$$\boxed{\underline{Q}: \; k(x,x) - 2k(x,y) + k(y,y) > 0 \; ?}$$

$$K = \left(k(x_i, x_j) \right)_{i,j} \succeq 0$$

$$K = \begin{pmatrix} k(x,x) & k(x,y) \\ k(y,x) & k(y,y) \end{pmatrix}$$

c = coef. vector $\quad c^T K c \geq 0$

$$c = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \quad (c_1 \;\; c_2)\, K \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} =$$

$$= c_1^2 k(x,x) + 2c_1 c_2 k(x,y) + c_2^2 k(y,y)$$

$$\geq 0$$

$c_1 = 1,\ c_2 = -1$: $\boxed{k(x,x) - 2k(x,y) + k(y,y) \geq 0}$

$$\begin{cases} K = \left(k(x_i, x_j)\right)_{i,j \in [n]} & (\alpha \in \mathbb{R}^n) \\ f(x) = \sum_{i=1}^n \alpha_i k(x, x_i) \end{cases}$$

- $f(x_j) = \sum_{i=1}^n \alpha_i \underbrace{k(x_j, x_i)}_{j\text{-th line of } K}$

$$\boxed{f(x_j) = (K\alpha)_j}$$

$$(f(x_1), \dots, f(x_n))^T = K\alpha$$

- $\|f\|_{\mathcal{H}}^2 = \left\| \sum_{i=1}^n \alpha_i k(\cdot, x_i) \right\|_{\mathcal{H}}^2$

$$= \left\langle \sum_{i=1}^n \alpha_i k(\cdot, x_i), \sum_{i=1}^n \alpha_i k(\cdot, x_i) \right\rangle_{\mathcal{H}}$$

$$= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \langle k(\cdot, x_i), k(\cdot, x_j) \rangle_{\mathcal{H}}$$

reproducing prop.

$$= \sum \sum \alpha_i \alpha_j k(x_i, x_j)$$

$$\boxed{\|f\|_{\mathcal{H}}^2 = \alpha^T K \alpha}$$

$g(x) = \langle g, k_x \rangle$

$k(x, y) = k_y(x)$

$= \langle k_y, k_x \rangle$

Kernel ridge regression

(Christianini & Shaw-Taylor '00)

Let k be a p.s.d. kernel / take k = Gaussian kernel bw 3

$$(*)\quad \hat{f} \in \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}$$

Representer theorem applies: $\hat{f} = \sum_{i=1}^{n} \alpha_i k(x, x_i)$

$\longrightarrow$ only have to solve optim problem in dim n.

Use the facts we just proved:

$$(\hat{f}(x_1), \dots, \hat{f}(x_n))^T = K\alpha \;;\quad \|\hat{f}\|_{\mathcal{H}}^2 = \alpha^T K \alpha$$

$$\hat{\alpha} \in \arg\min_{\alpha \in \mathbb{R}^n} \underbrace{\left\{ \frac{1}{n} \|y - K\alpha\|^2 + \lambda \alpha^T K \alpha \right\}}_{\text{cvx, smooth as a f}^{\text{n}}\text{ of } \alpha} \quad (**)$$

Let us find a critical point.

$$\nabla_\alpha = \frac{1}{n}(-2)K(y - K\alpha) + n\frac{\lambda}{n}(2K)\alpha$$

$$= \frac{2}{n} K \left[(K + n\lambda I_n)\alpha - y \right]$$

$$\stackrel{?}{=} 0$$

$$\nabla = \frac{2}{n} K \left[\underline{(K + n\lambda I_n)\alpha - y} \right] = 0$$

$\lambda > 0$

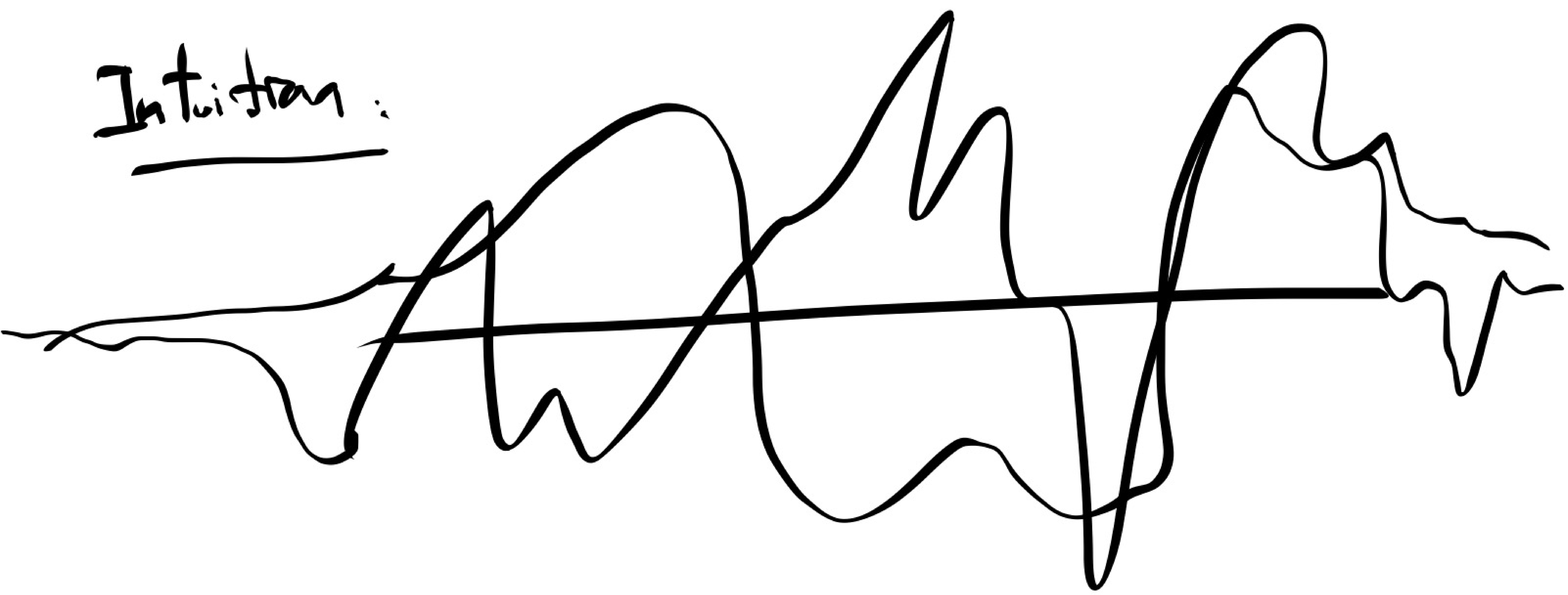
$\llcorner$ set to 0 is a possible solution

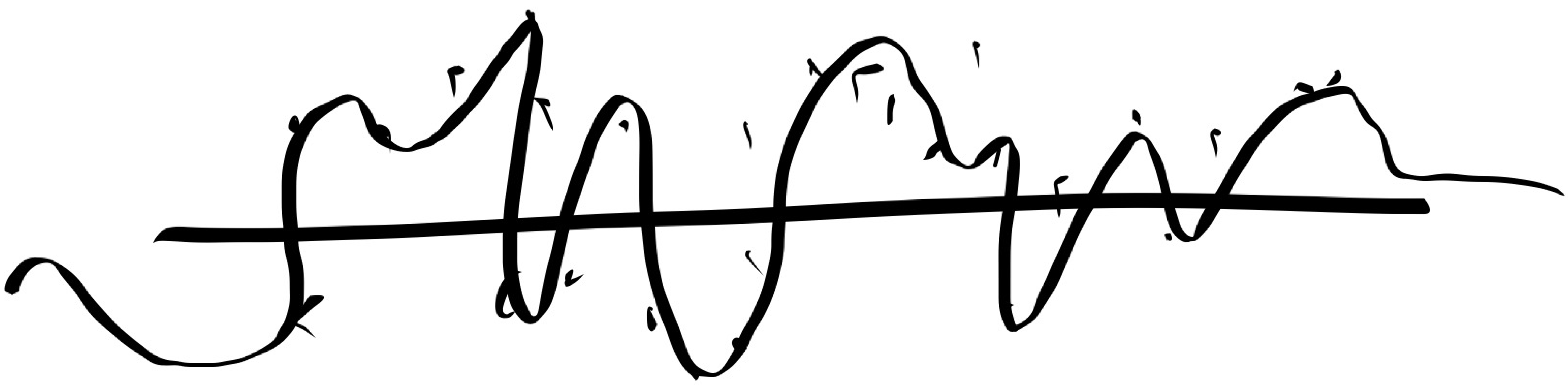
Again, $K + n\lambda I_n$ is invertible

$$\Rightarrow \boxed{\hat{\alpha} = (K + n\lambda I_n)^{-1} y}$$

$$\Rightarrow \boxed{\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_i k(x, x_i)}$$

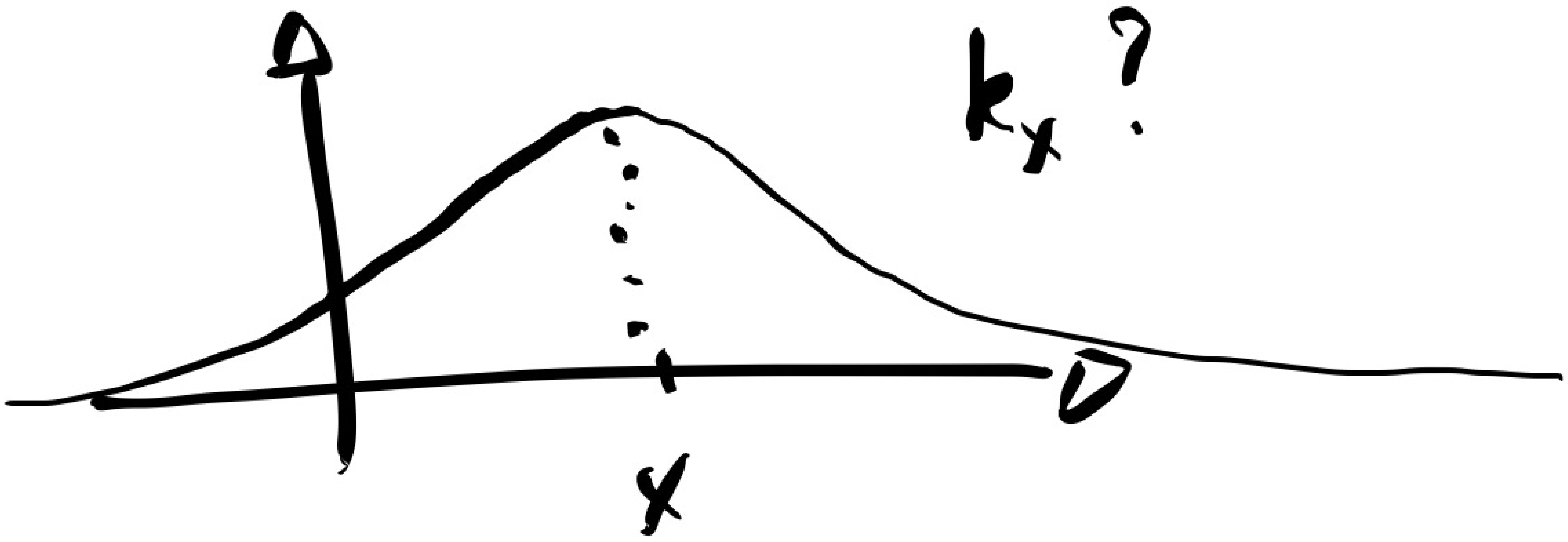
Intuition:





Gaussian kernel:

$$k(x,y) = \exp\left(-\frac{(x-y)^2}{2}\right)$$

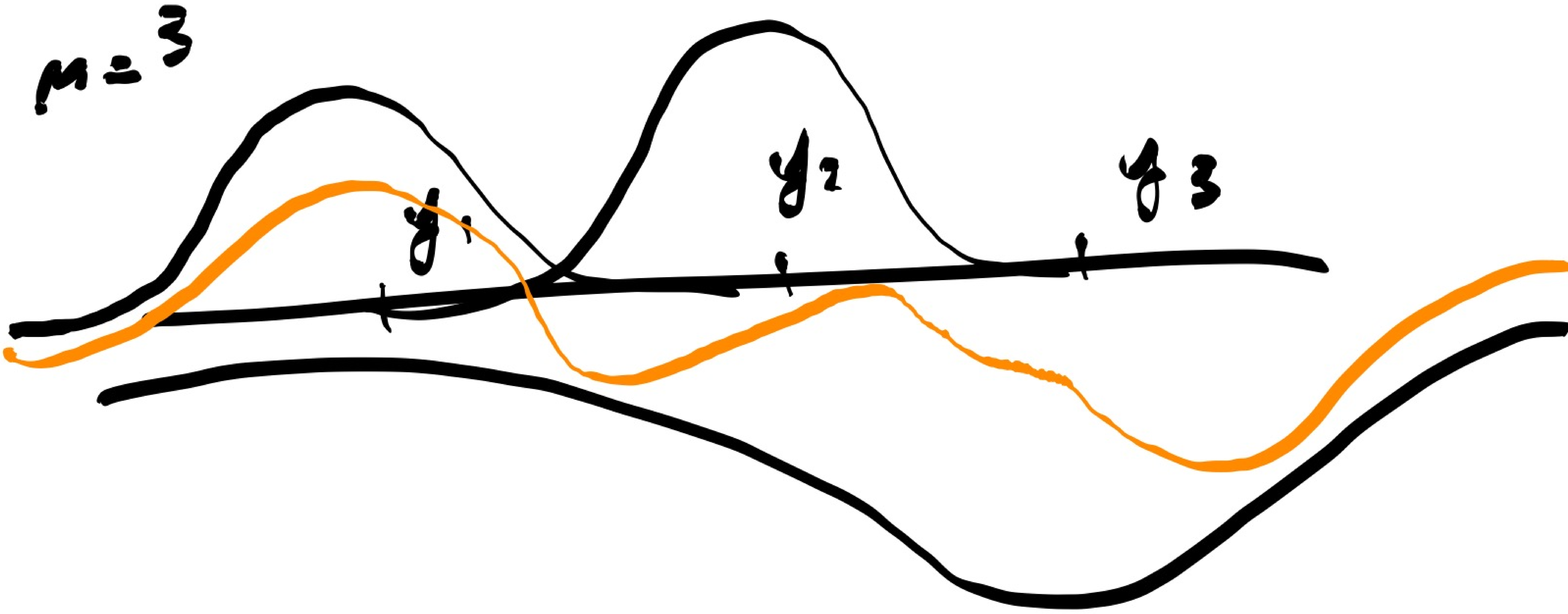


k_x ?

RKHS associated to k contains

$$\mathcal{H}_0 = \left\{ g(x) = \sum_{i=1}^{m} \alpha_i k_{y_i}(x), \text{ with } m > 0, \; y_1 \dots y_m \in \mathbb{R} \right\}$$

$m = 3$



regular ridge regression = kernel ridge with linear kernel

$$K = \begin{pmatrix} x_1^T x_1 & x_1^T x_2 & \cdots \\ x_2^T x_1 & \ddots & \\ \vdots & & \ddots \end{pmatrix}$$

columns

$$\left(x_1 \,\middle|\, x_2 \,\middle|\, \cdots \,\middle|\, x_n \right)$$

$$= \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}$$

$$\boxed{K = XX^T}$$

(Woodbury identity)

$$\hat{\alpha} = (K + n\lambda I_n)^{-1} y$$

$$= (XX^T + n\lambda I_n)^{-1} y$$

$$\text{pred: } K\hat{\alpha} = XX^T (XX^T + n\lambda I_n)^{-1} y$$

$$X\hat{\beta} = X(X^T X + n\lambda I)^{-1} X^T y$$

they are the same!

<u>Uniqueness</u>

remember, we had to solve

$$K\left[(K + n\lambda I_n)\alpha - y\right] = 0$$

if K is singular, other solution.

ε s.t. $K\varepsilon = 0$

$\hat{\alpha} = (K + n\lambda I_n)^{-1} y$

$\Rightarrow$ $\boxed{\hat{\alpha} + \varepsilon \text{ is also a solution}}$.

$$K\left[(K + n\lambda I)\varepsilon\right] = \cancel{K}\left(\cancel{K\varepsilon} + n\lambda\varepsilon\right)$$

$$= \cancel{K}\varepsilon = 0$$

Not a problem since $\hat{\alpha}$ and $\hat{\alpha} + \varepsilon$ correspond to the same element in $\mathcal{H}$.

$$\left\|\sum_i \hat{\alpha}_i k_{x_i} - \sum_i (\hat{\alpha}_i + \varepsilon_i) k_{x_i}\right\|_{\mathcal{H}}^2 =$$

$$= \left\|\sum_i (-\varepsilon_i) k_{x_i}\right\|_{\mathcal{H}}^2 = \varepsilon^T \underbrace{K\varepsilon}_{=0} = 0.$$