Kernel

Definition: is a real-valued function of two arguments. $\forall x, x' \in \mathcal{X} \neq \emptyset$, $\phi : \mathcal{X} \mapsto \mathcal{H}$(Hilbert Space),$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $k(x,x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$is a positive definite kernel

Properties Symmetric $k(x,x') = k(x',x)$ Positive $\| \sum_{i=1}^n \alpha_i \phi(x_i) \|_{\mathcal{H}}^2 \geq 0$

p.d.: $k_1 + k_2$; $k_1 \times k_2$; $ck_1, c > 0$; $\lim_{n\to\infty} k_n = k$; k^{-1}; $e^k = \lim_{n\to\infty} \sum_{i=0}^n \frac{k^i}{i!}$

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

Example 1: p.d. $\min(x,y) = \int_0^\infty \mathbb{I}_{t\leq x} \mathbb{I}_{t \leq y} dt$

$(x,y) \in \mathbb{R}^+ \times \mathbb{R}^+$, where $\mathbb{R}^+ = \{x \in \mathbb{R}; x \geq 0\}$

$\int_0^\infty \mathbb{I}_{t\leq x} \mathbb{I}_{t\leq y} dt = \int_0^\infty \mathbb{I}_{t \leq \min(x,y)} dt = \int_0^{\min(x,y)} dt = \min(x,y)$

$K(x,y) = \min(x,y) = \int_0^\infty \mathbb{I}_{t\leq x} \mathbb{I}_{t\leq y} dt = \min(y,x) = K(y,x)$ symmetric

$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \min(x,y) = \int_0^\infty \sum_{i=1}^n \alpha_i \mathbb{I}_{t\leq x} \sum_{j=1}^n \alpha_j \mathbb{I}_{t\leq y} dt = \int_0^\infty (\sum_{i=1}^n \alpha_i \mathbb{I}_{t\leq x})^2 dt \geq 0$

Example 2: not p.d. $\max(x,y)$ over $\mathbb{R}^+$.

Let $x_1 = 1, x_2 = 2, \alpha_1 = 2, \alpha_2 = 2$; det $\begin{bmatrix} 1 & 2 \\ 2 & 2 \end{bmatrix} = -2$

Example 3: p.d. $K_1(A,B) = P(A \cap B)$ $P(A) = E[\mathbb{I}_A]$

$K_1(A,B) = P(A\cap B) = P(B \cap A) = K_1(B,A)$ sym

$K_1(A,B) = P(A \cap B) = E[\mathbb{I}_A \mathbb{I}_B]$

$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j E[\mathbb{I}_{A_i} \mathbb{I}_{A_j}] = E[\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \mathbb{I}_{A_i} \mathbb{I}_{A_j}] = E[(\sum_{i=1}^n \alpha_i \mathbb{I}_{A_i})^2] \geq 0$

Example 4: p.d. $K_2(A,B) = P(A\cap B) - P(A)P(B)$

$K_2(A,B) = P(A \cap B) - P(A)P(B) = E[\mathbb{I}_A \mathbb{I}_B] - E[\mathbb{I}_A] E[\mathbb{I}_B] = Cov[\mathbb{I}_A, \mathbb{I}_B]$

$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j Cov[\mathbb{I}_{A_i}, \mathbb{I}_{A_j}] = Cov[\sum_{i=1}^n \alpha_i \mathbb{I}_{A_i}, \sum_{j=1}^n \alpha_j \mathbb{I}_{A_j}] = Var[\sum_{i=1}^n \alpha_i \mathbb{I}_{A_i}]$

Example 5: p.d.

$k(x,x') = \frac{1}{1-xx'} = \sum_{k=0}^\infty (xx')^k$;$x, x' \in (-1,1)$;

$\ln(1+xx')$, $x = (20,1)$, $\alpha = (0.5, -1)$ not p.d.

$\sum_{i,j}^n \alpha_i \alpha_j \frac{1}{1-x_i x_j} = \sum_{i,j}^n \alpha_i \alpha_j \sum_{k=0}^\infty (x_i x_j)^k = \sum_{k=0}^\infty (\sum_i^n \alpha_i x_i)^{2k}$

$2^{x+x'} = 2^x 2^{x'} = \phi(x)\phi(x')$

$2^{x+x'} = \exp[xx' \ln 2] = \lim_{n\to\infty} \sum_{i=0}^n \frac{(\ln 2)^i (xx')^i}{i!}$

Example 6: p.d.

$k(x,x') = \cos(x+x') = \cos(x)\cos(x') - \sin(x)\sin(x') = k(w,w') - k(v,v')$

In the region of $\cos(x) < \sin(x)$, $k(x,x') < 0$

$\cos(x-x') = \cos(x)\cos(x') + \sin(x)\sin(x') = k(w,w') + k(v,v')$ Sum of p.d. is still p.d.

$\sin(x+x') = \sin(x)\cos(x') + \cos(x)\sin(x')$; $2\sum_{i,j} \sin(x_i)\cos(x_j)$ not p.d.

Cauchy-Schwarts inequity for kernels $k^2(x,x') \leq k(x,x)k(x',x')$

Proof: $n = 2, x = (x_1, x_2), \alpha = (\alpha_1, \alpha_2)$

$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(x_i, x_j) \geq 0 \iff$ the Gram matrix $\begin{pmatrix} k(x_1,x_1) & k(x_1,x_2) \\ k(x_2,x_1) & k(x_2,x_2) \end{pmatrix}$

semi-positive definite or equivalent determinant ≥ 0

$k(x_1,x_1)k(x_2,x_2) - k(x_1,x_2)k(x_2,x_1) \geq 0 \Rightarrow k(x_1,x_1)k(x_2,x_2) \geq k^2(x_1,x_2)$

RKHS

Reproducing Kernel Hilbert Space

Hilbert Space is a complete inner product space;

Inner product space is a vector space with an inner product (dot product, scalar product), a vector space (H,+,$\cdot$) over $\mathbb{R}$ ($\cdot$ scalor multiplication);

Dot product $\vec{a}\vec{b} = a_x b_x + a_y b_y = |\vec{a}||\vec{b}|\cos(\theta)$is a mapping: $H \times H \to \mathbb{R}$

Aronsjar Theorem: A p.d. k, there exist $\mathcal{H}$ and ϕ such that $k(x,x') = \langle \phi(x), \phi(x')_{\mathcal{H}}$ is true.

Inverse: A function k: $\mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ verifies $k(x,x') = \langle \phi(x), \phi(x')_{\mathcal{H}}$, then it is a positive kernel. $x, x' \in \mathcal{X} \neq \emptyset, \phi \in \mathcal{H}$

RKHS construction

For constructing $t \mapsto k(t,x), x \in \mathbb{R}$; $f : \mathcal{X} \mapsto \mathbb{R}$, add linear combinations

$f(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$;

$g(x) = \sum_{j=1}^m \beta_j k(x, y_j)$;

$\langle f, g \rangle = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_i k(x_i, y_j)$

not depend on the "represenation" in term of $\left\{ \begin{smallmatrix} x_1, .., x_n \\ \alpha_1, .., \alpha_n \end{smallmatrix} \right\}$; $\left\{ \begin{smallmatrix} y_1, .., y_m \\ \beta_1, .., \beta_m \end{smallmatrix} \right\}$

Definition: $X \neq \emptyset$, $\mathcal{H}$ is a Hilbert Space of function $\mathcal{X} \mapsto \mathbb{R}$

$\mathcal{H}$ is a RKHS when there is a function k: $\mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ such that

1. $k(\cdot, x) : t \mapsto k(t,x) \in \mathcal{H}$ for all $x \in \mathcal{X}$
2. $\langle f, k(\cdot,x) \rangle_{\mathcal{H}} = f(x), \forall f \in \mathcal{H}, x \in \mathcal{X}$ Reproducing Property

f:function;k:argument

$\langle f, k(\cdot,x) \rangle = \sum_{i=1}^n \alpha_i k(x_i, x) = f(x)$, $f \in \left\{ \begin{smallmatrix} x_1, .., x_n \\ \alpha_1, .., \alpha_n \end{smallmatrix} \right\}$; $k(\cdot,x) = (x,1)^T$

$k(x,y) = \langle \phi(x), \phi(y) \rangle = \langle k(\cdot,y), k(\cdot,x) \rangle = \langle k(\cdot,x), k(\cdot,y) \rangle$

Properties

1. $\langle f, g \rangle = \langle g, f \rangle$ symmetry for any $f, g \in H$
2. $\langle \alpha f_1 + \beta f_2, g \rangle = \alpha \langle f_1, g \rangle + \beta \langle f_2, g \rangle$ for any $f, g \in H; \alpha, \beta \in \mathbb{R}$ Linearity
3. $\langle f, f \rangle \geq 0$ for all $f \in H$
4. $\|f\|^2 = \langle f, f \rangle = 0 \iff f = 0_H$; a Norm on H

- Proof

Step 1 check that $\langle f, g \rangle$ is p.d.;

$f_1, ..f_n$, scalar $\gamma_1, .., \gamma_n$

$\sum_{i=1}^n \sum_{j=1}^n \gamma_i \gamma_j \langle f_i, f_j \rangle = \langle \sum_{i=1}^n \gamma_i f_i, \sum_{j=1}^n \gamma_j f_j \rangle \geq 0, g \in H$

Step 2 Use Cauchy-Schwarz inequality for $\langle f, g \rangle$ $x \in \mathcal{X}, f \in \mathcal{H}$

$|f(x)|^2 = |\langle f, k(\cdot,x) \rangle|^2 \leq \|f\|^2 \|k(\cdot,x)\|^2 = \|f\|^2 k(x,x)$

then for any $x \in \mathcal{X}$, $\|f\|^2 = \langle f, f \rangle = 0 \implies |f(x)|^2 = 0 \implies f(x) = 0$

We have shown that $(H, \langle \cdot, \cdot \rangle)$ just constructed to a inner product space pre-Hilbert Space.

A metric space is complete for an inner product when it cantains the limit fo all the Cauchy sequences for this inner product.

It can be completed into a Hilbert Space by including the limits of convergent Cauchy sequances

Exapmple 1: RKHS over $\mathcal{X} \in \mathbb{R}^d, k(x,y) = x^T y$ The RKHS with kernel k is

$\mathcal{H} = \{ f_w : \mathbb{R}^d \mapsto \mathbb{R};\ f_w(x) = w^T x; \quad w \in \mathbb{R}^d \}$

$\langle f_v, f_w \rangle_{\mathcal{H}} = v^T w \implies \langle f_v, f_v \rangle = \|f_v\|_{\mathcal{H}}^2 = \|v\|^2$

$\mathcal{H}$ is the RKHS associated with k $t \mapsto k(t,x) = x^T t = (x^T t)^T = t^T x = f_t(x)$

$\langle f, k(\cdot,x) \rangle = \langle f_w, f_x \rangle = x^T w = (x^T w)^T = w^T x = f_w(x)$

Example 2: RKHS over $\mathcal{X} \in \mathbb{R}^d, \quad k(x,y) = x^T y + c, c > 0$

$\mathcal{H} = \{ f : \mathbb{R}^d \mapsto \mathbb{R};\ f(x) = w^T x + w_0;\ w \in \mathbb{R}^d, w_0 \in \mathbb{R} \}$

$\langle f_{v,v_0}, f_{w,w_0} \rangle_{\mathcal{H}} = v^T w + \frac{1}{c} v_0 w_0$

Inner product:$\langle f_{v,v_0}, f_{v,v_0} \rangle = \|f_{v,v_0}\|_{\mathcal{H}}^2 = \|v\|^2 + \frac{v_0^2}{c}$

$f_{w,w_0} \leftrightarrow (w, w_0)^T \in \mathbb{R}^{d+1}$ which is a Hilber Space

Reproducing property: $\mathcal{H}$ contains all the functions $k(\cdot,x)$ for $x \in \mathbb{R}^d$

$\langle f_{v,v_0}, k(\cdot,x) \rangle = \langle f_{v,v_0}, f_{x,c} \rangle = v^T x + \frac{1}{c} v_0 c = f_{v,v_0}(x)$

Example 3: RKHS over $\mathcal{X} \in \mathbb{R}^d$ $K(x,y) = (x^T y)^2$
$\mathcal{H} = \{f_S : f_S(x) = x^T S x; \underset{(d,d)}{S}\}$ symmetric

Inner product: $\langle f_{S_1}, f_{S_2}\rangle_{\mathcal{H}} = \langle S_1, S_2\rangle_{\mathcal{F}} = \mathrm{tr}(S_1^T S_2) = \sum_{i,j=1}^{n} [S_1]_{ij}[S_2]_{ij}$

$k(y,x) = (y^T x)(y^T x) = y^T \cdot x x^T \cdot y = f_{xx^T}(y) \in \mathcal{H}$; xx^Tsym
Reproducing property: $\mathcal{H}$ contains all the functions $k(\cdot, x)$ for $x \in \mathbb{R}^d$
$\langle f_S, k(\cdot,x)\rangle_{\mathcal{H}} = \langle f_S, f_{xx^T}\rangle_{\mathcal{H}} = \langle S^T, xx^T\rangle_{\mathcal{F}} = \mathrm{tr}[S^T x x^T] = \mathrm{tr}[x^T S x] = x^T S x = f_S(x)$

Example 3: RKHS over $\mathbb{R}^d$ $K(x,y) = (x^T y + c)^2$
$(x^T y + c)(x^T y + c) = x^T y x^T y + 2c x^T y + c^2 = x^T y y^T x + 2c x^T y + c^2$
$\mathcal{H} = \{f : f(x) = x^T S x + 2w^T x + w_0;\ S \in \mathbb{R}^{d\times d}, w \in \mathbb{R}^d, w_0 \in \mathbb{R}\}$
the inner product $\langle f_{S_1,s_1,s_{10}}, f_{S_2,s_2,s_{20}}\rangle_{\mathcal{H}} = \langle S_1, S_2\rangle_{\mathcal{F}} + \frac{2 s_{10} s_{20}}{c} s_1^T s_2 + (\frac{s_{10}s_{20}}{c})^2$
Reproducing property $\langle f_{S,w,w_0}, k(\cdot,y)\rangle_{\mathcal{H}} = \langle f_{S,w,w_0}, f_{yy^T,2cy,c^2}\rangle_{\mathcal{H}}$
$= \langle S, yy^T\rangle_{\mathcal{F}} + \frac{2cy^T w}{2c} + \frac{w_0 c^2}{c^2} = y^T S y + y^T w + w_0 = f_{S,w,w_0}(y)$

Definition2 $X \neq \emptyset$, $\mathcal{H}$ is a Hilbert Space of function $\mathcal{X} \mapsto \mathbb{R}$
$\mathcal{H}$ is a RKHS if and only if for any $f \in \mathcal{H}$, $x \in \mathcal{X}$
the evaluation function $\mathcal{H} \mapsto \mathbb{R}$: $F_x : f \mapsto f(x)$ is continuous
$f, g \in \mathcal{H}$ if $\|f - g\|$ is small then their different $|f(x) - g(x)|$ is small.
F_x is continuous. if $\|f-g\|_{\mathcal{H}} < \delta \implies |f(x) - g(x)| < \varepsilon$ (might depend on x)
F_x is *C-Lipschitz* continuous when $|f(x) - g(x)| \leq c\|f-g\|_{\mathcal{H}}, c > 0, \forall f,\ g \in \mathcal{H}$
C-Lipschitz $\implies$ continuity.
$|f(x) - g(x)| = |(f-g)(x)| = |\langle f - g, k(\cdot,x)\rangle_{\mathcal{H}}| \leq \|f-g\|_{\mathcal{H}} \underbrace{\langle k(\cdot,x), k(\cdot,x)\rangle^{\frac{1}{2}}}_{k^{\frac{1}{2}}(x,x)}$

Riesz Representation Theorem: In any Hilber Space of function $\mathcal{X} \mapsto \mathbb{R}$ for which F_x is continuous for each $x \in \mathcal{X}$, then there is an unique element of $\mathcal{H}$, notated g_x, for which $f(x) = \langle f, g_x\rangle_{\mathcal{H}}$ for each $f \in \mathcal{H}$, $\quad g_x(\cdot) = k(\cdot, x)$.

Create a vector space by adding all the finite linear combination of $k(\cdot,x), x \in \mathcal{X}$
$V = \{f : \mathcal{X} \to \mathbb{R},\ f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i); n \geq 1;\ x_1,..,x_n \in \mathcal{X}; \alpha_1,..,\alpha_n \in \mathbb{R}\}$
$f \in V \leftrightarrow \left\{\begin{smallmatrix} x_1,..,x_n \\ \alpha_1,..,\alpha_n\end{smallmatrix}\right\}$ $g \in V \leftrightarrow \left\{\begin{matrix} y_1,..,y_m \\ \beta_1,..,\beta_m\end{matrix}\right\}$ $f + g \leftrightarrow \left\{\begin{matrix} x_1,..,x_n, y_1,..,y_m \\ \alpha_1,..,\alpha_n, \beta_1,..,\beta_m\end{matrix}\right\}$
$\gamma f \leftrightarrow \left\{\begin{matrix} x_1,..,x_n \\ \gamma\alpha_1,..,\gamma\alpha_n\end{matrix}\right\}, \gamma \in \mathbb{R}$

$$\gamma_1 f + \gamma_2 g \leftrightarrow \left\{\begin{matrix} \overbrace{x_1,..,x_n}^{z_1,..,z_n}, \overbrace{y_1,..,y_m}^{z_{n+1},..,z_{n+m}} \\ \underbrace{\gamma_1\alpha_1,..,\gamma_1\alpha_n}_{\delta_1,..,\delta_n}, \underbrace{\gamma_2\beta_1,..,\gamma_2\beta_m}_{\delta_{n+1},..,\delta_{n+m}}\end{matrix}\right\} \leftrightarrow h(x) = \sum_{i=1}^{n+m} \delta_i k(x, z_i)$$

$(\gamma_1 f + \gamma_2 g)(x) = \gamma_1 \sum_{i=1}^{n} \alpha_i k(x, x_i) + \gamma_2 \sum_{i=1}^{m} \beta_i k(x, y_i) = \gamma_1 f(x) + \gamma_2 g(x)$
The representation $\left\{\begin{smallmatrix} x_1,..,x_n \\ \alpha_1,..,\alpha_n\end{smallmatrix}\right\}$ of a function in V is not necessary unique
Define $\langle f, g\rangle = \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{m} \beta_i k(x_i, y_j)$ is a function $\mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$
$f \in V \leftrightarrow \left\{\begin{smallmatrix} x_1,..,x_n \\ \alpha_1,..,\alpha_n\end{smallmatrix}\right\}$; $g \in V \leftrightarrow \left\{\begin{matrix} y_1,..,y_m \\ \beta_1,..,\beta_m\end{matrix}\right\}$

$$\langle f, g\rangle = \sum_{i=1}^{n} \alpha_i \underbrace{\sum_{j=1}^{m} \beta_i k(x_i, y_j)}_{g(x_i)} = \sum_{i=1}^{n} \alpha_i g(x_i) = \sum_{j=1}^{m} \beta_i \underbrace{\sum_{i=1}^{n} \alpha_i k(y_j, x_i)}_{f(y_j)} = \sum_{j=1}^{m} \beta_i f(y_j)$$

$\langle f, g\rangle$ does not depend on the particular representation of (f, g)
So it is a function $\mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$
$\langle f, k(\cdot, x)\rangle = \sum_{i=1}^{n} \alpha_i k(x_i, x) = f(x)$; $\langle k(\cdot,y), k(\cdot,x)\rangle = k(x,y)$

Example 1: $\underset{(n,n)}{k}$; $[k]_{ij} = k(x_i, x_j), k(x,y) = x^T y$, $\mathcal{X} = \{x_1, ..x_n\}$; $k = (k_1, ..k_n)$

$f : \mathcal{X} \mapsto \mathbb{R}; \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n)\end{bmatrix} \subset \mathbb{R}^n; k(\cdot, x_i) = \begin{bmatrix} k_{1i} \\ \vdots \\ k_{ni}\end{bmatrix} = k_i;\ \alpha_1,..,\alpha_n \in \mathbb{R}$
$\mathcal{H} = \{\alpha_1 k_1 + \cdots + \alpha_n k_n\} = \mathrm{Span}\{k_1, ..k_n\} = \mathbb{R}^n$ is a vector space.
$\langle f, g\rangle_{\mathcal{H}} = f^T k^{-1} g$; $\langle f, k(\cdot, x_i)\rangle = \langle f, k e_i = f^T \underbrace{k^{-1} k}_{I} e_i = f^T e_i = f(x_i)$

Decomposition $K = U\Lambda U^T = LL^T = k^{1/2} k^{1/2}$; $k^{1/2} = U \Lambda^{1/2} U^T$
$k_{ij} = \phi^T(x_i)\phi(x_j) = (\Lambda^{1/2} U_i)^T (\Lambda^{1/2} U_j)$

Orthogonality $u^T v = 0$; $A^T = A^{-1} \Rightarrow$normal and diagonalizable
Let $v = \mathrm{span}[k(\cdot, x_i), .., k(\cdot, x_n)]$ $\mathcal{V}$ is closed linear subspace of $\mathcal{H}$.
Then all minimizers of $J \in \mathcal{V}$, there is an unique decomposition
$g = g_v + g_\perp$ with $g_v \in \mathcal{V}$ $\forall g \in \mathcal{V}$, $\langle g_\perp, f\rangle = 0$
$\|g\|^2_{\mathcal{H}} = \|g_v + g_\perp\|^2_{\mathcal{H}} = \langle g_v + g_\perp, g_v + g_\perp\rangle = \langle g_v, g_v\rangle + \langle g_\perp, g_\perp\rangle + \underbrace{2\langle g_v, g_\perp\rangle}_{0} = \|g_v\|^2_{\mathcal{H}} + \|g_\perp\|^2_{\mathcal{H}}$

$g(x_i) = \langle g, k(\cdot, x_i)\rangle = \langle g_v + g_\perp, k(\cdot, x_i)\rangle = \langle g_v, k(\cdot, x_i)\rangle + \underbrace{\langle g_\perp, k(\cdot, x_i)\rangle}_{0} = g_v(x_i)$

$J(\theta, g) - J(\theta, g_v) = [g(x_i)] + \lambda\|g\|^2_{\mathcal{H}} - [g_v(x_i)] - \lambda\|g_v\|^2_{\mathcal{H}} = \lambda\|g_\perp\|^2_{\mathcal{H}} \geq 0$
is free of θ,g is strictly increasing

Representer theorem $\alpha = (\alpha_1, .., \alpha_n)^T \in \mathbb{R}$ is the solution of min $J(\theta, g)$
$\forall\theta$, the function $g(.) = \sum_{i=1}^{n} \alpha_i k(x_i, .), g \in \mathcal{H}$ min $J(\theta, g)$
$\|g\|^2_{\mathcal{H}} = \langle g, g\rangle_{\mathcal{H}} = \langle \sum_{i=1}^{n} \alpha_i k(\cdot, x_i), \sum_{j=1}^{n} \alpha_j k(\cdot, x_j)\rangle_{\mathcal{H}} = \sum_{i,j=1}^{n} \alpha_i \alpha_j \underbrace{\langle k(\cdot, x_i), k(\cdot, x_j)\rangle_{\mathcal{H}}}_{k(x_i, x_j)} = \underset{(1,n)}{\alpha^T} \underset{(n,n)}{K} \underset{(n,1)}{\alpha}$

$g(x_i) = \sum_{j=1}^{n} \alpha_j k(x_i, x_j) = \sum_{j=1}^{n} \alpha_j [\underset{(n,n)}{K}]_{i,j} = [K\alpha]_i$

Matrix
transpose: $[A^{\mathrm{T}}]_{ij} = [A]_{ji}$;conjungate transpose / adjugate: $A^* = (\overline{A})^{\mathrm{T}} = \overline{A^{\mathrm{T}}}$
$\mathrm{tr}(A) = a_{11} + a_{22} + \cdots + a_{nn}$(sum of the elements on the main diagonal)
$\mathrm{span}(A) = \{\lambda_1 v_1 + \cdots + \lambda_r v_r \mid \lambda_1, \ldots, \lambda_r \in \mathbb{R}\}$the set of all finite linear combinations of elements of $A.v_1, \ldots, v_r$ be the column vectors of A.
$(A^T)^T = A$; $(AB)^T = A^T B^T$; $det(A^T) = det(A)$; $(A^T)^{-1} = (A^{-1})^T$
$A = U\Lambda U^{-1}$. Λ diag $A^n = U\Lambda^n U^{-1}$; $[AB]_{ij} = \sum_k A_{ik}B_{kj}$; $[ABC]_{ij} = \sum_{kl} A_{ik} B_{kl} C_{lj}$
invertible $\underset{n,n}{A}$ $AA^{-1} = I$;$det(A^{-1}) = \frac{1}{det(A)}$; $(A^{-1})^{-1} = A$; $(A^T)^{-1} = (A^{-1})^T$
$\nabla_x \underset{(m,n)}{A} \underset{(n,1)}{X} = \underset{(n,m)}{A^T}$; $\nabla_x \|g(x)\|^2 = \nabla_x \langle g(x), g(x)\rangle = 2[\underset{(n,m)}{\nabla_x g(x)}]\underset{(m,1)}{g(x)}$
$\nabla_x \langle g_1, g_2\rangle = \underset{(n,m)}{\nabla_x g_1} \underset{(m,1)}{g_2} + \underset{(n,m)}{\nabla_x g_2} \underset{(m,1)}{g_1}$; $\nabla_x \underset{(1,n)}{X^T} \underset{(n,n)}{B} \underset{(n,1)}{X} = 2 \underset{(n,n)}{B} \underset{(n,1)}{X}$;
$\nabla_x \|y - k\alpha\|^2 = 2k^T(k\alpha - y)$

Distance in feature space
$D_{k(x_1,x_2)} = \|\phi(x_1) - \phi(x_2)\|^2 = \langle \phi(x_1) - \phi(x_2), \phi(x_1) - \phi(x_2)\rangle =$
$\langle \phi(x_1), \phi(x_1)\rangle + \langle \phi(x_2), \phi(x_2)\rangle - 2\langle \phi(x_1), \phi(x_2)\rangle = k(x_1, x_1) + k(x_2, x_2) - 2k(x_1, x_2)$
Point to set:
$D_{k(x,S)} = \|\phi(x) - \mu\| = \|\phi(x) - \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)\| = \sqrt{k(x,x) - \frac{2}{n}\sum_{i=1}^{n} k(x, x_i) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j)}$
Centering data:
$k^c_{i,j} = \langle \phi(x_i) - \mu, \phi(x_j) - \mu\rangle = \langle \phi(x_i), \phi(x_j)\rangle - 2\langle \mu, \phi(x_i) + \phi(x_j)\rangle + \langle \mu, \mu\rangle$
$= k_{ij} - \frac{1}{n}\sum_{k=1}^{n}(K_{i,k} + K_{j,k}) + \frac{1}{n^2}\sum_{k,l=1}^{n} K_{k,l} = K - UK - KU + UKU = (I-)K(I-U)$

Kernel Function

Linear kernel $\mathcal{X} = \mathbb{R}^d; \mathcal{X} \times \mathcal{X} = \mathbb{R}; x \in \mathbb{R}^2, x = (x_1, x_2)^T;$
$k(x,x') = \langle x, x'\rangle_{\mathbb{R}^d} = x^T x' = x_1 x_1' + x_2 x_2' + \cdots$
$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(x_i, x_j) = \alpha^T x^T x \alpha = \|\alpha x\|^2 \geq 0$ is p.d.
$k(\cdot, x) : y \mapsto yx; f(x) = w^T x; \|f\| = \|w\|_2$
$k(x,x') = \langle \phi(x), \phi(x')\rangle_{\mathcal{H}}$

Binary $g : \mathbb{R}^2 \to \mathbb{R}; f(x) = \begin{cases} +1 & \text{if } g(x) \geq 0 \\ -1 & \text{if } g(x) < 0 \end{cases}$
Training set:$T = \{(x_i, y_i); x_i \in \mathcal{X}, y_i \in \{-1;+1\}\}$
$I_+ = \{i; y_i = +1\},\ I_- = \{i; y_i = -1\};$
of $I_+ = n_+, I_- = n_-; T = n = n_+ + n_-$
$C_+ = \frac{1}{n_+} \sum_{i \in I_+}^{n} x_i; C_- = \frac{1}{n_-} \sum_{i \in I_-}^{n} x_i; C = \frac{1}{2}(C_+ + C_-)$
$g(x) = \langle C_+ - C_-, X - C\rangle_{\mathbb{R}^2} = (X - C)^T (C_+ - C_-) = \langle X, C_+\rangle - \langle X, C_-\rangle + b$
$= \sum_{i=1}^n \alpha_i \langle x_i, x\rangle + b$
$\langle C_+, X\rangle = \frac{1}{n_+} \sum_{i \in I_+}^{n} \langle x_i, x\rangle; \langle C_-, X\rangle = \frac{1}{n_-} \sum_{i \in I_-}^{n} \langle x_i, x\rangle; \alpha_i = \begin{cases} \frac{1}{n_+} & y_i = +1 \\ \frac{-1}{n_-} & y_i = -1 \end{cases}$
$b = \langle C_-, C\rangle - \langle C_+, C\rangle = \frac{1}{2n_-^2} \sum_{(i,j) \in I_-} \langle x_i, x_j\rangle - \frac{1}{2n_+^2} \sum_{(i,j) \in I_+} \langle x_i, x_j\rangle$
$\langle C_+, C\rangle = \langle C_+, \frac{1}{2} C_+\rangle + \langle C_+, \frac{1}{2} C_-\rangle = \frac{1}{2n_+^2} \sum_{(i,j) \in I_+} \langle x_i, x_j\rangle + \frac{1}{2}\langle C_+, C_-\rangle$
$\langle C_-, C\rangle = \langle C_-, \frac{1}{2} C_+\rangle + \langle C_-, \frac{1}{2} C_-\rangle = \frac{1}{2}\langle C_+, C_-\rangle + \frac{1}{2n_-^2} \sum_{(i,j) \in I_-} \langle x_i, x_j\rangle$

Polynomial Kernel $k(x,y) = \langle \phi(x), \phi(y)\rangle_{\mathbb{R}} = (a x^T y + c)^d$
$x = (x_1, x_2) \in \mathbb{R}^2$
$\phi : \mathbb{R}^3 \to \mathbb{R}^2, \phi(x) = (x_1^2, \sqrt{2} x_1 x_2, x_2^2)^T$
$\phi(x)^T \phi(y) = \begin{bmatrix} x_1^2 \\ \sqrt{2} x_1 x_2 \\ x_2^2 \end{bmatrix}_{3 \times 1} \begin{bmatrix} y_1^2 & \sqrt{2} y_1 y_2 & y_2^2 \end{bmatrix}_{1 \times 3} = (x^T y)^2$
$\phi : \mathbb{R}^4 \to \mathbb{R}^2, \phi(x) = (x_1^2, x_1 x_2, x_2 x_1, x_2^2)^T$
$\langle \phi(x), \phi(y)\rangle_{\mathbb{R}} = x_1^2 y_1^2 + 2 x_1 x_2 y_1 y_2 + x_2^2 y_2^2 = \langle x, y\rangle_{\mathbb{R}}^2 = (x_1 y_1 + x_2 y_2)^2$
$\phi : \mathbb{R}^{2^3} \to \mathbb{R}^2, \phi(x) = (x_1^3, x_1^2 x_2, x_2 x_1^2, x_1 x_2 x_1, x_2 x_1 x_2, x_1 x_2^2, x_2^2 x_1, x_2^3)^T$
$\langle \phi(x), \phi(y)\rangle_{\mathbb{R}} = x_1^3 y_1^3 + 3 x_1^2 x_2 y_1^2 y_2 + 3 x_1 x_2^2 y_1 y_2^2 + x_2^3 y_2^3 = \langle x, y\rangle_{\mathbb{R}}^3 = (x_1 y_1 + x_2 y_2)^3$

$\mathcal{X} = \mathbb{R}^n$
$\phi(x) = \{x_{j1}, .., x_{jd}; 1 \leq j1, ..jd \leq n\}$, n^d iterms
$k(x,y) = \langle \phi(x), \phi(y)\rangle_{\mathbb{R}} = \langle x, y\rangle_{\mathbb{R}}^d$

$\phi(x)^T \phi(y) = (x^T y + 1)^2$
$\phi : \mathbb{R}^2 \mapsto \mathbb{R}^6, \phi(x) = (x_1^2, \sqrt{2} x_1, \sqrt{2} x_1 x_2, \sqrt{2} x_2, x_2^2, 1)$
$\phi : \mathbb{R}^2 \mapsto \mathbb{R}^9, \phi(x_1, x_2) = (x_1^2, x_2^2, 1, x_1, x_1, x_2, x_2, x_1 x_2, x_1 x_2)^T$

$K(x,y) = (1 + x^T y)^d$ **for** $d = 1, 2 \ldots$ is p.d.
$k(x,y) = (1 + x^T y)^d = (1 + y^T x)^d = k(y,x)$ is sym
Ture for $d = 0, 1$. $k_{d+1} = k_d k$ is p.d.
$k_d(x,y) = \phi_d^T(x) \phi_d(y)$
$\phi_{d+1}^T(x) \phi_{d+1}(y) = \phi_d^T(x) \phi_d(y) + [x_1 \phi_d(x)]^T y_1 \phi_d(y) + [x_2 \phi_d(x)]^T y_2 \phi_d(y)$
$= k_d(x,y) k_1(x,y) = k_{d+1}(x,y)$

$\phi : \mathbb{R}^2 \mapsto H, k(x,y) = \langle \phi(x), \phi(y)\rangle_H = x^T y - 1$
Choose $n = 1, x_1 = (1/2, 0)^T, \alpha_1 = 1$
$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(x_i, x_j) = k(x_1, x_1) = 1/4 - 1 = -3/4.$

Gaussian kernel
$K(X, X') = \exp[-\frac{1}{2}(X - X')^T \Lambda^{-1} (X - X')]$
$k(\cdot, x) : y \mapsto \exp[-\frac{1}{2\sigma^2}(y - x)^2]$ with parametor σ^2
ARD kernel= $\exp[-\frac{1}{2} \sum_{j=1}^{D} \frac{1}{\sigma_j^2}(x_j - x_j')^2]$ if Λ is diagonal
Isotropic kernel = $\exp\left(-\frac{\|X - X'\|^2}{2\sigma^2}\right)$ if Λ is spherical

Fisher kernel
$\phi(x, \theta_0) = \frac{d}{d\theta} \ln p_\theta(x); I(\theta) = E[\phi^2(X, \theta)]; k(x, x') = \frac{\phi(x,\theta_0)\phi(x',\theta_0)}{I(\theta_0)}$
$\psi(x, \theta_0) = \frac{\phi(x,\theta_0)}{\sqrt{I(\theta_0)}}, k(x, x') = \langle \psi(x, \theta_0), \psi(x', \theta_0)\rangle$ is p.d.
$X \sim Bern(\theta), 0 < \theta < 1; p_\theta(x) = \theta^x (1 - \theta)^{(1-x)}, x \in \{0, 1\},$
$E[X] = \sum_{x=0}^1 x p(x) = \theta, E[X^2] = \sum_{x=0}^1 x^2 p(x) = \theta,$
$Var[X] = E[(X - \theta)^2] = \theta(1 - \theta)$
$\phi(X, \theta) = \frac{d}{d\theta} \ln p_\theta(x) = \frac{x}{\theta} + \frac{1-x}{1-\theta} = \frac{x - \theta}{\theta(1-\theta)}$
$I(\theta) = E[\phi^2(X, \theta)] = \frac{E[(X-\theta)^2]}{\theta^2(1-\theta)^2} = \frac{V[X]}{\theta^2(1-\theta)^2} = \frac{\theta(1-\theta)}{\theta^2(1-\theta)^2} = \frac{1}{\theta(1-\theta)}$
$k(x, x') = \frac{\phi(x,\theta_0)\phi(x',\theta_0)}{I(\theta_0)} = \frac{(x-\theta_0)(x'-\theta_0)}{\theta_0^2(1-\theta_0)^2} \theta_0(1 - \theta_0) = \frac{(x-\theta_0)(x'-\theta_0)}{\theta_0(1-\theta_0)}$
$X = (X_1, X_2), x = (x_1, x_2)$ with $x_1 \in \{0, 1\}$ and $x_2 \in \{0, 1\}$.
$p_\theta(\vec{x}) \underset{x_1 \perp x_2}{=} p_\theta(x_1) p_\theta(x_2) = \theta^{x_1 + x_2}(1 - \theta)^{2 - x_1 - x_2}$
$\ln p_\theta(x) = (x_1 + x_2) \ln \theta + (2 - x_1 - x_2) \ln(1 - \theta)$
$\phi(\vec{x}, \theta) = \frac{d}{d\theta} \ln p_\theta(x) = \frac{x_1 + x_2}{\theta} + \frac{2 - x_1 - x_2}{1 - \theta} = \frac{2(\bar{x} - \theta)}{\theta(1-\theta)}$
$E[\bar{X}] = \theta, Var[\bar{X}] = E[(\bar{X} - \theta)^2] = \frac{\theta(1-\theta)}{2}$
$I(\theta) = \frac{4}{\theta^2(1-\theta)^2} \frac{\theta(1-\theta)}{2} = \frac{2}{\theta(1-\theta)}$
$k(x, x') = \frac{\phi(\vec{x},\theta_0)\phi(\vec{x}',\theta_0)}{I(\theta_0)} = \frac{2(\bar{x}-\theta_0) 2(\bar{x}'-\theta_0)}{\theta_0^2(1-\theta_0)^2} \frac{\theta_0(1-\theta_0)}{2} = \frac{2(\bar{x}-\theta_0)(\bar{x}'-2\theta_0)}{\theta_0(1-\theta_0)}$

Multivariate Fisher kernel
$x \in \mathbb{R}^d, \theta \in \mathbb{R}^k. \phi : \mathbb{R}^d \to \mathbb{R}^k.$
$\underset{(k,k)}{I} = E_{p_{\theta_0}}[\phi(X) \phi^T(X)]$ is p.d.
$K(x,y) = \phi(x)^T I^{-1} \phi(y)$ for any couple of points (x, y), both in $\mathbb{R}^d$
$I = (P^T \Lambda P; P^T P = Id; (I^{-1})^T = (P^T \Lambda^{-1} P)^T = P^T \Lambda^{-1} P = I^{-1}$
$K(x,y) = \phi(x)^T I^{-1} \phi(y) = [\phi(x)^T I^{-1} \phi(y)]^T = \phi(y)^T (I^{-1})^T \phi(x) = K(y,x)$
Choose $x_1, ..x_n \in \mathcal{X}; \alpha_1, ..\alpha_n \in \mathbb{R}$
$\sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) = \sum_{i,j=1}^n \alpha_i \alpha_j (I^{-\frac{1}{2}} \phi(x_i))^T (I^{-\frac{1}{2}} \phi(x_j)) = [\sum_{i=1}^n \alpha_i I^{-\frac{1}{2}} \phi(x_i)]^2 \geq 0$
$\phi(x) = \nabla_\theta \ln p_\theta(x)|_{\theta=\theta_0} = \nabla_\theta[-\frac{d}{2} \ln(2\pi) + \frac{1}{2}(\det \Lambda) - \frac{1}{2}(x - \theta_0)^T \Lambda (x - \theta_0)] =$
$\frac{2}{2}\Lambda(x - \theta_0) = \Lambda(x - \theta_0)$
$E[\phi(x) x \phi(x) y] = x^T E[(\phi(x))^T \phi(x)] y = x^T [E[\phi(x)^2] + \theta\theta^T] y$
$I = E[\phi(X) \phi^T(X)] = E[(\Lambda(x - \theta))^T \Lambda(x - \theta)] = \Lambda E[(x - \theta)^T (x - \theta)] \Lambda =$
$\Lambda \Lambda^{-1} \Lambda = \Lambda$
$K(x,y) = \phi(x)^T I^{-1} \phi(y) = (\Lambda(x - \theta))^T \Lambda^{-1} \Lambda(x - \theta)$
$= (x - \theta)^T \Lambda (\Lambda)^{-1} \Lambda (y - \theta) = (x - \theta)^T \Lambda (y - \theta)$

KRR
$J(g) = \sum_{i=1}^n (y_i - g(x_i))^2 + \lambda \|g\|_{\mathcal{H}}^2; \min_{g \in \mathcal{H}} J(g);$

$J(\sum_{i=1}^n \alpha_i k(x_i,\cdot)) = \| \underset{(n,1)}{Y} - \underset{(n,n)}{K} \underset{(n,1)}{\alpha} \|^2 + \underset{(1,1)}{\lambda} \underset{(1,n)}{\alpha^T} \underset{(n,n)}{K} \underset{(n,1)}{\alpha}$
$\nabla_\alpha J = \frac{\partial}{\partial\alpha}\|K\alpha - Y\|^2 + \lambda \frac{\partial}{\partial\alpha}\langle \alpha, K\alpha\rangle = 2(K\alpha - Y)K^T + \lambda(IK\alpha + K^T\alpha)$
$= 2(K\alpha - Y)K^T + 2\lambda K\alpha = 2K[(K+\lambda I)\alpha - Y]$
p.d. K, X sym, $K = K^T$, $X = X^T$; $K = P\Lambda P^T$; $I = PP^T$; Λdiag $\gamma_1, .., \gamma_n$.
$\lambda > 0, \gamma_i > 0$, $K + \lambda I = P(\Lambda + \lambda I)P^T$ is inversible.
$\nabla_\alpha J \overset{set}{=} 0 \implies \alpha^\star = (K+\lambda I)^{-1} Y$

$J(\vec{w}) = \frac{1}{N}\sum_{i=1}^N (y_i - (w_0 + \vec{w}^T\vec{x}_i))^2 + \lambda\|\vec{w}\|^2, \lambda \triangleq \frac{\sigma^2}{\tau^2}$
$\hat{\vec{w}}_{\text{ridge}} = (\lambda\vec{I}_D + \vec{X}^T\vec{X})^{-1}\vec{X}^T\vec{y}$; $\underset{(n,d)}{X} = (x_1, , x_n)^T$
Numerically stable computation: $\hat{\vec{w}}_{\text{ridge}} = \vec{V}(\vec{Z}^T\vec{Z} + \lambda\vec{I}_N)^{-1}\vec{Z}^T\vec{y}$
Primal problem: $J(\vec{w}) = (\vec{y} - \vec{X}\vec{w})^T(\vec{y} - \vec{X}\vec{w}) + \lambda\|\vec{w}\|^2$
Primal variables $\vec{w} = \vec{X}^T\vec{\alpha} = \sum_{i=1}^N \alpha_i\vec{x}_i$
Dual problem: $\vec{w} = X^T \underbrace{(XX^T + \lambda I_N)^{-1}}_{n,n}\vec{y} = \underbrace{(X^TX + \lambda I_N)^{-1}}_{d,d} X^T\vec{y}$
Dual variables: $\vec{\alpha} = (\vec{K} + \lambda\vec{I}_N)^{-1}\vec{y}$
Predictive mean: $y = f(\vec{x}) = \sum_{i=1}^N \alpha_i\vec{x}_i^T\vec{x} = \sum_{i=1}^N \alpha_i\kappa(\vec{x}_i, \vec{x})$
Weighted: $\arg\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n W_i(y_i - f(x_i))^2 + \lambda\|f\|^2$
$\arg\min_{\alpha\in\mathbb{R}^d} \frac{1}{n}(K\alpha - y)^T W(K\alpha - y)$
$\nabla_\alpha J(\alpha) = \frac{2}{n}(KWK\alpha - KWy) + 2\lambda K\alpha = \frac{2}{n}KW^{\frac{1}{2}}[(W^{\frac{1}{2}}KW^{\frac{1}{2}} + n\lambda I)W^{-\frac{1}{2}}\alpha - W^{\frac{1}{2}}y]$
$(W^{\frac{1}{2}}KW^{\frac{1}{2}} + n\lambda I)W^{-\frac{1}{2}}\alpha = W^{\frac{1}{2}}y$; $\alpha = W^{\frac{1}{2}}(W^{\frac{1}{2}}KW^{\frac{1}{2}} + n\lambda I)^{-1}W^{\frac{1}{2}}y$

String kernels

$\phi(\vec{x})$ # of times that substrings appears in string $\vec{x}$, $X = (\phi(x_1), , \phi(x_n))^T$.
$K(x, x') = \sum_{i\in\mathcal{A}^*} \lambda_i\phi_i(\vec{x})\phi_i(\vec{x}') = \sum_i \lambda_i w^T\phi(x_i))$ $\lambda_i \geq 0$; $\mathcal{A}^*$ set of strings.
$J(w) = \sum_i \lambda_i(y_i - w^TX_i)^2 = (y - Xw)^T\Lambda(y - Xw)$
$\nabla_w J(w) = \nabla_w(y^T\Lambda y + (Xw)^T\Lambda Xw - 2X^T\Lambda yw) = 2X^T\Lambda Xw - 2X^T\Lambda y$
$w^\star = (X^T\Lambda X)^{-1}X^T\Lambda y$
$J(w) = (y - Xw - w_0\mathbf{1})^T\Lambda(y - Xw - w_0\mathbf{1}) =$
$y^Ty - 2y^TXw + (Xw)^TXw - 2w_0\mathbf{1}(y - Xw) + w_0^2\mathbf{1}\mathbf{1}^T$
$\nabla_w J(w) = -2(y^TX)^T + 2X^TXw + 2w_0\mathbf{1}^TX \overset{set}{=} 0$; $w = (X^TX)^{-1}X^Ty$
$\nabla_{w_0} J(w_0) = -2\mathbf{1}^T(y - Xw) + 2nw_0 = -2n\bar{y} + 2nw_0 \overset{set}{=} 0$; $\mathbf{1}^TX = \sum x_i = 0$
$f = \sum_{i=1}^n a_i k_i^{1/2} = K^{1/2}a$; $g = K^{1/2}b$; $\langle f, g\rangle = f^TK^{-1}g = a^Tb$

Semi-parametric regression

Training set $\mathcal{D} = \{(x_i, y_i), 1 \leq i \leq n\}$, $x_i \in \mathbb{R}^d$ is a feature vector, and $y_i \in \mathbb{R}$;
$f(x) = \theta^Tx + g(x)$ where $\theta \in \mathbb{R}^d$ is a vector of parameters and $g : \mathbb{R}^d \mapsto \mathbb{R}$ belongs to a RKHS with kernel $k(.,.)$; $\min_{g\in\mathcal{H}} J(\theta, g)$
$J(\theta, g) = \sum_{i=1}^n \left(y_i - \theta^Tx_i - g(x_i)\right)^2 + \lambda\|g\|_{\mathcal{H}}^2$
$J(\theta, \sum_{i=1}^n \alpha_i k(x_i,\cdot)) = \| \underset{(n,1)}{Y} - \underset{(n,d)}{X}\underset{(d,1)}{\theta} - \underset{(n,n)}{K}\underset{(n,1)}{\alpha} \|^2 + \underset{(1,1)}{\lambda}\underset{(1,n)}{\alpha^T}\underset{(n,n)}{K}\underset{(n,1)}{\alpha}$
$\nabla_\alpha J = \frac{\partial}{\partial\alpha}\|K\alpha + X\theta - Y\|^2 + \lambda\frac{\partial}{\partial\alpha}\langle\alpha, K\alpha\rangle = 2K(K\alpha + X\theta - Y) + \lambda(IK\alpha + K^T\alpha)$
$= 2K(K\alpha + X\theta - Y) + 2\lambda K\alpha = 2K[(K+\lambda I)\alpha + X\theta - Y]$
p.d. K, X sym, $K = K^T$, $X = X^T$; $K = P\Lambda P^T$; $I = PP^T$; Λdiag $\gamma_1, .., \gamma_n$.
$\lambda > 0, \gamma_i > 0$, $K + \lambda I = P(\Lambda + \lambda I)P^T$ is inversible.
$\nabla_\alpha J \overset{set}{=} 0 \implies \alpha^\star = (K+\lambda I)^{-1}(Y - X\theta^\star)$
Also $K\alpha^\star + X\theta - Y = -\lambda\alpha^\star$. Let $G = (K+\lambda I)^{-1}$
$\nabla_\theta J = 2X^T(K\alpha + X\theta - Y) = 2X^T(-\lambda\alpha^\star) \overset{set}{=} 0$
$X^TG(Y - X\theta^\star) = 0$; $X^TGY = X^TGX\theta^\star$; $\theta^\star = (X^TGX)^{-1}X^TGY$

Optimal ordering training set is $\mathcal{D} = \{(x_i, y_i), 1 \leq i \leq n\}$, $x_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$
$I_- = \{i, 1 \leq i \leq n, y_i = -1\}$; $I_+ = \{i, 1 \leq i \leq n, y_i = +1\}$; $n_- + n_+ = n$
$J(f) = \frac{1}{n_-n_+}\sum_{i\in I_-}\sum_{j\in I_+}\left(1 - \left(f(x_j) - f(x_i)\right)\right) + \lambda||f||_H^2$
Notate $v = \text{span}[k(x_i,\cdot), 1 \leq i \leq n]$ $\mathcal{V} \subset \mathcal{H}$, RKHS of f. one can project $f \in \mathcal{H}$ onto $\mathcal{V}$ and write in an unique way $f = f_v + f_\perp$ with $f_v \in \mathcal{V}$ $\forall g \in \mathcal{V}$, $\langle f_\perp, g\rangle = 0$
$\|f\|_{\mathcal{H}}^2 = \|f_v + f_\perp\|_{\mathcal{H}}^2 = \langle f_v + f_\perp, f_v + f_\perp\rangle = \langle f_v, f_v\rangle + \langle f_\perp, f_\perp\rangle + 2\underbrace{\langle f_v, f_\perp\rangle}_{0} = \|f_v\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2$
$f(x_i) = \langle f, k(\cdot, x_i)\rangle = \langle f_v + f_\perp, k(\cdot, x_i)\rangle = \langle f_v, k(\cdot, x_i)\rangle + \underbrace{\langle f_\perp, k(\cdot, x_i)\rangle}_{0} = f_v(x_i)$
$J(f) - J(f_v) = \lambda\|f\|_{\mathcal{H}}^2 - \lambda\|f_v\|_{\mathcal{H}}^2 = \lambda\|f_\perp\|_{\mathcal{H}}^2 \geq 0$
the function $f(x) = \sum_{i=1}^n \alpha_i k(x_i, x)$ is the solution of $\min_{f\in\mathcal{H}} J(f)$
$\|f\|_{\mathcal{H}}^2 = \langle f, f\rangle_{\mathcal{H}} = \langle\sum_{i=1}^n \alpha_i k(\cdot, x_i), \sum_{j=1}^n \alpha_j k(\cdot, x_j)\rangle_{\mathcal{H}}$
$= \sum_{i,j=1}^n \alpha_i\alpha_j \underbrace{\langle k(\cdot, x_i), k(\cdot, x_j)\rangle_{\mathcal{H}}}_{k(x_i, x_j)} = \underset{(1,n)}{\alpha^T}\underset{(n,n)}{K}\underset{(n,1)}{\alpha}$
$f(x_i) = \sum_{j=1}^n \alpha_j k(x_i, x_j) = \sum_{j=1}^n \alpha_j[\underset{(n,n)}{K}]_{i,j} = K_i^T\alpha$
$J(\sum_{i=1}^n \alpha_i k(x_i,\cdot)) = \frac{1}{n_-n_+}\sum_{i\in I_-}\sum_{j\in I_+}[1 - (K_j\alpha - K_i\alpha)] + \lambda\alpha^TK\alpha$
$K_- = \frac{1}{n_-}\sum_{i\in I_-} K_i$; $K_+ = \frac{1}{n_+}\sum_{i\in I_+} K_i$;
$J(\alpha) = 1 - [K_+ - K_-]\alpha + \lambda\alpha^TK\alpha$
p.d. K, X is symmetric, $K = K^T$, $X = X^T$; $K = P\Lambda P^T$; $I = PP^T$; Λ is diagonal matrix with $\gamma_1, .., \gamma_n$.
$\lambda > 0, \gamma_i > 0$, $K + \lambda I = P(\Lambda + \lambda I)P^T$ is inversible.
$\nabla_\alpha J = -[K_+ - K_-] + \lambda(IK\alpha + K^T\alpha) = -[K_+ - K_-] + 2\lambda K\alpha \overset{set}{=} 0$
$\alpha^\star = (2\lambda K)^{-1}[K_+ - K_-]$
$x_- = \frac{1}{n_-}\sum_{i\in I_-} x_i$; $x_+ = \frac{1}{n_+}\sum_{i\in I_+} x_i$
$K = XX^T$, $\underset{(n,d)}{X} = (x_1^T, .., x_n^T)^T$; $K_i = Xx_i$;
$K_+ = \frac{1}{n_+}\sum_{j\in I_+} Xx_i = Xx_+$; $K_- = Xx_-$
$f(x) = \sum_{i=1}^n \alpha_i x_i^T x = [\sum_{i=1}^n \alpha_i x_i]^T x = (X^T\alpha)^T x$
$\alpha = (2\lambda XX^T)^{-1}X[x_+ - x_-]$
$f(x) = (2\lambda)^{-1}[x_+ - x_-]^T x$
$J(w) = 1 - (w^Tx_+ - w^Tx_-) + \lambda w^Tw$
$\nabla_w J(w) = -[x_+ - x_-] + 2\lambda w \overset{set}{=} 0$
$f(x) = (2\lambda)^{-1}[x_+ - x_-]^T x$