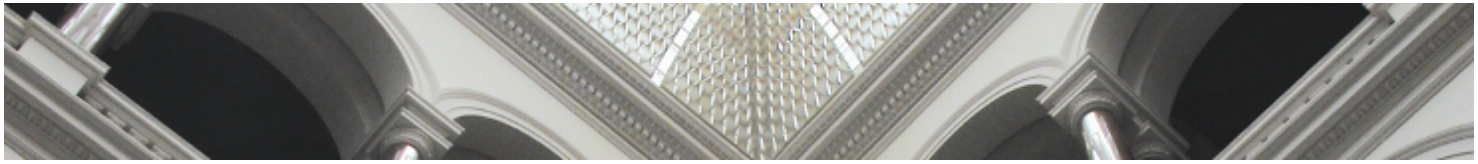
Albeit many possibilities exists for defining an proper kernel K, the following kernels appear frequently:

Linear kernel $K(x, x') := \langle x, x' \rangle$;

Polynomial kernel $K(x, x') := (\langle x, x' \rangle + c)^d$ with $c > 0$ and $d \in \mathbb{N}$;

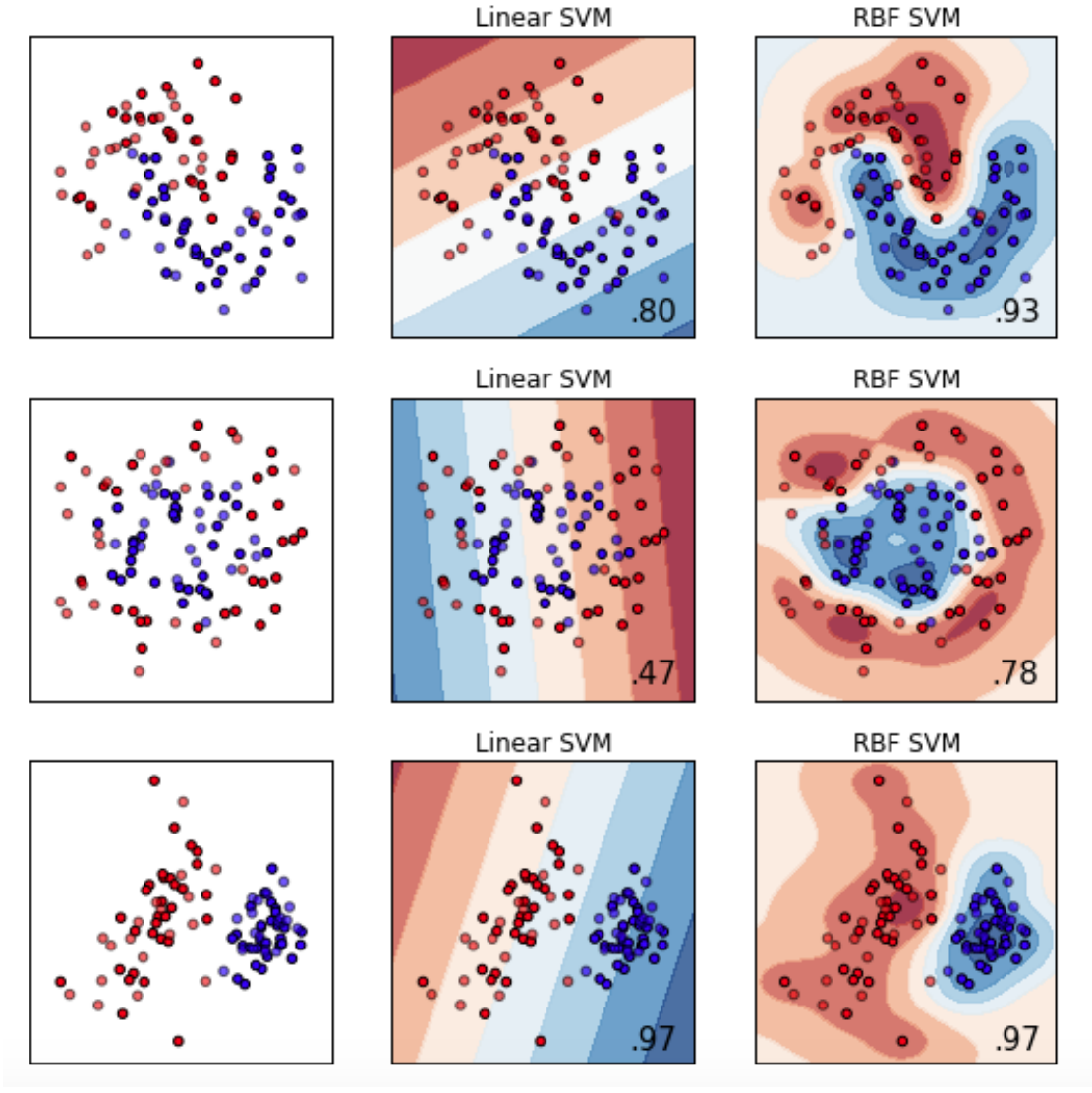
Radial basis function (RBF) kernel $K(x, x') := \exp(-\frac{1}{2\sigma^2} \|x - x'\|^2)$.

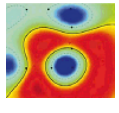


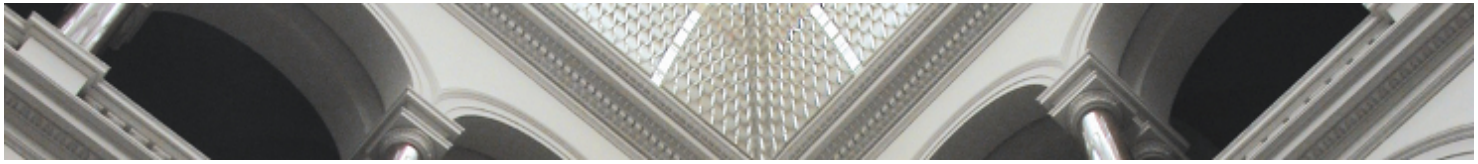


Technische Universität Berlin

Example: RBF kernel







Technische Universität Berlin

Agenda

Feature Maps

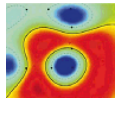
- Input Space vs. Feature Space
- Example I: The XOR Problem
- Example II: Ordinary Least Squares

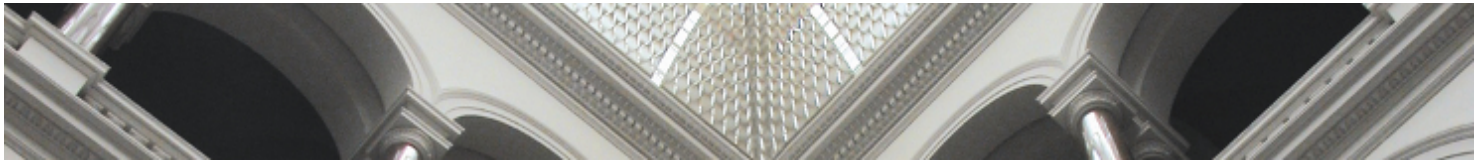
Kernels

- From Feature Space to Kernels
- Examples of Kernels
- Mercer Kernels and PSD Kernels
- Calculation Rules
- Representer Theorem

Reproducing Kernel Hilbert Spaces

- Hilbert Space and RKHS
- Feature Maps, Kernels, and RKHS: The connection





Technische Universität Berlin

When is a kernel a kernel?

Albeit this definition allows to encode arbitrary functions, in order to ensure that a decomposition into feature vectors $K(x, x') = \langle \phi(x), \phi(x') \rangle$ exists, K needs to satisfy the following condition:

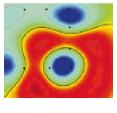
Theorem (Mercer's condition)

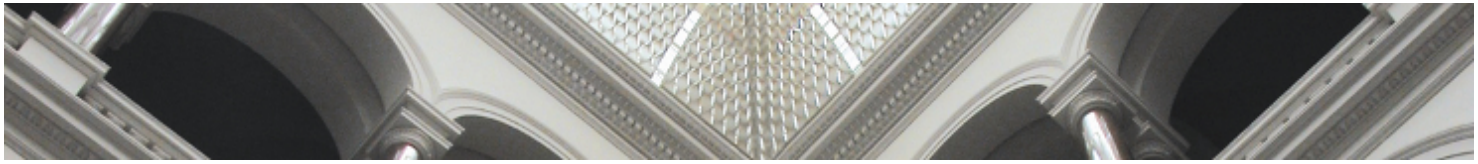
Let $\mathcal{X} \subset \mathbb{R}^N$ be a compact set and let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous and symmetric function. Then, K admits a uniformly convergent expansion of the form

$$K(x, x') = \sum_{i=0}^{\infty} a_i \phi_i(x) \phi_i(x') \,,$$

with $a_i > 0$ iff for any square integrable function c ($c \in L_2(\mathcal{X})$), the following condition holds:

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





Technische Universität Berlin

When is a kernel a kernel?

We present another, slightly more general and approachable definition which ensures the existence of the decomposition $K(x, x') = \langle \phi(x), \phi(x') \rangle$:

Definition (Positive definite symmetric kernels)

A kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be positive definite symmetric (PDS) if for any $\{x_1, \ldots, x_n\} \subseteq \mathcal{X}$, the matrix $K = [K(x_i, x_j)]_{ij} \in \mathbb{R}^{n \times n}$ is symmetric positive semidefinite (SPSD).

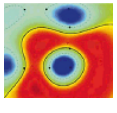
Symmetric Matrix:

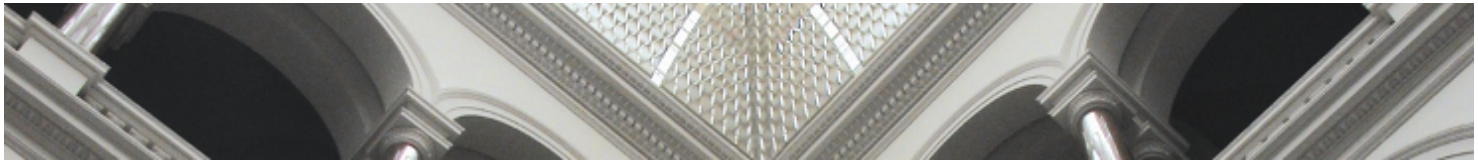
$$A = A^\mathsf{T}$$

Positive semi-definite Matrix:

$$x^\mathsf{T} A x \geq 0 \quad \forall\, x$$

alternatively, all Eigenvalues λ of A non-negative.





Technische Universität Berlin

Agenda

Feature Maps

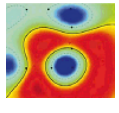
- Input Space vs. Feature Space
- Example I: The XOR Problem
- Example II: Ordinary Least Squares

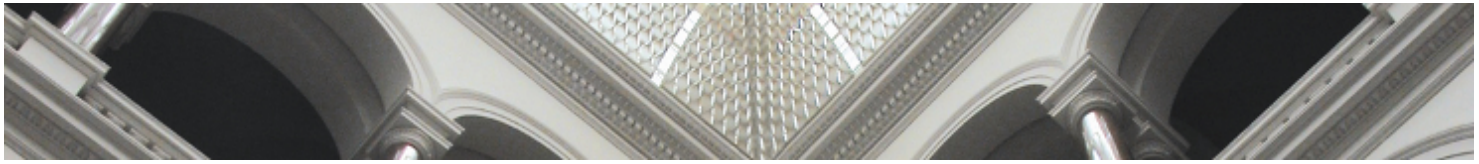
Kernels

- From Feature Space to Kernels
- Examples of Kernels
- Mercer Kernels and PSD Kernels
- Calculation Rules
- Representer Theorem

Reproducing Kernel Hilbert Spaces

- Hilbert Space and RKHS
- Feature Maps, Kernels, and RKHS: The connection



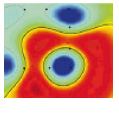


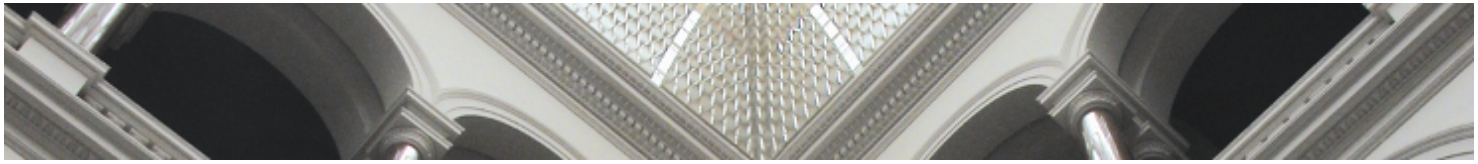
Technische Universität Berlin

Inner Products and Norms

$$\|\phi(x)\| = \sqrt{\langle \phi(x), \phi(x) \rangle} = \sqrt{K(x,x)}$$

$$\|\phi(x) - \phi(y)\| = \sqrt{K(x,x) - 2K(x,y) + K(y,y)}$$



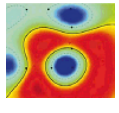


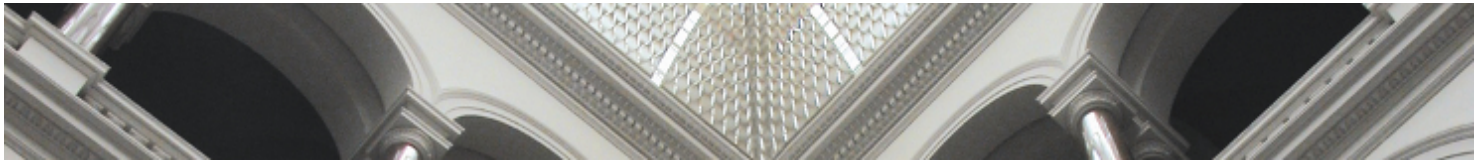
Technische Universität Berlin

Constructing kernels from kernels

If $K_1 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $K_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are Mercer kernels, then

$$K(x, y) = K_1(x, y) + K_2(x, y)$$
$$K(x, y) = cK_1(x, y) + K_2(x, y), \quad c \in \mathbb{R}^+$$
$$K(x, y) = K_1(x, y) + c, \quad c \in \mathbb{R}^+$$
$$K(x, y) = K_1(x, y)K_2(x, y)$$
$$K(x, y) = f(x)f(y), \quad f : \mathcal{X} \to \mathbb{R}$$
$$K(x, y) = (K_1(x, y) + c)^d, \quad c \in \mathbb{R}^+, \text{ and } d \in \mathbb{N}$$
$$K(x, y) = \exp(K_1(x, y)/\sigma^2), \quad \sigma \in \mathbb{R}$$
$$K(x, y) = \exp(-(K_1(x, x) - 2K_1(x, y) + K_1(y, y))/2\sigma^2), \quad \sigma \in \mathbb{R}$$
$$K(x, y) = K_1(x, y)/\sqrt{K_1(x, x)K_1(y, y)}$$



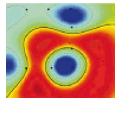


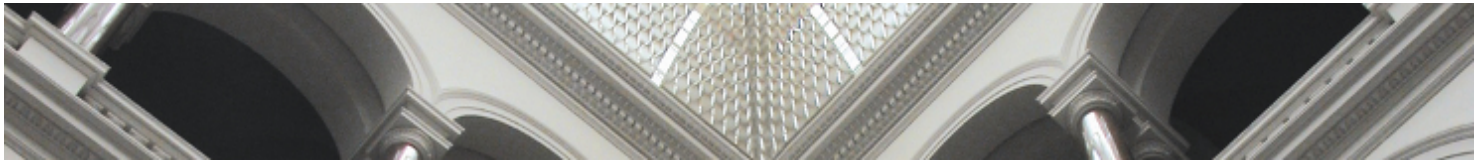
Spherical Normalization

We can not only use those rules to construct new kernels but also to manipulate data points in feature space. In fact, the last point actually normalizes data points in feature space:

$$
\begin{aligned}
K(x,y) &= \frac{K_1(x,y)}{\sqrt{K_1(x,x)K_1(y,y)}} \\
&= \frac{\langle \phi(x), \phi(y) \rangle}{\sqrt{\|\phi(x)\|^2 \|\phi(y)\|^2}} \\
&= \frac{\langle \phi(x), \phi(y) \rangle}{\|\phi(x)\| \|\phi(y)\|} \\
&= \langle \frac{\phi(x)}{\|\phi(x)\|}, \frac{\phi(y)}{\|\phi(y)\|} \rangle
\end{aligned}
$$

So, all feature vectors have the same distance from the origin and hence, the data points lie on the surface of a hypersphere.





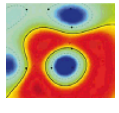
Multiplicative Normalization

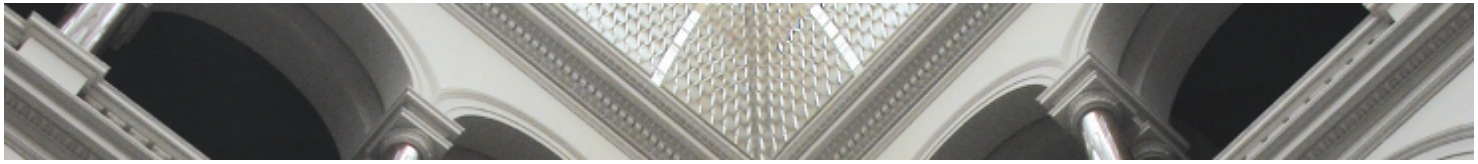
Spherical normalization removes the 'length' information of features. Sometimes this is not intended. Ong and Zien (2008) introduced the multiplicative normalization where features are normalized to have uniform variance:

$$\frac{1}{n}\sum_{i=1}^{n}\|\phi(x_i)-\phi(\bar{x})\|^2 = 1$$

$$K(x,y) = \frac{K_1(x,y)}{\frac{1}{n}\sum_i K_1(x_i,x_i) - \frac{1}{n^2}\sum_i\sum_j K_1(x_i,x_j)}$$

So, all feature vectors have the same distance from the origin and hence, the data points lie on the surface of a hypersphere.





Technische Universität Berlin

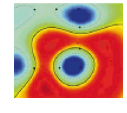
Centering

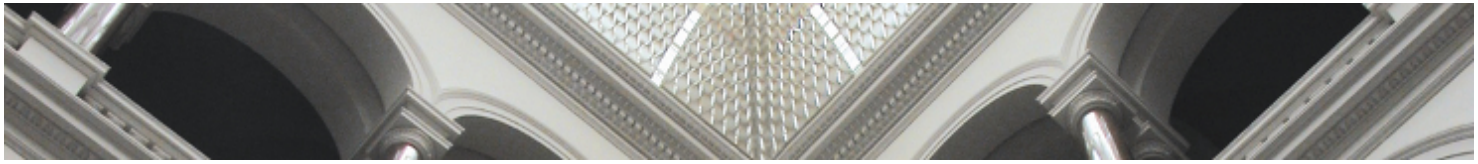
We now want to center the data in **feature space**, i.e. subtracting the mean of the dataset ($\frac{1}{n}\sum_{i=1}^{n}\phi(x_i)$) from each entry:

$$K(x,y) = \langle \phi(x) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i), \phi(y) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\rangle$$

$$= \langle \phi(x), \phi(y)\rangle - \langle \frac{1}{n}\sum_{i=1}^{n}\phi(x_i), \phi(x)\rangle - \langle \frac{1}{n}\sum_{i=1}^{n}\phi(x_i), \phi(y)\rangle + \langle \frac{1}{n}\sum_{i=1}^{n}\phi(x_i), \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)$$

$$= K(x,y) - \frac{1}{n}\sum_{i=1}^{n}K(x_i,x) - \frac{1}{n}\sum_{i=1}^{n}K(x_i,y) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K(x_i,x_j)$$





Technische Universität Berlin

Agenda

Feature Maps

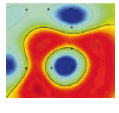
- Input Space vs. Feature Space
- Example I: The XOR Problem
- Example II: Ordinary Least Squares

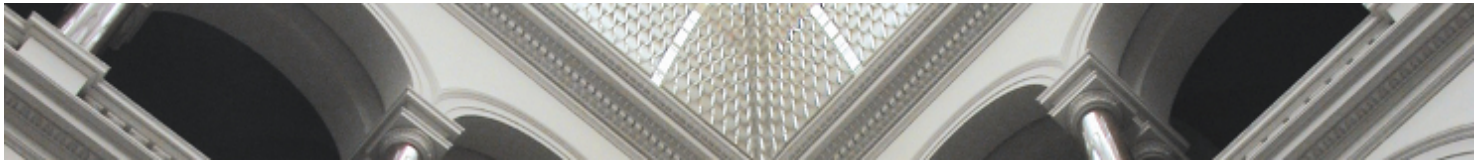
Kernels

- From Feature Space to Kernels
- Examples of Kernels
- Mercer Kernels and PSD Kernels
- Calculation Rules
- Representer Theorem

Reproducing Kernel Hilbert Spaces

- Hilbert Space and RKHS
- Feature Maps, Kernels, and RKHS: The connection





Technische Universität Berlin

Representer Theorem

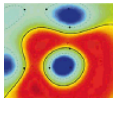
Finally, we can give a concise description of the existence of the expansion in the previous OLS with gradient descent example. The representer theorem states that if a given optimization problem can be rephrased in a specific form, then the optimal solution of this optimization problem must live in the span of the data.

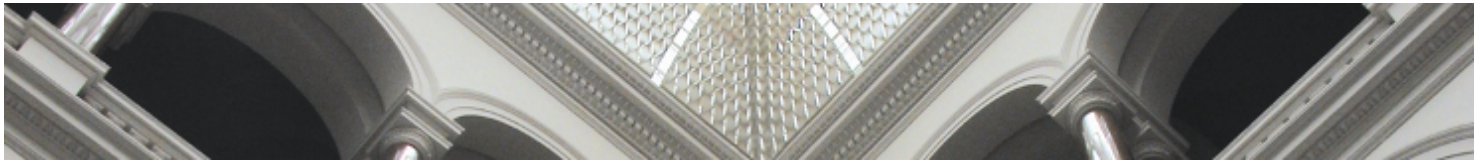
Theorem (Representer Theorem)

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel and $\mathcal{F}$ its corresponding RKHS. Then, for any non-decreasing function $G : \mathbb{R} \to \mathbb{R}$ and any loss function $L : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, the optimization problem

$$\arg\min_{f \in \mathcal{F}} G(\|f\|_{\mathcal{F}}) + L\left(f(x_1), \ldots, f(x_n)\right)$$

admits a solution of the form $f^* = \sum_{i=1}^n \alpha_i k(x_i, \cdot) = \sum_{i=1}^n \alpha_i \phi(x_i)$. If G is further assumed to be increasing, then any solution has this form.





Technische Universität Berlin

Agenda

Feature Maps

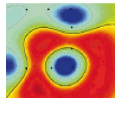
- Input Space vs. Feature Space
- Example I: The XOR Problem
- Example II: Ordinary Least Squares

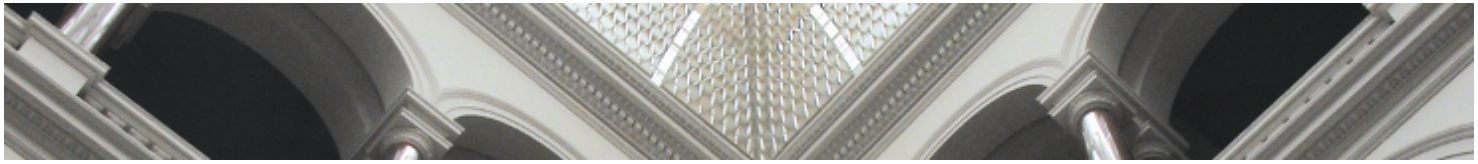
Kernels

- From Feature Space to Kernels
- Examples of Kernels
- Mercer Kernels and PSD Kernels
- Calculation Rules
- Representer Theorem

Reproducing Kernel Hilbert Spaces

- Hilbert Space and RKHS
- Feature Maps, Kernels, and RKHS: The connection





Technische Universität Berlin

Agenda

Feature Maps

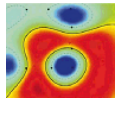
- Input Space vs. Feature Space
- Example I: The XOR Problem
- Example II: Ordinary Least Squares

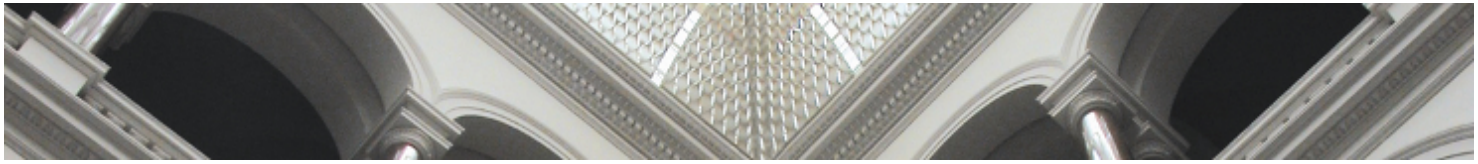
Kernels

- From Feature Space to Kernels
- Examples of Kernels
- Mercer Kernels and PSD Kernels
- Calculation Rules
- Representer Theorem

Reproducing Kernel Hilbert Spaces

- Hilbert Space and RKHS
- Feature Maps, Kernels, and RKHS: The connection





Technische Universität Berlin

Hilbert Space

Definition (Hilbert Space)

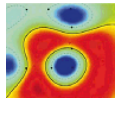
A Hilbert Space is a complete vector space equipped with an inner product.

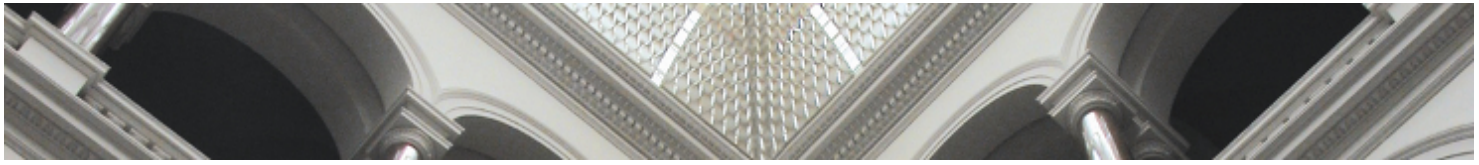
Properties (Inner Product)

The inner product $\langle f, g \rangle$ has the following properties:

- Symmetry: $\langle f, g \rangle = \langle g, f \rangle$
- Linearity: $\langle \alpha f_1 + \beta f_2, g \rangle = \alpha \langle f_1, g \rangle + \beta \langle f_2, g \rangle$
- Non-negativity: $\langle f, f \rangle \geq 0$
- Zero: $\langle f, f \rangle = 0 \Rightarrow f = 0$

Basically a ”nice” infinite dimensional vector space, where lots of things behave like the finite case (e.g. using inner product we can define ”norm” or ”orthogonality”)





Technische Universität Berlin

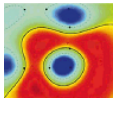
Reproducing Kernel Hilbert Space (RKHS)

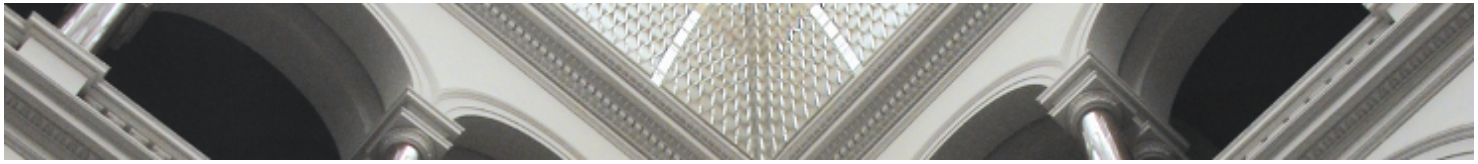
Now that the relation between feature maps and kernels is clear, we only need a relation between those entities with their respective reproducing kernel Hilbert space (RKHS) which comes in the form of the following definition:

Theorem (Reproducing kernel Hilbert space (first definition))

*Let $\mathcal{F}$ be a Hilbert space of $\mathbb{R}$-valued functions defined on a non-empty set $\mathcal{X}$. A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **(reproducing) kernel** of $\mathcal{F}$, and $\mathcal{F}$ is a reproducing kernel Hilbert space, if K satisfies*

1. $\forall\, x \in \mathcal{X},\ K(x, \cdot) \in \mathcal{F}$
2. $\forall\, x \in \mathcal{X}, \forall\, f \in \mathcal{F},\ \langle f, K(x, \cdot)\rangle_{\mathcal{F}} = f(x)$ *(the reproducing property)*





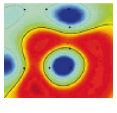
The connection: Kernel and RKHS

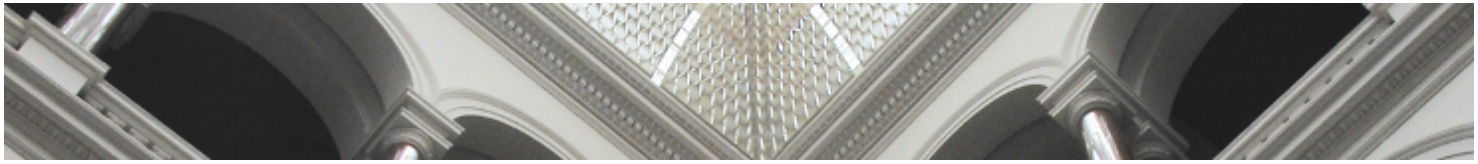
The reproducing kernel Hilbert space (RKHS) is a space of **functions** $\phi(x) = K(x, \cdot)$ and can be constructed by

$$\mathcal{F} = \{\sum_{i=1}^{\ell} \alpha_i K(x_i, \cdot) | \ell \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \mathcal{X}, i = 1, \ldots, \ell\}$$

Theorem (Moore-Aronszajn)

Every positive semi-definite kernel corresponds to a unique RKHS, and every RKHS is associated with a unique positive semi-definite kernel.





Technische Universität Berlin

Agenda

Feature Maps

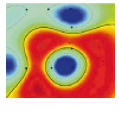
- Input Space vs. Feature Space
- Example I: The XOR Problem
- Example II: Ordinary Least Squares

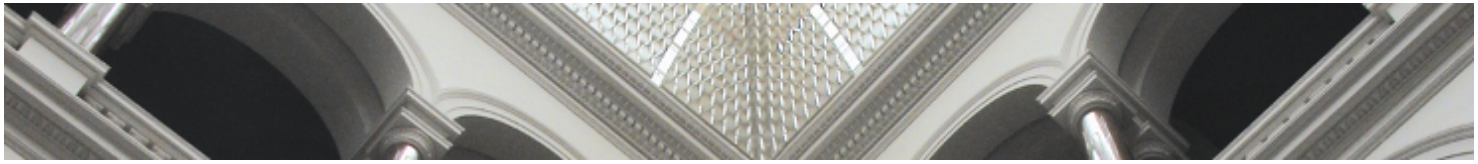
Kernels

- From Feature Space to Kernels
- Examples of Kernels
- Mercer Kernels and PSD Kernels
- Calculation Rules
- Representer Theorem

Reproducing Kernel Hilbert Spaces

- Hilbert Space and RKHS
- Feature Maps, Kernels, and RKHS: The connection

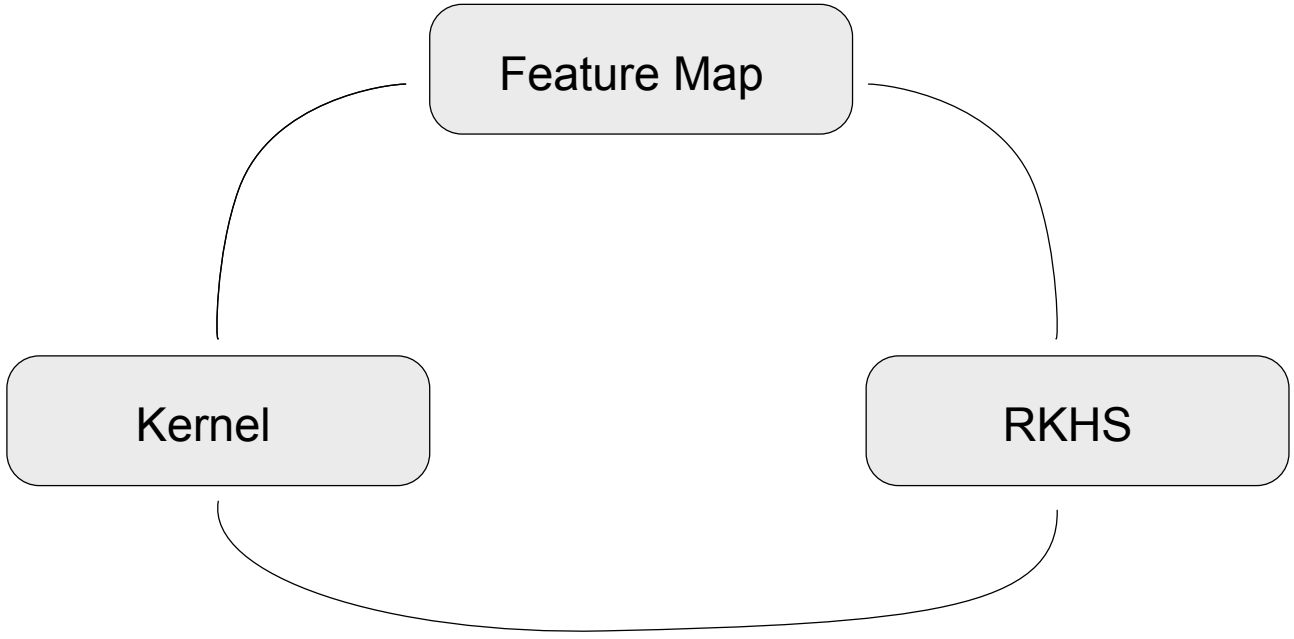


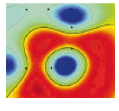


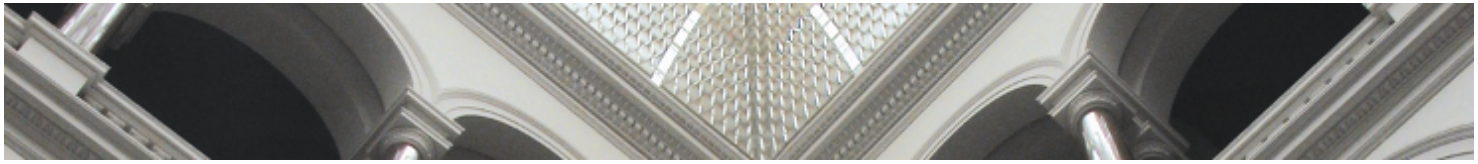
Technische Universität Berlin

Feature Maps, Kernels, and RKHS: The connection

There is a strong connection between feature maps, kernels, and the reproducing kernel Hilbert space (RKHS).







Technische Universität Berlin

Thank you!