φ

Bonus: Hilbert spaces

$(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$

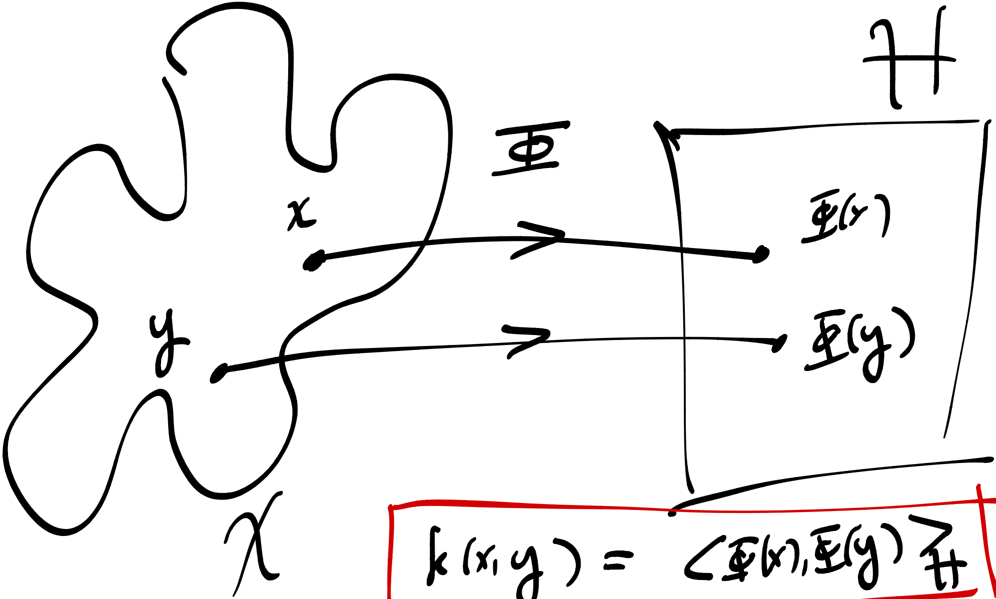
Definition: A *Hilbert space* is a real or complex $\boxed{\text{vector space}}$ which is also a complete metric space with respect to the distance function induced by the inner product.

- **Remark:** recall the linear kernel, all we used were properties of inner product
- let $\Phi : \mathcal{X} \to \mathcal{H}$ be some mapping, $\mathcal{H}$ a Hilbert space with scalar product $\langle \cdot, \cdot \rangle$
- then $\boxed{k(x, y) = \langle \Phi(x), \Phi(y) \rangle}$ is positive definite:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \langle \Phi(x), \Phi(y) \rangle = \left\| \sum_{i=1}^{n} c_i \Phi(x_i) \right\|^2 \geq 0\,,$$

by linearity of the inner product.

$\to$ large class of kernels: "nice" vector space $\mathcal{H}$ + embedding of $\mathcal{X}$ in $\mathcal{H}$



$$\boxed{k(x,y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}}$$

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, sym., Gram matrix K is psd.

Kernel as inner products

all $k(x_i, x_j)$ $(1 \leq i,j \leq n)$.

- **Remarkable fact:** the converse statement is true!

> **Theorem:**[9] For any kernel k on $\mathcal{X}$, there exists a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ and a mapping $\Phi : \mathcal{X} \to \mathcal{H}$ such that
>
> $$\forall x, y \in \mathcal{X}, \qquad \boxed{k(x, y) = \langle \Phi(x), \Phi(y) \rangle} \, .$$

$(\Phi = \varphi)$

- **Reminder:** Hilbert space = inner product + *complete* for the associated norm (Cauchy sequences converge in $\mathcal{H}$)
- **Consequence:** we can think of any kernel as a dot product in the *feature space*
- **Main idea:** forget about Φ and work only with kernel evaluations (more on that later)

[9] Aronszajn, *Theory of reproducing kernels*, Transactions of the American Mathematical Society, 1950

$h(x, y) = k(x_i, x_j)$

Proof in the finite case

- assume that $\boxed{\mathcal{X} = \{x_1, \dots, x_N\}}$ is finite of size N
- any kernel k is entirely defined by the $N \times N$ positive semi-definite matrix $K := (k(x_i, x_j))_{i,j=1}^{N}$

spectral theorem

- we can diagonalize K in an orthonormal basis $(u_1, \dots, u_N)$ with associated (non-negative) eigenvalues $\lambda_1, \dots, \lambda_N$: $K = U \Lambda U^\top$, with $U_{:,i} = u_i$, $\Lambda = \operatorname{diag}(\lambda)$, $UU^\top = U^\top U = \mathrm{I}$
- then we write

$$
\begin{aligned}
k(x_i, x_j) &= \left(\sum_{\ell=1}^{N} \lambda_\ell u_\ell u_\ell^\top \right)_{i,j} \\
&= \sum_{\ell=1}^{N} \lambda_\ell (u_\ell)_i (u_\ell)_j = \langle \Phi(x_i), \Phi(x_j) \rangle ,
\end{aligned}
$$

with

$$
\Phi(x_i) := \left(\sqrt{\lambda_1} (u_1)_i, \cdots, \sqrt{\lambda_n} (u_N)_i \right)^\top .
$$

□

Reminder: $\boxed{K = U \Lambda U^T}$ $\in \mathbb{R}^{N \times N}$

real ✓

with $U \in \mathbb{R}^{N \times N}$ $UU^T = U^T U = I_N$ (identity) since K sym.

additionally, $K \succeq 0$

$\Downarrow$

$\forall i \in [N], \lambda_i \geq 0$

$$\Lambda = \mathrm{Diag}(\lambda_1, \ldots, \lambda_N) = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{pmatrix}$$

$$\Rightarrow \quad K = \sum_{l=1}^{N} \underbrace{\lambda_l}_{\text{eigenvalues of } K} u_l u_l^T \quad \text{with } u_l = l\text{th column of } U.$$

$$U\Lambda = \begin{pmatrix} u_1 & \Big| & u_2 & \Big| & \cdots & \Big| & u_N \end{pmatrix} \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_N \end{pmatrix}$$

$$= \begin{pmatrix} \lambda_1 u_1 & \Big| & \lambda_2 u_2 & \Big| & \cdots & \Big| & \lambda_N u_N \end{pmatrix}$$

$$K = U \Lambda U^T = \left(\lambda_1 u_1 \middle| \cdots \middle| \lambda_N u_N \right) \begin{pmatrix} u_1^T \\ \vdots \\ u_N^T \end{pmatrix} = \sum_{l=1}^{N} \lambda_l u_l u_l^T$$

posi.

$$\mathbb{R}^{N \times N} \ni u_i u_j^T \;?$$

$$U U^T = I_N$$

$$\left(u_1 \middle| \cdots \middle| u_N \right) \begin{pmatrix} u_1^T \\ \vdots \\ u_N^T \end{pmatrix} \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$$

$$u_i u_j^T = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

$$k(x_i, x_j) = K_{i,j} = \left(\sum_{l=1}^{N} \lambda_l u_l u_l^T\right)_{i,j}$$

$$= \sum_{l=1}^{N} \lambda_l \left(u_l u_l^T\right)_{i,j}$$

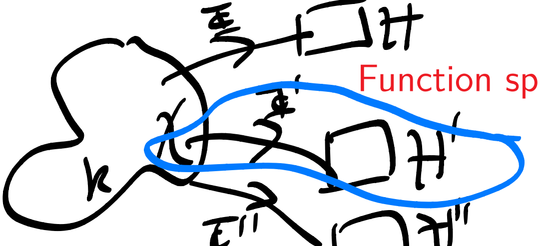
$$\begin{array}{cc} & j \\ i & \begin{pmatrix} u_{l,1} \\ \vdots \\ u_{l,N} \end{pmatrix} \begin{pmatrix} u_{l,1} & \cdots & u_{l,N} \end{pmatrix} \end{array} \quad u_{l,i} \times u_{l,j}$$

$$= \sum_{l=1}^{N} \lambda_l u_{l,i} u_{l,j}$$

$$k(x_i, x_j) = \sum_{l=1}^{N} \sqrt{\lambda_l} u_{l,i} \times \sqrt{\lambda_l} u_{l,j} = \left\langle \begin{pmatrix} \sqrt{\lambda_1} u_{1,i} \\ \vdots \\ \sqrt{\lambda_N} u_{N,i} \end{pmatrix}, \begin{pmatrix} \sqrt{\lambda_1} u_{1,j} \\ \vdots \\ \sqrt{\lambda_N} u_{N,j} \end{pmatrix} \right\rangle \quad \square$$

5.2. Reproducing kernel Hilbert spaces

(RKHS)



Function spaces

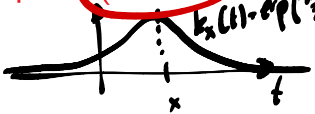
- among all spaces in the previous statement, one of them has ~~interesting~~ very nice properties
- in particular, it is a **space of functions**
- *i.e.*, we can map each point $x \in \mathcal{X}$ to a *function* $\Phi(x) = k_x \in \mathcal{H}$
- **Example:** $\mathcal{X} = \mathbb{R}$, we map each x to the function $t \mapsto xt$
- $\to$ space of linear functions
- more complicated in general...

$\cancel{\Phi(x)(y)}$

Reproducing Kernel Hilbert Space (RKHS)

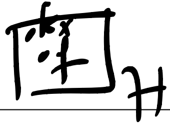
Ex: $k(x,y) = \exp(-(x-y)^2)$

$k_x(t) = \exp\left(-\frac{(x-t)^2}{2}\right)$



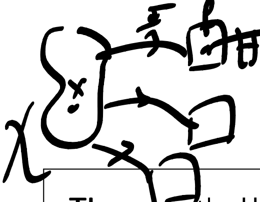
Definition: let $\mathcal{X}$ be a set and $\mathcal{H}$ be a function space forming a Hilbert space with inner product $\langle \cdot, \cdot \rangle$. The function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *reproducing kernel* of $\mathcal{H}$ if

- $\mathcal{H}$ contains all functions of the form $k_x : t \mapsto k(x,t)$
- for every $x \in \mathcal{X}$ and $f \in \mathcal{H}$, the *reproducing property* holds:



$$f(x) = \langle f, k_x \rangle .$$

- if a reproducing kernel exists, then $\mathcal{H}$ is called a *reproducing kernel Hilbert space* (RKHS)



$f(x) \in \mathbb{R}$

Equivalent definition

Theorem: the Hilbert space $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ is a RKHS if, and only if, for any $x \in \mathcal{X}$, the mapping $f \mapsto f(x)$ is continuous.

BONUS

- *Proof of* $\Rightarrow$*:* let k be a reproducing kernel, $x \in \mathcal{X}$ and $f_n \to f$ in $\mathcal{H}$
- we write

reproducing prop: $f_n(x) = \langle f_n, k_x \rangle$

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

$$
\begin{aligned}
|f_n(x) - f(x)| &= |\langle f_n - f, k_x \rangle| \\
&\leq \underbrace{\|f_n - f\|_{\mathcal{H}}}_{\to 0} \cdot \|k_x\|_{\mathcal{H}}
\end{aligned}
$$

$\to$ fixed, finite

by Cauchy-Schwarz inequality.

- $\|f_n - f\| \to 0$ and we can conclude
- **Remark:** $\|k_x\|^2 = \langle k_x, k_x \rangle = k(x, x)$, thus $|f(x)| \leq \|f\|_{\mathcal{H}} \cdot \cancel{k(x,x)^{1/2}}$

$|f(x)| \leq \|f\|_{\mathcal{H}} \cdot \|k_x\|_{\mathcal{H}}$

$\exp(-(x-x)^2) = 1$ $\quad = 1 \quad$ 1

Continuity ctd.

Bonus.

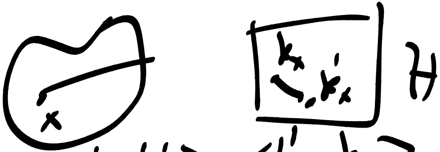
- *Proof of* $\Leftarrow$*:* let $x \in \mathcal{X}$
- by the reproducing property, $L : x \mapsto f(x)$ is a *linear functional*
- Riesz theorem there exists ℓ_x such that $L(x) = \langle f, \ell_x \rangle$
- define $k(x, y) := \ell_y(x)$
- one can check readily the RKHS properties. □

Riesz-Fréchet

Uniqueness

Theorem: if $\mathcal{H}$ is a RKHS, then it has a unique reproducing kernel. Conversely, a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ can be the reproducing kernel of at most one RKHS.

- we talk about *the* RKHS associated to k
- *Proof:* let k and k' be two reproducing kernels
- then for all $x \in \mathcal{X}$,

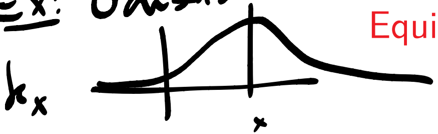


$$\begin{aligned}\|k_x - k'_x\|^2 &= \langle k_x - k'_x, k_x - k'_x \rangle = \langle k_x, k_x \rangle - \langle k'_x, k_x \rangle - \langle k_x, k'_x \rangle + \langle k'_x, k'_x \rangle \\ &= \cancel{k_x(x)} - \cancel{k'_x(x)} - \cancel{k_x(x)} + \cancel{k'_x(x)} \\ &= 0 .\end{aligned}$$

rep. prop

$\square$

Ex: Gaussian kernel.



Equivalence psd / RKHS

Theorem: a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if, and only if, it is a reproducing kernel. $((\mathcal{H}, \langle \cdot, \cdot \rangle))$

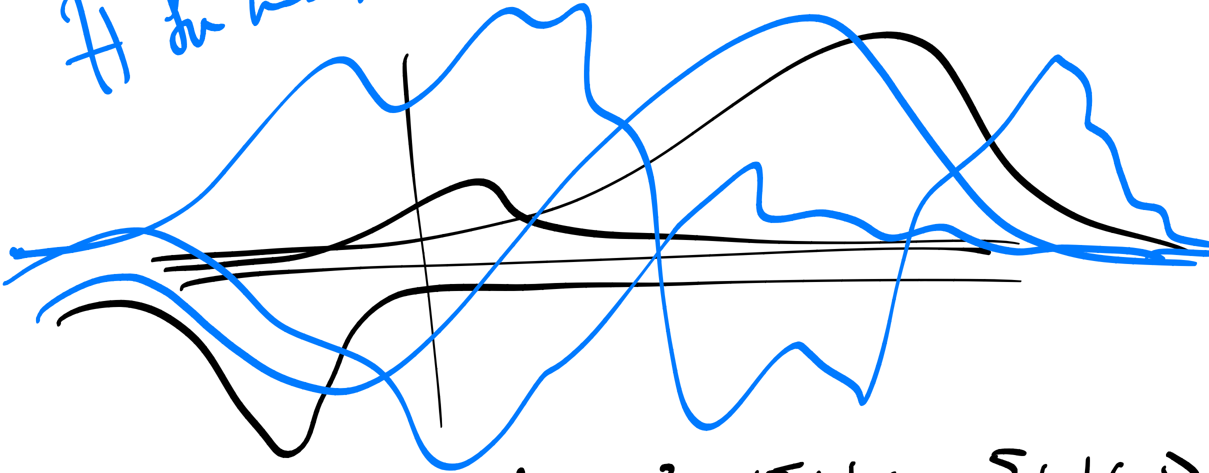
- **Idea:** build $\mathcal{H}$ as the ~~completion~~ of

$$\mathcal{H}_0 := \left\{ \sum_{i=1}^{n} \alpha_i \underbrace{k(\cdot, x_i)}_{k_{x_i}}, n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \mathcal{X} \right\}$$

- **Remark:** showing that a kernel is positive definite is enough to get Φ and $\mathcal{H}$ with the reproducing property "for free"

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ $\mathcal{H}$

$\mathcal{H}$ for functions.



If $\mathcal{H}_0 = \{ \sum \alpha_i k(\cdot, x_i) \}$

$$\|f\|_{\mathcal{H}}^2 = \langle \sum \alpha_i k(\cdot, x_i), \sum \alpha_j k(\cdot, x_j) \rangle$$

$$= \sum \sum \alpha_i \alpha_j \langle k(\cdot, x_i), \rightarrow$$

$$\forall f \in \mathcal{H}_0, \quad \|f\|^2_{\mathcal{H}} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \langle k_{x_i}, k_{x_j} \rangle$$

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

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K_{i,j} \longleftarrow \text{Gram matrix}$$

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

Mahalanobis.

Example

- **Example:** polynomial kernel of degree 2:

$$k(x,y) = (x^\top y)^2 \,.$$

- proved during the exercise session:

$$k(x,y) = \langle xx^\top, yy^\top \rangle_F \,,$$

thus k is positive definite

- **Question:** what is the RKHS?
- we know that $\mathcal{H}$ contains all the functions

$$f(x) = \sum_i a_i k(x_i, x) = \sum_i a_i \langle x_i x_i^\top, xx^\top \rangle = \langle \sum_i a_i x_i x_i^\top, xx^\top \rangle$$

Example, ctd.

- spectral theorem: any symmetric matrix can be decomposed as $\sum_i a_i x_i x_i^\top$
- candidate RKHS: set a quadratic functions

$$f_S(x) = \langle S, xx^\top \rangle = x^\top S x \,,$$

  with S symmetric matrix of size $d \times d$
- inner product on $\mathcal{H}$:

$$\langle f_S, f_{S'} \rangle = \langle S, S' \rangle_F \,.$$

- we can check that $\mathcal{H}$ is a Hilbert space (isomorphic to $\mathcal{S}^{d \times d}$)
- finally, we check the reproducing property

5.3. More examples

Elementary properties

Proposition: Let $k_i : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a (potentially infinite) family of p.d. kernels. Then

- for any $\lambda_1, \ldots, \lambda_p \geq 0$, the sum $\sum_{i=1}^{p} \lambda_i k_i$ is positive definite (can add)
- for any $a_1, \ldots, a_p \in \mathbb{N}$, the product $k_1^{a_1} \cdots k_p^{a_p}$ is positive definite (can multiply)
- if it exists, the limit $k = \lim_{p \to +\infty} k_p$ is positive definite

Moreover, let $\mathcal{X}_i$ be a sequence of sets and k_i positive kernels on each $\mathcal{X}_i$. Then

$$k((x_1, \ldots, x_p), (y_1, \ldots, y_p)) := \prod_{i=1}^{p} k_i(x_i, y_i)$$

$k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$

$x = (x_1, x_2), y = (y_1, y_2)$

and

$$k((x_1, \ldots, x_p), (y_1, \ldots, y_p)) := \sum_{i=1}^{p} k_i(x_i, y_i)$$

$\tilde{k}(x, y) = k(x_1, y_1)$ $(+ \text{ or}) \times k(x_2, y_2)$

are positive definite kernels.

$\left(-\|x-y\|^2 \right.$ NOT a kernel :C $\left. \right)$

Taking the exponential

$p! = 1 \cdot 2 \cdot \ldots \cdot p$
≥ 0

Theorem: if k is a positive definite kernel, then e^k as well.

- *Proof:* we write

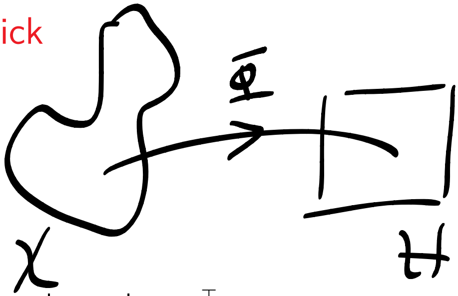
$$\mathrm{e}^{k(x,y)} = \lim_{n \to +\infty} \sum_{p=0}^{n} \frac{k(x,y)^p}{p!} ,$$

  then reason step by step.
- by the product property, $k(x,y)^p$ is a kernel for any $p \geq 0$
- as a positive linear combination of kernels, $\sum_{p=0}^n \frac{k(x,y)^p}{p!}$ is a kernel for all $n \geq 1$
- finally, e^k is a kernel as a limit of kernels. □

5.4. The kernel trick and applications

generally, $\dim \mathcal{H} = +\infty$

The kernel trick



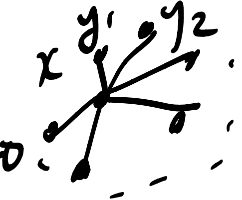
- input data $x_1, \ldots, x_n \in \mathcal{X}$
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel with associated RKHS $\mathcal{H}$
- we call $\Phi : \mathcal{X} \to \mathcal{H}$ the feature map
- **Idea:** imagine that our algorithm only depends on scalar products $x_i^\top x_j$
- then we can map the x_i to $\mathcal{H}$ and replace the inner products by kernel evaluations, since

$$\langle \Phi(x_i), \Phi(x_j) \rangle = k(x_i, x_j)\,.$$

- simple "trick" with many, many applications

$$x_i^T x_j \longleftrightarrow \langle \Phi(x_i), \Phi(x_j) \rangle$$

Example



- **Example:** computing distances → 1-NN
- suppose that our algo relies on distance computation
- that is, $\|x-y\|^2$ = $\langle x\cdot y, x\cdot y\rangle$
- we can write

1-NN in the feature space

$$\begin{aligned}\|\Phi(x)-\Phi(y)\|^2 &= \langle\Phi(x)-\Phi(y),\Phi(x)-\Phi(y)\rangle\\ &= \langle\Phi(x),\Phi(x)\rangle - 2\langle\Phi(x),\Phi(y)\rangle + \langle\Phi(y),\Phi(y)\rangle\end{aligned}$$

$$\|\Phi(x)-\Phi(y)\|^2 = k(x,x) - 2k(x,y) + k(y,y)\,.$$

$$d(x,y_i)^2 = k(x,x) - 2k(x,y_i) + k(y_i,y_i)$$

- in other words,

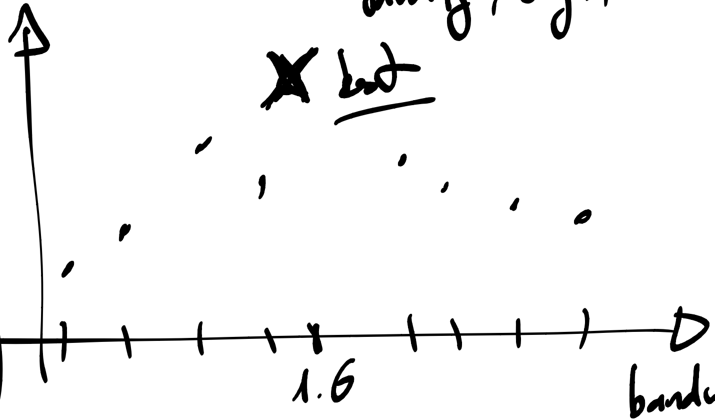
dist. in feature space ←
$$d_{\mathcal{H}}(x,y) = \sqrt{k(x,x)-2k(x,y)+k(y,y)}\,.$$

- as promised, **we do not need to know Φ!**

Q: How to choose the kernel?

hyperparameters of kernel family, e.g., Gaussian

acc. on val set for 1-NN with k_σ



$$\left(\exp\left(\frac{-\|x-y\|^2}{2\sigma^2}\right)\right)$$

5.5. The representer theorem

Motivation

$\|f\|^2 = \langle f, f \rangle$

- let us imagine that we take $\mathcal{H}$ as hypothesis class
- starting from regularized ERM, our optimization problem will look like

$$\arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2 \right\} . \qquad (\star)$$

- we penalize by the norm because it is an indicator of the *smoothness* of f
- **Why?** Cauchy-Schwarz + exercise:

$$|f(x) - f(y)| = |\langle f, k_x - k_y \rangle| \leq \|f\| \cdot \|k_x - k_y\| = \|f\| \cdot d_{\mathcal{H}}(x, y) .$$

- Eq. $(\star)$ is a complicate problem, potentially *infinite-dimensional*
- **Question:** how to solve it in practice?