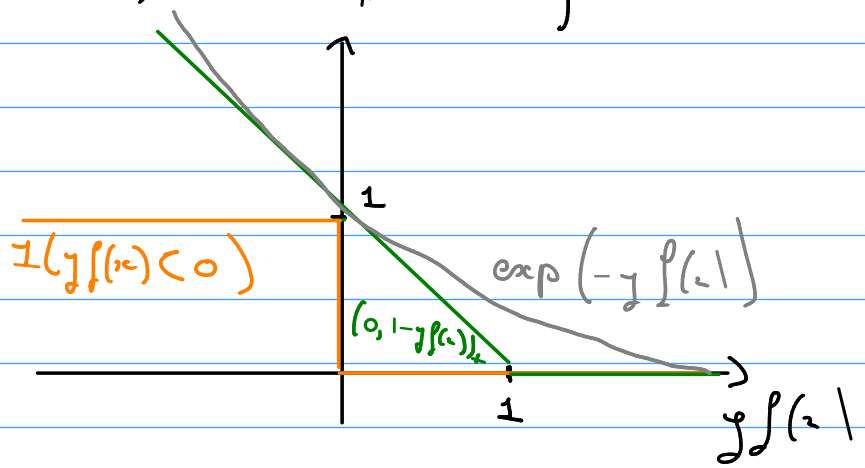
TD 7 : SVM, RKHS

Exercise 11 : SVM

$$\mathcal{F} = \{ f_w : \|w\| \leq M \} \qquad \hat{R}_{\varphi, \mathcal{F}}(x) = \operatorname{sign}\left(\hat{f}_{\varphi, \mathcal{F}}(x) \right)$$

where $\hat{f}_{\varphi, \mathcal{F}} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} \varphi(-y_i f(x_i))$

with $\varphi(x) = (1+x)_+$ the hinge loss.



Relaxation :

$$\hat{f}_{\varepsilon, \mathcal{F}} = \operatorname{argmin} \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - y_i f_w(x_i))_+ + \lambda \|w\|^2 \right\}$$

1. At optimum, $\partial_w \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - y_i f_w(x_i))_+ + \lambda \|w\|^2 \right\} \ni 0$

- $\partial_w \lambda \|w\|^2 = \{ 2\lambda w \}$
- Let $i \in \{1, \dots, n\}$,
  - if $y_i \langle w, x_i \rangle > 1$, $\partial_w (y_i \langle w, x_i \rangle) = \{0\}$
  - if $y_i \langle w, x_i \rangle < 1$, $\partial_w (y_i \langle w, x_i \rangle) = \{ -y_i x_i \}$

- if $y_i \langle \omega, x_i \rangle = 1$, $\partial_\omega (y_i \langle \omega, x_i \rangle) = y_i x_i [-1, 0]$

So that proves that $\hat{\omega} \in \text{Span}\{x_i : i = 1, \dots, n\}$

2. Let's note $\hat{\omega} = X^T \beta$, Then the problem becomes:

$$\begin{aligned}
&\underset{\beta}{\text{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - y_i \langle X^T \beta, x_i \rangle)_+ + \lambda \| X^T \beta \|^2 \right\} \\
&= \underset{\beta}{\text{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(1 - y_i \beta (X x_i)^T\right)_+ + \lambda \beta^T X X^T \beta \right\} \\
&= \underset{\beta}{\text{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(1 - y_i (K\beta)_i\right)_+ + \lambda \beta^T K \beta \right\}
\end{aligned}$$

where $K = \left[\langle x_i, x_j \rangle \right]_{1 \le i,j \le n}$

3. The problem is equivalent to

$$\begin{aligned}
\hat{\beta} = \underset{\substack{\beta, \xi \in \mathbb{R}^n \text{ s.t} \\ y_i (K\beta)_i \ge 1 - \xi_i \\ \xi_i \ge 0}}{\text{argmin}} & \left\{ \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \beta^T K \beta \right\}
\end{aligned}$$

4. Let's write the Lagrangian:

$$\begin{aligned}
\mathcal{L}(\beta, \xi, \alpha, \nu) &= \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \beta^T K \beta - \sum_{i=1}^{n} \alpha_i (y_i (K\beta)_i + \xi_i - 1) \\
&\quad - \sum_{i=1}^{n} \nu_i \xi_i \\
&= \xi^T \frac{\mathbf{1}}{n} + \lambda \beta^T K \beta - \left(\text{diag}(y)\alpha\right)^T K \beta - (\mu + \alpha)^T \xi + \mu^T \mathbb{1}
\end{aligned}$$

Then, $\nabla_\beta L = 2\lambda K\beta - K\,\mathrm{diag}(y)\alpha = K(2\lambda\beta - \mathrm{diag}(y)\alpha)$

So $\beta = \dfrac{\mathrm{diag}(y)\alpha}{2\lambda}$ is solution to $\nabla_\beta L = 0$

$\nabla_\xi L = \dfrac{\mathbb{1}}{n} - \alpha - \nu$

So $\alpha + \nu = \dfrac{\mathbb{1}}{n}$ In order to have $\nabla_\xi L = 0$

- The primal feasibility gives
  $-y_i(K\hat{\beta})_i - (1-\hat{\xi}_i) \geq 0$
  $-\hat{\xi}_i \geq 0$

- The dual feasibility gives
  $-\alpha \geq 0$
  $-\nu \geq 0$

So, all in all,

$$\hat{\beta}_i = y_i\hat{\alpha}_i/(2\lambda)$$
$$\min(\hat{\alpha}_i, y_i(K\hat{\beta})_i - (1-\hat{\xi}_i)) = 0$$
$$\min(1/n - \hat{\alpha}_i, \hat{\xi}_i) = 0$$

5. Complementary slackness:

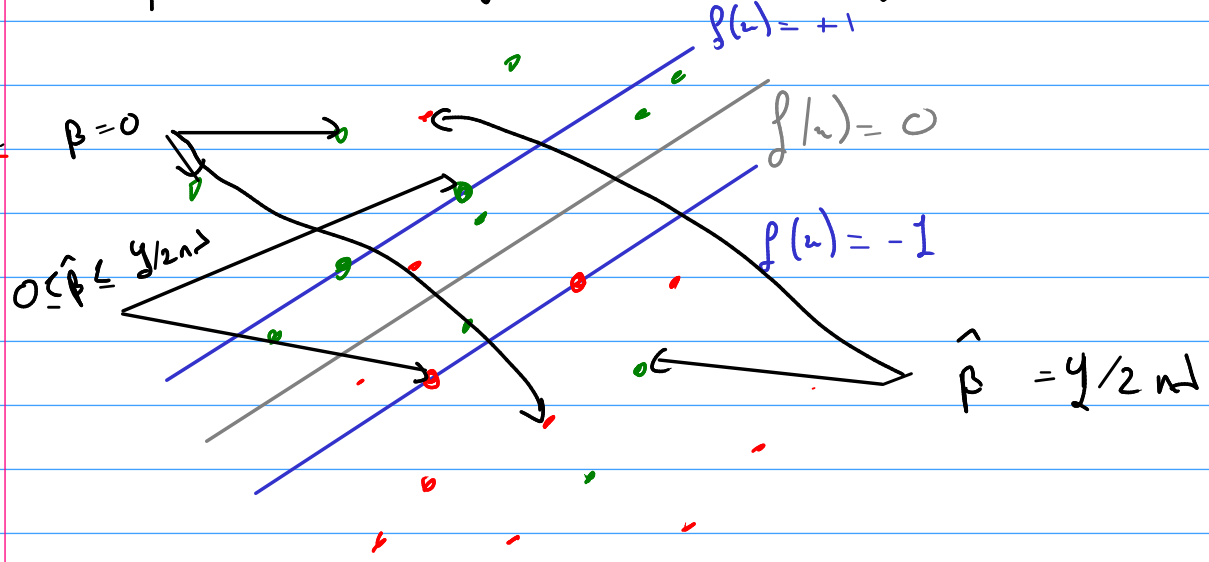
$$\forall i,\ \alpha_i(y_i f(x_i) + \xi_i - 1) = 0 \text{ and } \nu_i\xi_i = 0$$

And so, $\forall i,\ \beta_i(y_i f(x_i) + \xi_i - 1) = 0$ and $\left(\beta_i - \dfrac{y_i}{2\lambda n}\right)\xi_i = 0$

- If $\beta_i = 0$, then $\xi_i = 0$ and $y_i f(x_i) \geq 0$
- If $0 < \beta_i < \dfrac{y_i}{2\lambda n}$ then $\xi_i = 0$ and $y_i f(x_i) + \xi_i - 1 = 0$
  So $y_i f(x_i) = 1$

- if $\beta_i = \frac{y_i}{2\lambda n}$ then $y_i f(x_i) + \xi_i = 1$ so $y_i f(x_i) \leq 1$

6.



7. If we use the conditions found above we can minimize wrt β and ξ and by strong duality, we obtain that the problem is equivalent to

$$\hat{\alpha} = \underset{0 \leq \alpha_i \leq 1/n}{\arg\max} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{4\lambda} \sum_{i,j=1}^{n} K_{ij} y_i y_j \alpha_i \alpha_j \right\}$$

QP problem

Link with kernels: $f \in \mathcal{H}$: RKHS

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

Representer theorem: $\hat{f}(x) = \sum_{i=1}^{n} \hat{\beta}_i K(x_i, x)$

Exercise n° 12 : MKMS

1. Let k_1 and k_2 be two PSD kernels on $\mathcal{X}$.

Let $\lambda, \mu > 0$.

Let $x_1, \dots, x_n \in \mathcal{X}$, $c_1, \dots, c_n \in \mathbb{R}$,

$$\sum_i \sum_j c_i c_j (\lambda k_1 + \mu k_2)(x_i, x_j)$$

$$= \lambda \underbrace{\left(\sum_i \sum_j c_i c_j k_1(x_i, x_j) \right)}_{\geq 0} + \mu \underbrace{\left(\sum_i \sum_j c_i c_j k_2(x_i, x_j) \right)}_{\geq 0}$$

$$\geq 0 .$$

2. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ s.t

- $k(x, y) = k(y, x)$
- $\forall x, y, \; k(x, y) \geq 0$

Let's try to find a counter example with $n = 2$

$$\det\left(\begin{pmatrix} a & c \\ c & b \end{pmatrix} - X I_2 \right) = (a - X)(b - X) - c^2$$

And we can see that -1 is an eigenvalue of $\begin{pmatrix} a & c \\ c & b \end{pmatrix}$ if $c = \sqrt{(a+1)(b+1)}$

So in general, k is not a PSD Kernel

3. Let k_1 and k_2 be two PSD kernels on $\mathcal{X}$.

Let $x_1, \dots, x_n \in \mathcal{X}$,

Let us consider the matrices $(K_1)_{i,j} = \left(k_1(x_i, x_j)\right)_{i,j}$

$(K_2)_{i,j} = \left(k_2(x_i, x_j)\right)_{i,j}$

K_1 and K_2 are PSD (Let us consider K_x the matrix of product kernel)

Hint: use the $X^T X$ decomposition!

$K_1 = Y_1^T Y_1 \quad K_2 = Y_2^T Y_2$

$K_{x\,i,j} = K_{1\,i,j}\, K_{2\,i,j} = K_{1\,i,j}\, K_{2\,j,i}$ (since PSD)

$= Y_{1\,i}\, Y_{1\,j}^T\, Y_{2\,j}\, Y_{2\,i}^T$

$= \mathrm{tr}\left(Y_{1\,i}\, Y_{1\,j}^T\, Y_{2\,j}\, Y_{2\,i}^T\right)$

$= \mathrm{tr}\left(\left(Y_{2\,i}^T Y_{1\,i}\right)\left(Y_{2\,j}^T Y_{2\,j}\right)\right)$

$= \left(Y_{2\,i}^T Y_{1\,i}\,,\, Y_{2\,j}^T Y_{2\,j}\right)_F$

Here K_x is a Hadar Matrix and is PSD

So $k_1 k_2$ is PSD

Bonus: Show that the limit of PSD kernels is PSD

Bonus: Show that if k is PSD, $\exp(k)$ is PSD.

4. $(\Omega, \mathcal{E}, P)$: Probability space.

$k : \mathcal{E} \times \mathcal{E} \longrightarrow \mathbb{R}$

$A, B \mapsto P(A \cap B) - P(A)P(B)$

Let $A, B \in \mathcal{E}$,

$$P(A \cap B) - P(A)P(B) = E(\mathbb{1}_A \mathbb{1}_B) - E(\mathbb{1}_A)E(\mathbb{1}_B)$$

Let $A_1, \dots, A_n \in \mathcal{E}$, $c_1, \dots, c_n \in \mathbb{R}$,

$$\sum_{i,j} c_i c_j \left[P(A_i \cap A_j) - P(A_i)P(A_j)\right]$$

$$= \sum_{i,j} c_i c_j \left[E(\mathbb{1}_{A_i}\mathbb{1}_{A_j}) - E(\mathbb{1}_{A_i})E(\mathbb{1}_{A_j})\right]$$

$$= E\left(\sum_{i,j} c_i c_j \mathbb{1}_{A_i}\mathbb{1}_{A_j}\right) - \sum_{i,j} c_i c_j E(\mathbb{1}_{A_i})E(\mathbb{1}_{A_j})$$

$$= E\left(\left(\underbrace{\sum_i c_i \mathbb{1}_{A_i}}_{Y}\right)^2\right) - \left(\sum_i c_i E(\mathbb{1}_{A_i})\right)^2$$

$$= E(Y^2) - E(Y)^2 \geq 0$$

So k is PSD

5. k PSD on $\mathcal{E}$, $\bar{k} : \mathcal{P}(\mathcal{E}) \times \mathcal{P}(\mathcal{E}) \longrightarrow \mathbb{R}$

$A, B \mapsto \sum_{a \in A, b \in B} k(a,b)$

By Aronszajn's Theorem, there exists a Hilbert space $\mathcal{H}$ and a mapping $\phi : \mathcal{E} \to \mathcal{H}$ s.t. $\forall a, b \in \mathcal{E}$, $k(a,b) = (\phi(a), \phi(b))_{\mathcal{H}}$

Let's introduce $\bar{\phi} : \mathcal{P}(\mathcal{E}) \to \mathcal{H}$

$A \mapsto \sum_{a \in A} \phi(a)$

then $\forall A, B \in \mathcal{P}(\Sigma)$ $\bar{k}(A,B) = \langle \bar{\phi}(A), \bar{\phi}(B) \rangle_{\mathcal{D}}$

So $\bar{k}$ is a PSD kernel.

Bonus Kernels:

- Pointwise limit of PSD kernels
- Exp of PSD kernels.
- min on $\mathbb{R}_+$
- $1/\max$ on $\mathbb{R}_+$

  } min / max on $\mathbb{R}_+$

- GCD on $\mathbb{N}$
- $1/\text{LCM}$ on $\mathbb{N}$

  } GCD / LCM on $\mathbb{N}$

- $k(x,y) = 1/(1-xy)$ on $(-1,1)$
- $k(x,y) = e^{-(x-y)^2}$ (Gaussian kernel)
- $k(x,y) = \cos(x-y)$