ECE595 / STAT598: Machine Learning I

Lecture 3.1: Regression with Kernels - Kernel Method

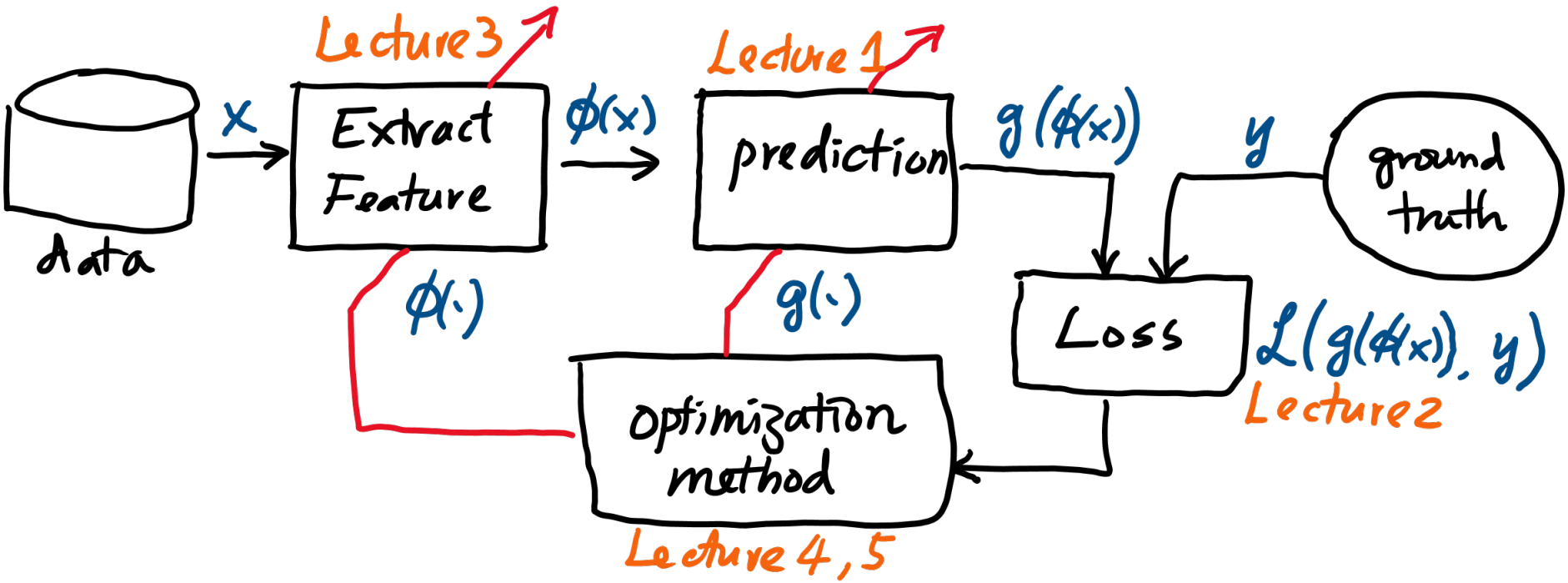
Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE
UNIVERSITY

Outline



Outline

Mathematical Background

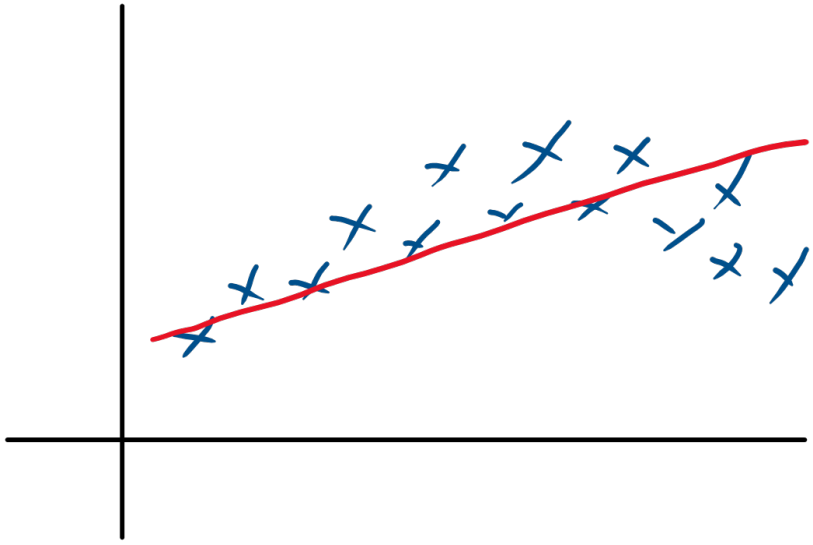
- Lecture 1: Linear regression: A basic data analytic tool
- Lecture 2: Regularization: Constraining the solution
- Lecture 3: Kernel Method: Enabling nonlinearity

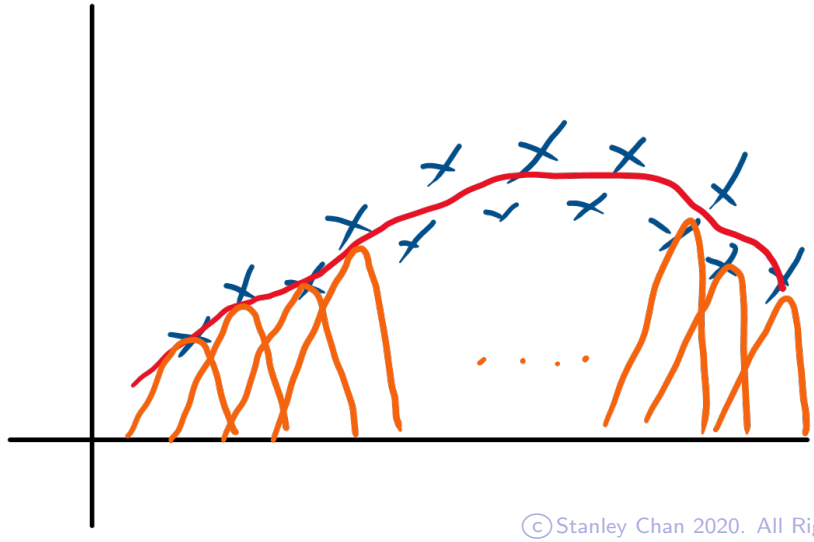
Lecture 3: Kernel Method

- Kernel Method
  - Dual Form
  - Kernel Trick
  - Algorithm
- Examples
  - Radial Basis Function (RBF)
  - Regression using RBF
  - Kernel Methods in Classification

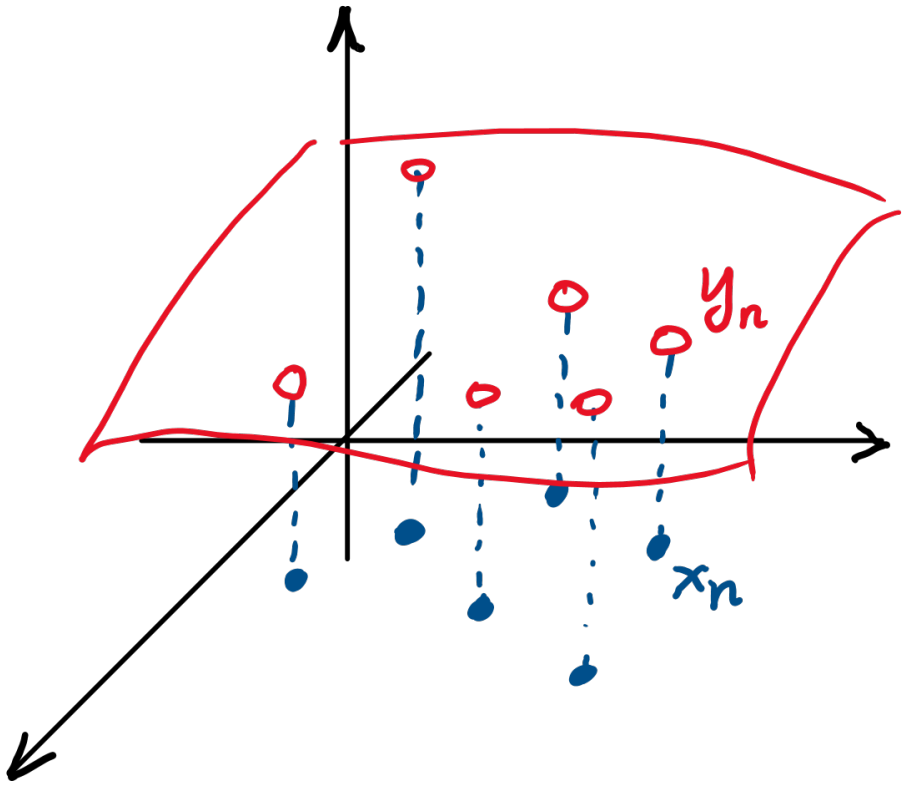
Why Another Method?

- Linear regression: Pick a **global** model, best fit globally.
- Kernel method: Pick a **local** model, best fit locally.
- In kernel method, instead of picking a line / a quadratic equation, we pick a **kernel**.
- A kernel is a measure of **distance** between **training samples**.
- Kernel method buys us the ability to handle nonlinearity.
- Ordinary regression is based on the **columns** (features) of $\boldsymbol{A}$.
- Kernel method is based on the **rows** (samples) of $\boldsymbol{A}$.





Pictorial Illustration



goal: learn the surface

prediction: When new sample comes, Interpolate on the surface

Overview of the Method

Model Parameter:

- We **want** the model parameter $\widehat{\boldsymbol{\theta}}$ to look like: (How? Question 1)

$$\widehat{\boldsymbol{\theta}} = \sum_{n=1}^{N} \alpha_n \boldsymbol{x}^n.$$

- This model expresses $\widehat{\boldsymbol{\theta}}$ as a combination of the **samples**.
- The trainable parameters are α_n, where $n = 1, \ldots, N$.
- If we can make α_n **local**, i.e., non-zero for only a few of them, then we can achieve our goal: localized, sample-dependent.

Predicted Value

- The predicted value of a new sample $\boldsymbol{x}$ is

$$\widehat{y} = \widehat{\boldsymbol{\theta}}^T \boldsymbol{x} = \sum_{n=1}^{N} \alpha_n \langle \boldsymbol{x}, \boldsymbol{x}^n \rangle.$$

- We **want** this model to encapsulate nonlinearity. (How? Question 2)

Dual Form of Linear Regression

Goal: Addresses Question 1: Express $\widehat{\boldsymbol{\theta}}$ as

$$\widehat{\boldsymbol{\theta}} = \sum_{n=1}^{N} \alpha_n \boldsymbol{x}^n.$$

We start by listing out a technical lemma:

Lemma

For any $\boldsymbol{A} \in \mathbb{R}^{N \times d}$, $\boldsymbol{y} \in \mathbb{R}^d$, *and* $\lambda > 0$,

$$(\boldsymbol{A}^T\boldsymbol{A} + \lambda \boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{y} = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}. \tag{1}$$

Proof: See Appendix.

Remark:

- The dimensions of $\boldsymbol{I}$ on the left is $d \times d$, on the right is $N \times N$.
- If $\lambda = 0$, then the above is true only when $\boldsymbol{A}$ is invertible.

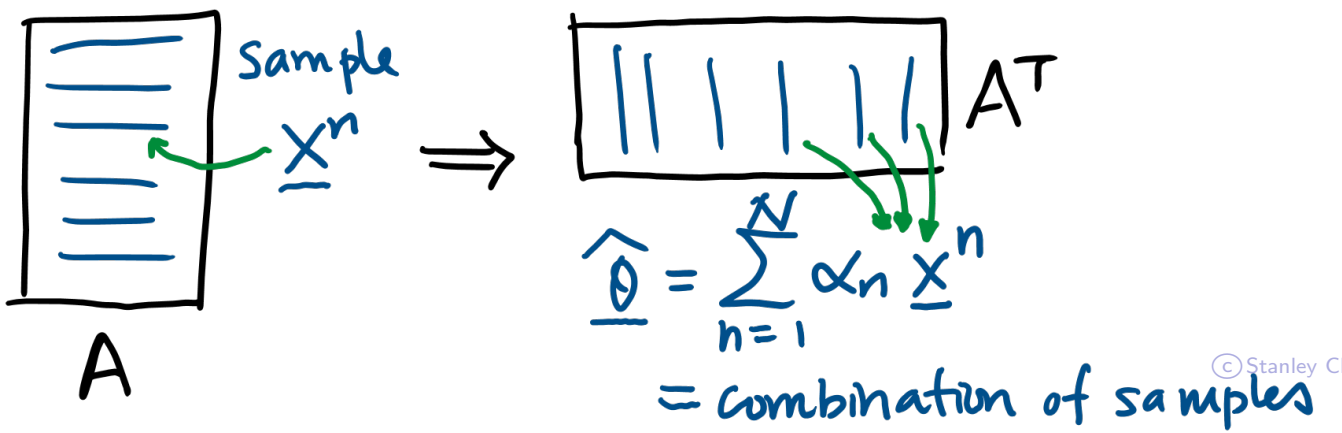
Dual Form of Linear Regression

- Using the Lemma, we can show that

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{A}^T\boldsymbol{A} + \lambda \boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{y} \qquad \text{(Primal Form)}$$

$$= \boldsymbol{A}^T \underbrace{(\boldsymbol{A}\boldsymbol{A}^T + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}}_{\stackrel{\text{def}}{=}\boldsymbol{\alpha}} \qquad \text{(Dual Form)}$$

$$= \begin{bmatrix} — & (\boldsymbol{x}^1)^T & — \\ — & (\boldsymbol{x}^2)^T & — \\ & \vdots & \\ — & (\boldsymbol{x}^N)^T & — \end{bmatrix}^T \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} = \sum_{n=1}^{N} \alpha_n \boldsymbol{x}^n, \quad \alpha_n \stackrel{\text{def}}{=} [(\boldsymbol{A}\boldsymbol{A}^T + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}]_n$$



The Kernel Trick

Goal: Addresses Question 2: Introduce nonlinearity to

$$\widehat{y} = \widehat{\boldsymbol{\theta}}^T \boldsymbol{x} = \sum_{n=1}^{N} \alpha_n \langle \boldsymbol{x}, \boldsymbol{x}^n \rangle.$$

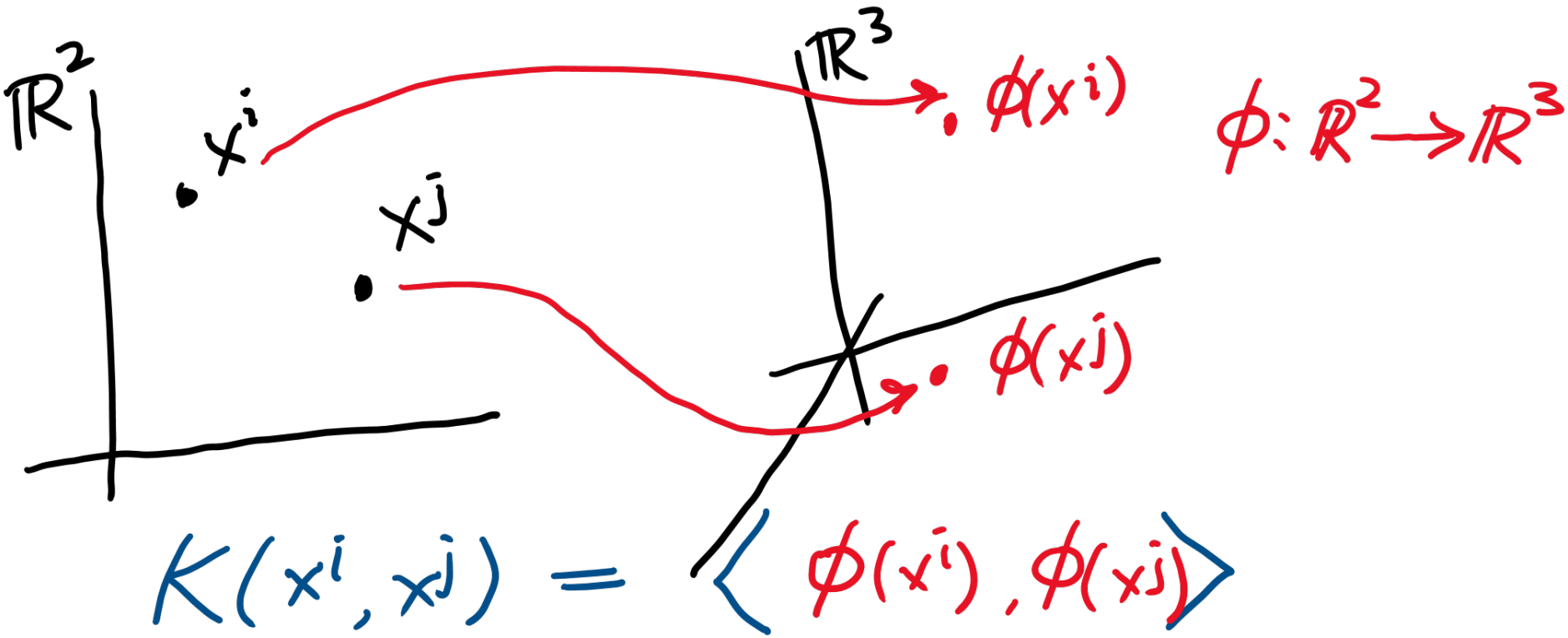
The Idea:

- Replace the inner product $\langle \boldsymbol{x}, \boldsymbol{x}^n \rangle$ by $k(\boldsymbol{x}, \boldsymbol{x}^n)$:

$$\widehat{y} = \widehat{\boldsymbol{\theta}}^T \boldsymbol{x} = \sum_{n=1}^{N} \alpha_n k(\boldsymbol{x}, \boldsymbol{x}^n).$$

- $k(\cdot, \cdot)$ is called a **kernel**.
- A kernel is a measure of the **distance** between two samples $\boldsymbol{x}^i$ and $\boldsymbol{x}^j$.
- $\langle \boldsymbol{x}^i, \boldsymbol{x}^j \rangle$ measure distance in the ambient space, $k(\boldsymbol{x}^i, \boldsymbol{x}^j)$ measure distance in a **transformed** space.
- In particular, a valid kernel takes the form $k(\boldsymbol{x}^i, \boldsymbol{x}^j) = \langle \phi(\boldsymbol{x}^i), \phi(\boldsymbol{x}^j) \rangle$ for some nonlinear transforms ϕ.

Kernels Illustrated



- A kernel typically lifts the ambient dimension to a **higher** one.
- For example, mapping from $\mathbb{R}^2$ to $\mathbb{R}^3$

$$\boldsymbol{x}^n = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \text{and} \qquad \phi(\boldsymbol{x}_n) = \begin{bmatrix} x_1^2 \\ x_1 x_2 \\ x_2^2 \end{bmatrix}$$

Relationship between Kernel and Transform

Consider the following kernel $k(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u}^T \boldsymbol{v})^2$. What is the transform?

- Suppose $\boldsymbol{u}$ and $\boldsymbol{v}$ are in $\mathbb{R}^2$. Then $(\boldsymbol{u}^T \boldsymbol{v})^2$ is

$$(\boldsymbol{u}^T \boldsymbol{v})^2 = \left(\sum_{i=1}^{2} u_i v_i \right) \left(\sum_{j=1}^{2} u_j v_j \right)$$

$$= \sum_{i=1}^{2} \sum_{j=1}^{2} (u_i u_j)(v_i v_j) = \begin{bmatrix} u_1^2 & u_1 u_2 & u_2 u_1 & u_2^2 \end{bmatrix} \begin{bmatrix} v_1^2 \\ v_1 v_2 \\ v_2 v_1 \\ v_2^2 \end{bmatrix}.$$

- So if we define ϕ as

$$\boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \quad \mapsto \quad \phi(\boldsymbol{u}) = \begin{bmatrix} u_1^2 \\ u_1 u_2 \\ u_2 u_1 \\ u_2^2 \end{bmatrix}$$

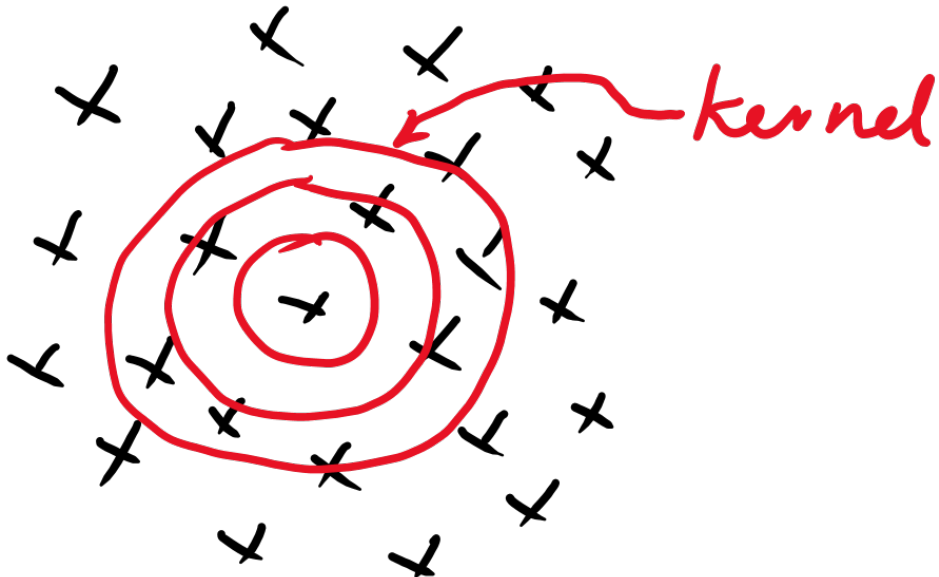
then $(\boldsymbol{u}^T \boldsymbol{v})^2 = \langle \phi(\boldsymbol{u}), \phi(\boldsymbol{v}) \rangle$.

Radial Basis Function

A useful kernel is the **radial basis kernel** (RBF):

$$k(\boldsymbol{u}, \boldsymbol{v}) = \exp\left\{-\frac{\|\boldsymbol{u}-\boldsymbol{v}\|^2}{2\sigma^2}\right\}.$$

- The corresponding nonlinear transform of RBF is **infinite dimensional**. See Appendix.
- $\|\boldsymbol{u}-\boldsymbol{v}\|^2$ measures the distance between two data points $\boldsymbol{u}$ and $\boldsymbol{v}$.
- σ is the std dev, defining "far" and "close".
- RBF enforces **local** structure; Only a few samples are used.



Kernel Method

Given the choice of the kernel function, we can write down the algorithm as follows.

1. Pick a kernel function $k(\cdot, \cdot)$.
2. Construct a kernel matrix $\boldsymbol{K} \in \mathbb{R}^{N \times N}$, where $[\boldsymbol{K}]_{ij} = k(\boldsymbol{x}^i, \boldsymbol{x}^j)$, for $i = 1, \ldots, N$ and $j = 1, \ldots, N$.
3. Compute the coefficients $\boldsymbol{\alpha} \in \mathbb{R}^N$, with

$$\alpha_n = [(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{y}]_n.$$

4. Estimate the predicted value for a new sample $\boldsymbol{x}$:

$$g_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{n=1}^{N} \alpha_n k(\boldsymbol{x}, \boldsymbol{x}^n).$$

Therefore, the choice of the regression function is shifted to the choice of the kernel.