# MAP569 Machine Learning II

## PC3 : Kernels and RKHS

## 1 Kernel of a Sobolev space

We consider the Sobolev space

$$\mathcal{H} = \{f : [0,1] \to \mathbb{R}, \text{ continuous, differentiable a.e., } f' \in L^2([0,1]),\ f(0) = 0\}\ ,$$

endowed with the Hilbert norm

$$|f|_{\mathcal{H}} = \sqrt{\int_0^1 (f')^2}.$$

1. Prove that if $\mathcal{H}$ is a RKHS with kernel $k$, then for all $x \in [0,1]$ we have

$$f(x) = \int_0^1 f'(y)\, \frac{\partial}{\partial y} k(x,y)\, \mathrm{d}y \quad \text{and} \quad f(x) = \int_0^1 f'(y) \mathbf{1}_{\{y \leq x\}}\, \mathrm{d}y\ .$$

2. What is the reproducing kernel $k$ associated with $\mathcal{H}$ ?
3. What is the shape of $\phi(x) = k(x, .)$ ?

## 2 Kernels for proteins or genetic sequences

Proteins or genetic sequences can be represented by words of varying length based on a finite alphabet $\mathcal{A}$. We want to apply some supervised learning algorithms to such objects. Classical algorithms like Ridge regression or SVM take as input vectors in $\mathbb{R}^p$, not words of varying length. The recipe for applying some supervised learning algorithms to proteins or genetic sequences is to map them to a feature space via a symmetric and positive kernel and apply the algorithm in the feature space. The kernel value $k(x, y)$ between two words $x, y$ will then be a measure of the proximity between the two words $x, y$.

### 2.1 Spectral kernel

A basic kernel to measure the proximity between two words $x, y$ is to count the number of common subwords of a given length $d$. More precisely, for $x \in \bigcup_n \mathcal{A}^n$ and $s \in \mathcal{A}^d$, set

$$N_s(x) = \text{number of occurence of } s \text{ in } x.$$

Define the spectral kernel on $\cup_n \mathcal{A}^n$ by

$$k(x,y) = \sum_{s \in \mathcal{A}^d} N_s(x) N_s(y) \text{ for all } x, y \in \bigcup_n \mathcal{A}^n\ .$$

1. Is $k$ positive semi-definite ?
2. Propose an algorithm to compute $k(x, y)$.
3. What is the complexity of your algorithm ?

### 2.2 Substring kernel

Instead of counting common subwords, we can count common substrings. For $0 < \alpha < 1$, a word $x$ and a string $s$ of length $|s| = d$, define

$$\phi_s(x) = \sum_{i_1 < \ldots < i_d} \mathbf{1}_{x[i_1,\ldots,i_d]=s} \ \alpha^{i_d - i_1 + 1} ,$$

and the kernel

$$k_d^{\phi}(x,y) = \sum_{s \in \mathcal{A}^d} \phi_s(x)\phi_s(y) . \tag{1}$$

Direct computation of $k(x,y)$ by enumerating all substring is computationally intensive. Below, we explain how to compute $k(x,y)$ with an algorithm based on dynamic programming.

1. Let us consider
$$\psi_s(x) = \sum_{i_1 < \ldots < i_d} \mathbf{1}_{x[i_1,\ldots,i_d]=s} \ \alpha^{|x| - i_1 + 1} .$$
Prove that for a word $v$ and two letters $a, b$ we have
$$\phi_{vb}(xa) = \phi_{vb}(x) + \alpha \mathbf{1}_{a=b} \ \psi_v(x) \quad \text{and} \quad \psi_{vb}(xa) = \alpha\, \psi_{vb}(x) + \alpha\, \mathbf{1}_{a=b} \ \psi_v(x) .$$
2. Check that we also have
$$\phi_{va}(x) = \alpha \sum_i \mathbf{1}_{x[i]=a} \psi_v(x[1:i-1]) \quad \text{and} \quad \psi_{va}(x) = \sum_i \mathbf{1}_{x[i]=a} \ \psi_v(x[1:i-1]) \ \alpha^{|x|-i+1} .$$
3. We now prove that $k_d^{\phi}(x,y)$ can be computed recursively from $k_{d-1}^{\psi}$, where $k_{d-1}^{\psi}$ is defined by (1) with $\phi$ replaced by $\psi$ and $d$ replaced by $d-1$. Prove that
$$\begin{aligned} k_d^{\phi}(xa,y) &= k_d^{\phi}(x,y) + \alpha \sum_{v \in \mathcal{A}^{d-1}} \psi_v(x)\phi_{va}(y) , \\ &= k_d^{\phi}(x,y) + \alpha^2 \sum_i \mathbf{1}_{y[i]=a} \ k_{d-1}^{\psi}(x, y[1:i-1]) . \end{aligned}$$
4. So, all we need is to compute $k_{d-1}^{\psi}$. This computation can be performed recursively according to the next two formulas (check them!) :
$$(i) \quad k_d^{\psi}(xa,y) = \alpha k_d^{\psi}(x,y) + \sum_i \mathbf{1}_{y[i]=a} \ k_{d-1}^{\psi}(x, y[1:i-1]) \alpha^{|y|-i+2} ,$$
$$(ii) \quad k_d^{\psi}(xa,yb) = \alpha k_d^{\psi}(x,yb) + \alpha k_d^{\psi}(xa,y) - \alpha^2 k_d^{\psi}(x,y) + \mathbf{1}_{a=b} \ \alpha^2 k_{d-1}^{\psi}(x,y) .$$
5. Check that the overall complexity is $O(d|x||y|)$.

# 3 RKHS associated to the Gaussian kernel

We consider the Gaussian kernel $k_\sigma(x,y) = \exp\left(\frac{-\|x-y\|^2}{2\sigma^2}\right)$. We denote the Fourier transform in $\mathbb{R}^d$ by

$$\mathbf{F}[f](\omega) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(t) e^{-\mathrm{i}\langle \omega, t\rangle} \mathrm{d}t, \quad \text{for } f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) \text{ and } \omega \in \mathbb{R}^d .$$

The linear span

$$\mathcal{H}_\sigma = \left\{ f \in C_0(\mathbb{R}^d) \cap L^1(\mathbb{R}^d) \text{ such that } \int_{\mathbb{R}^d} \left|\mathbf{F}[f](\omega)\right|^2 e^{\sigma|\omega|^2/2} \, \mathrm{d}\omega < +\infty \right\}$$

is endowed with the scalar product

$$\langle f, g \rangle_{\mathcal{H}_\sigma} = (2\pi\sigma^2)^{-d/2} \int_{\mathbb{R}^d} \overline{\mathbf{F}[f](\omega)} \mathbf{F}[g](\omega) e^{\sigma|\omega|^2/2} \, \mathrm{d}\omega .$$

1. Check that
$$\langle k_\sigma(x,.), f\rangle_{\mathcal{H}_\sigma} = \mathbf{F}^{-1}\big[\mathbf{F}[f]\big](x) = f(x) \quad \forall\ f \in \mathcal{H}_\sigma, \forall\ x \in \mathbb{R}^d.$$
2. What is the RKHS associated to $k_\sigma$ ?
3. How does $\mathcal{H}_\sigma$ evolve when $\sigma$ grows ?