Kernel Methods Notes

Part I: Kernels basics, Kernel PCA & Ridge regression

- A **Kernel** is a function $K: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ such that there exists a Hilbert space $\mathcal{H}$ and mapping $\phi: \mathcal{X} \rightarrow \mathcal{H}$ where $K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$

- A **Hilbert Space** is a vector space on which an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$ is defined, this having the following properties:
  - $\langle a f_1 + b f_2, g \rangle_{\mathcal{H}} = a \langle f_1, g \rangle_{\mathcal{H}} + b \langle f_2, g \rangle_{\mathcal{H}}$ (linear
  - $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$ (Symmetr
  - $\langle f, f \rangle_{\mathcal{H}} \geq 0$, $= 0$ only when $f = 0$

- All kernels $K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$ are positive definite functions:
  - given arbitrary $a_1, \ldots, a_n \in \mathbb{R}$, $x_1, \ldots, x_n \in \mathcal{X}$

$$\sum_i \sum_j a_i a_j K(x_i, x_j) = \sum_i \sum_j \langle a_i \phi(x_i), a_j \phi(x_j) \rangle_{\mathcal{H}}$$

$$= \left\langle \sum_i a_i \phi(x_i), \sum_j a_j \phi(x_j) \right\rangle_{\mathcal{H}}$$

$$= \left\| \sum_i a_i \phi(x_i) \right\|_{\mathcal{H}}^2 \geq 0 \quad \dashv$$

- It turns out that the opposite direction holds as well: all positive definite functions are kernels!

- Therefore, all sums of kernels $K(x, x') = K_1(x, x') + K_2$ are kernels: for arbitrary $a_1, \ldots, a_n \in \mathbb{R}$, $x_1, \ldots, x_n \in \mathcal{X}$

$$\sum_i \sum_j a_i a_j K(x_i, x_j) = \sum_i \sum_j a_i a_j \left(K_1(x_i, x_j) + K_2(x_i, x_j) \right)$$

$$= \left\| \sum_i a_i \phi_1(x_i) \right\|_{\mathcal{H}_1}^2 + \left\| \sum_i a_i \phi_2(x_i) \right\|_{\mathcal{H}_2}^2 \geq$$

$\Rightarrow$ positive-definite $\therefore$ a kernel

- All products of kernels $K(x, x') = K_1(x, x') K_2(x, x')$ are kernels:

$$K_1(x, x') K_2(x, x') = \langle \phi_1(x), \phi_1(x') \rangle_{\mathcal{H}_1} \langle \phi_2(x), \phi_2(x') \rangle_{\mathcal{H}_2}$$

$$= \phi_1(x')^T \phi_1(x^*) \phi_2(x)^T \phi_2(x')$$

can always take trace of a scalar

$$= \phi_1(x')^T \phi_1(x^*) \, \mathrm{Trace}\left[\phi_2(x') \phi_2(x)^T \right]$$

can move a scalar into a trace

$$= \mathrm{Tr}\left[\underbrace{\phi_2(x') \phi_1(x')^T}_{A^T} \underbrace{\phi_1(x^*) \phi_2(x)^T}_{B} \right]$$

Frobenius product

$$= \mathrm{Tr}\left[A^T B \right]$$

$$= \mathrm{vec}(A)^T \mathrm{vec}(B)$$

$$= \left\langle \mathrm{vec}\left(\phi_1(x') \phi_2(x')^T \right), \mathrm{vec}\left(\phi_1(x^*) \phi_2(x)^T \right) \right\rangle_{\mathcal{H}}$$

$$= \langle \psi(x'), \psi(x) \rangle_{\mathcal{H}} = K(x, x') \checkmark$$

- Every Kernel is associated with a unique RKHS $\mathcal{H}$, which has the following properties:
    - $\forall x \in \mathcal{X}$, $K(\cdot, x) \in \mathcal{H}$
    - $\forall x \in \mathcal{X}$, $\forall f \in \mathcal{H}$, $\underline{\langle f, K(\cdot, x)\rangle = f(x)}$

reproducing property

- Ex. RKHS defined by a Fourier Series

consider the space of all periodic functions on $[-\pi, \pi]$:

$$f(x) = \sum_{l=-\infty}^{\infty} \hat{f}_l e^{ilx}$$

We can then define the ∞-D vector space spanned by the orthonormal basis $\{e^{ilx}\}_{l=-\infty}^{\infty}$, $x \in \mathbb{R}$ together with the standard L2 dot product $\langle \cdot, \cdot \rangle$, to give us a Hilbert space $\mathcal{H}$, where $\langle f, g\rangle_{L_2} = \sum_{l=-\infty}^{\infty} \hat{f}_l \bar{\hat{g}}_l$.

Is $\mathcal{H}$ an RKHS? Let $K(x,y) = K(x-y)$

$$= \sum_{l=-\infty}^{\infty} \hat{K}_l e^{il(x-y)}$$

$$= \sum_{l=-\infty}^{\infty} \hat{K}_l e^{ily} e^{-il}$$

we check for the reproducing property:

$$\langle f, K(\cdot, x)\rangle_{L_2} = \sum_{l=-\infty}^{\infty} \hat{f}_l \overline{\hat{K}_l e^{ily}}$$

$$= \sum_{l=-\infty}^{\infty} \hat{K}_l \hat{f}_l e^{ilx} \neq f(x)$$

Given this kernel, what is the dot product of the associated RKHS?

So $\mathcal{H}$ is not an RKHS. But we can easily modify it so that it is: $\mathcal{H}^*$ with $\langle f, g\rangle_{\mathcal{H}^*} = \sum_{l=-\infty}^{\infty} \frac{\hat{f}_l \bar{\hat{g}}_l}{\hat{K}_l}$

Now, $\langle f, K(\cdot,x)\rangle_{\mathcal{H}^*} = \sum_{l=-\infty}^{\infty} \frac{\hat{f}_l \cancel{\hat{K}_l} e^{ilx}}{\cancel{\hat{K}_l}} = \sum_{l=-\infty}^{\infty} \hat{f}_l e^{ilx} = f(x)$

$$\langle K(\cdot,x), K(\cdot y)\rangle_{\mathcal{H}^*} = \sum_{l=-\infty}^{\infty} \frac{\hat{K}_l e^{-ilx} \cancel{\hat{K}_l} e^{ily}}{\cancel{\hat{K}_l}} = \sum_{l=-\infty}^{\infty} \hat{K}_l e^{il(y-x)} = K(y-x$$

Importantly, $\langle f, f\rangle_{\mathcal{H}^*} = \|f\|^2_{\mathcal{H}^*} = \sum_{l=-\infty}^{\infty} \frac{|\hat{f}_l|^2}{\hat{K}_l}$, so the kernel enforces smoothness since any $f \in \mathcal{H}^*$ must have $\hat{f}_l$ that decay faster than $\hat{K}_l$ for $\|f\|^2_{\mathcal{H}^*} < \infty$, i.e. $f(\cdot)$ must be at least as smooth (low amplitudes at higher frequencies) as $K(\cdot)$.

- Kernel PCA: just like normal PCA but performed in feature space, via the reproducing property:

$$f^* = \underset{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}}=1}}{\operatorname{argmax}}$$ variance of data projected into $\mathcal{H}$ via feature map $\phi(x) = K(x,\cdot)$ along unit vector f

~~$= \underset{\|f\|_{\mathcal{H}}=1}{\operatorname{argmax}} \langle f, C \rangle_{\mathcal{H}}$, $C = \frac{1}{N}\sum_i \left(\phi(x_i) - \left(\frac{1}{N}\sum_j \phi(x_j)\right)\right)^2$~~

$$= \underset{\|f\|_{\mathcal{H}}=1}{\operatorname{argmax}} \frac{1}{N}\sum_{i=1}^{N} \langle f, \phi(x_i) - \bar{\phi}\rangle^2_{\mathcal{H}}$$

~~$= \frac{1}{N}\sum_i \tilde{\phi}(x_i)^2$~~

$$\bar{\phi} = \frac{1}{N}\sum_i \phi(x_i)$$

$$\tilde{\phi}(x_i) = \phi(x_i) - \bar{\phi}$$

$$\frac{1}{N}\sum_i \langle f, \tilde{\phi}(x_i)\rangle \langle f, \tilde{\phi}(x_i)\rangle$$

$$\frac{1}{N}\sum_i \langle f, \tilde{\phi}(x_i) \otimes \tilde{\phi}(x_i)\, f\rangle$$

$$= \underset{\|f\|_{\mathcal{H}}}{\operatorname{argmax}} \langle f, Cf\rangle, \quad C = \frac{1}{N}\sum_{i=1}^{N} \tilde{\phi}(x_i) \otimes \tilde{\phi}(x_i)$$

$$\Rightarrow \frac{\partial}{\partial f}\left[\langle f, Cf\rangle_{\mathcal{H}} + \lambda\left(\langle f,f\rangle_{\mathcal{H}} - 1\right)\right] = 0$$

$$\Leftrightarrow \quad Cf = \lambda f$$

$\Rightarrow f^{*}$ = Largest e-vector C

↳ but this requires computing C, which pt lives in $\mathbb{R}^{\infty \times \infty}$

→ How can we avoid feature space?

$\Rightarrow$ We can always express f as a linear combination of data points, without loss of generality, since any dimensions orthogonal to the space spanned by $\{\tilde{\phi}(x_i)\}_{i=1}^{n}$ will disappear in the first line $\langle f, \tilde{\phi}(x_i)\rangle_{\mathcal{H}}$, thus rendering them irrelevant to the optimization:

$$f = \sum_{i=1}^{N} \alpha_i \tilde{\phi}(x_i) \qquad \tilde{K}(x,x') = \langle \tilde{\phi}(x), \tilde{\phi}(x')\rangle_{\mathcal{H}}$$

$$\Leftrightarrow f(\cdot) = \sum_{i=1}^{N} \alpha_i \tilde{K}(x_i, \cdot) \qquad (\text{by reproducing property})$$

Thus we need only solve for the α's:

$$Cf = \frac{1}{N}\sum_{i=1}^{N} \tilde{\phi}(x_i) \sum_{j=1}^{N} \alpha_j \langle \tilde{\phi}(x_i), \tilde{\phi}(x_j)\rangle_{\mathcal{H}} \qquad \alpha_j \tilde{K}(x$$

$$= \frac{1}{N}\sum_{i=1}^{N} \tilde{\phi}(x_i) \sum_{j=1}^{N} \alpha_j \tilde{K}(x_i, x_j) \Rightarrow \langle \tilde{\phi}(x_k), Cf\rangle = \frac{1}{N}\sum_{i} K(x_k, x_i)\sum_j$$

$$\langle \tilde{\phi}(x_k), \lambda f\rangle_{\mathcal{H}} = \lambda \sum_i \alpha_i \tilde{K}(x_k, x_i) \qquad \Rightarrow \frac{1}{N}\tilde{K}\tilde{K}\alpha = \lambda \tilde{K}\alpha$$

where $\tilde{K}_{ij} = \tilde{K}(x_i, x_j)$. Since this matrix is symmetri and positive semidefinite, its inverse exists, so we get the following eigenvalue equation:

$$\tilde{K}\alpha = N\lambda\alpha$$

So we can solve for α by constructing the Gram matrix $\tilde{K}$ and solving the e-value equation, giving us the directions f of greatest variance without having to work at all in feature space. (ie. biggest e

Importantly, f is a function, so kernel PCA, as opposed to regular PCA, can give us non-linear principal subspaces rather than just hyperplanes (depending on the kernel).

- Kernel Ridge Regression: ridge regression in feature spa

$$y = w^T \phi(x) + \epsilon, \quad \phi(x) \in \mathcal{H}$$

$$\Rightarrow w^* = \underset{w \in \mathcal{H}}{\arg\min} \left[\sum_{i=1}^{N} \left(y_i - \langle w, \phi(x_i) \rangle_{\mathcal{H}} \right)^2 + \lambda \|w\|_{\mathcal{H}}^2 \right]$$

$$= \underset{w \in \mathcal{H}}{\arg\min} \left[\|Y - X^T w\|_{\mathcal{H}}^2 + \lambda \|w\|_{\mathcal{H}}^2 \right], \quad X = \begin{bmatrix} \phi(x_1) & \cdots & \phi(\end{bmatrix}$$

$$= \underset{w \in \mathcal{H}}{\arg\min} \left[Y^T Y - 2 Y^T X^T w + w^T (XX^T + \lambda I) w \right]$$

completing the square

$$= \underset{w \in \mathcal{H}}{\arg\min} \left[Y^T Y + \| (XX^T + \lambda I)^{\frac{1}{2}} w - (XX^T + \lambda I)^{-\frac{1}{2}} X Y \|_{\mathcal{H}}^2 - \| (XX^T + \lambda I)^{-\frac{1}{2}} X \right.$$

(we could've done this by taking derivatives, but derivatives don't necessarily exist for discrete x_i, y_i,

$$= (XX^T + \lambda I)^{-1} XY$$

To avoid having to do anything in feature space, we rewrite this in terms of the Gram matrix $K = X^T X$:

$K_{ij} = K(x_i, x_j)$

① Via SVD:

$$\underset{D\times N}{X} = \underset{D\times D}{\left[\tilde{U}\right]} \underset{D\times N}{\begin{bmatrix}\tilde{S}\\ 0\end{bmatrix}} \underset{N\times N}{\left[\tilde{V}\right]}$$

(orthogonal) (diagonal) (orthogonal)

Let
$$U = \tilde{U} \qquad D \times D$$
$$S = \begin{bmatrix}\tilde{S} & 0\\ 0 & 0\end{bmatrix} \qquad D \times D$$
$$V = \begin{bmatrix}\tilde{V} & 0\end{bmatrix} \qquad N \times D$$

such that $X = USV^T$

we then have:

$$w^* = (US^2U^T + \lambda I)^{-1} USV^T Y$$
$$= U(S^2 + \lambda I)^{-1} U^T USV^T Y$$
$$= US(S^2 + \lambda I)^{-1} V^T Y$$
$$= USV^T V(S^2 + \lambda I)^{-1} V^T Y$$
$$= USV^T (V^T S^2 V + \lambda I)^{-1} Y$$
$$= X(X^T X + \lambda I)^{-1} Y$$
$$= X(K + \lambda I)^{-1} Y$$

(can do this since S is diagonal and square (hence the change from the usual SVD))

② Via Woodbury Identity:

$$w^* = (XX^T + \lambda I)^{-1} XY$$

$$= (\lambda^{-1} I - \lambda^{-1} X(\lambda^{-1} X^T X + I)^{-1} X^T \lambda^{-1}) XY$$

$$= [\lambda^{-1} X - \lambda^{-1} X(\lambda^{-1} X^T X + I)^{-1} \lambda^{-1} X^T X] Y$$

$$= [\lambda^{-1} X + \lambda^{-1} X(\lambda^{-1} X^T X + I)^{-1} - \lambda^{-1} X(\lambda^{-1} X^T X + I)^{-1} - \lambda^{-1} X(\lambda^{-1} X^T X + I)^{-1} \lambda^{-1} X^T X] Y$$

$$= [\cancel{\lambda^{-1} X} + \lambda^{-1} X(\lambda^{-1} X^T X + I)^{-1} \cancel{- \lambda^{-1} X(\lambda^{-1} X^T X + I)^{-1}(\lambda^{-1} X^T X + I)}]$$

$$= \lambda^{-1} X(\lambda^{-1} X^T X + I)^{-1} Y$$

$$= \underline{X(\underbrace{X^T X}_{K} + \lambda I)^{-1} Y}$$

Thus, Our optimal weights are a weighted sum of the data points: $w^* = \sum_i \alpha_i \phi(x_i)$, $\underline{\alpha} = (K + \lambda I)^{-1} Y$

Note that w^* is a function in $\mathcal{H}$, such that its smoothne is constrained by the kernel since $\|w^*\|_{\mathcal{H}}^2 < \infty$. The larger our regularizing ~~the~~ constant λ, the smoother our resulti regression function $\langle w^*, \phi(x) \rangle_{\mathcal{H}} \doteq w^*(x)$ will be

Part II : MMD, HSIC, COCO

- Just like the "kernel trick" allows us to express functions in terms of feature space:

$$f(x) = \langle f, K(x,\cdot)\rangle_{\mathcal{H}}$$

the "mean trick" allows us to do the same with expectations:

$$\mathbb{E}_{x\sim p}\left[f(x)\right] = \langle \mu_p, f\rangle_{\mathcal{H}}$$

By the reproducing property,

[probability feature map] (analog to feature map $\phi(x)$)

$$\mu_p(x) = \langle \mu_p, K(x,\cdot)\rangle_{\mathcal{H}} = \mathbb{E}_{x\sim p}\left[K(x,X)\right]$$

so we can estimate it empirically just like ~~[illegible]~~ usual!

$$\hat{\mu}_p(a) = \frac{1}{N}\sum_i \langle K(x_i,\cdot), K(a,\cdot)\rangle_{\mathcal{H}}$$

$$= \frac{1}{N}\sum_i K(x_i,a)$$

mean embedding

We can prove that μ_p exists in feature space (i.e. prove that the "mean trick" works) via the Riesz representation theorem:

any bounded linear operator $A: \mathcal{H}\to\mathbb{R}$, i.e.

$$|Af| \le \lambda_A \|f\|_{\mathcal{H}} \quad \forall f\in\mathcal{H}$$

can be expressed as

$$Af = \langle f, g_A\rangle_{\mathcal{H}} \quad \text{for some } g\in\mathcal{H}$$

Thus, if we prove that the expectation operator $\mathbb{E}_p$ is bounded, then $\mu_p \in \mathcal{H}$:

assuming $\mathbb{E}_p\left[\sqrt{k(x,x)}\right] < \infty$

$$\left|\mathbb{E}_p f(x)\right| \overset{\text{Jensen}}{\leq} \mathbb{E}_p\left|f(x)\right| = \mathbb{E}_p\left[\langle f, k(x,\cdot)\rangle_{\mathcal{H}}\right] \overset{\text{Cauchy-Schwartz}}{\leq} \mathbb{E}_p\left[\|k(x,\cdot)\|_{\mathcal{H}}\|f\|_{\mathcal{H}}\right]$$

$$= \mathbb{E}_p\left[\sqrt{\langle k(x,\cdot), k(x,\cdot)\rangle}\right]\|f\|_{\mathcal{H}}$$

$$= \mathbb{E}_p\left[\sqrt{k(x,x)}\right]\|f\|_{\mathcal{H}}$$

$$= \lambda \|f\|_{\mathcal{H}}$$

$$\therefore \underline{\mathbb{E}_p f(x) = \langle f, \mu_p\rangle_{\mathcal{H}}, \quad \mu_p \in \mathcal{H}}$$

- To compare means, we use the <u>Max. Mean Discrepancy</u>

$$MMD(P, Q; \mathcal{H}) = \sup_{f\in\mathcal{H}}\left[\mathbb{E}_p f(X) - \mathbb{E}_Q f(Y)\right] \quad \text{for } X\sim P,\ Y\sim Q$$

where $\mathcal{H}$ is the unit ball in RKHS, i.e. $\|f\|_{\mathcal{H}} \leq 1$

$$= \sup_{f\in\mathcal{H}} \langle f, \mu_p - \mu_Q\rangle_{\mathcal{H}}$$

$$= \|\mu_p - \mu_Q\|_{\mathcal{H}}$$

which we estimate empirically by:

$$MMD^2 = \langle \mu_p, \mu_p\rangle_{\mathcal{H}} + \langle \mu_Q, \mu_Q\rangle_{\mathcal{H}} - 2\langle \mu_p, \mu_Q\rangle_{\mathcal{H}}$$

$$= \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i} k(x_i, x_j) + \frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{j\neq i} k(y_i, y_j)$$

$$- \frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} k(x_i, y_j)$$

exclude repeating the same data point to ensure estimator remains unbiased one-to-one mapping b/w P and μ

probability distribution

- MMD $= 0$ iff $P = Q$ whenever $\mathcal{H}$ is a <u>characteristic RKHS</u> with <u>characteristic kernel</u> $K(\cdot,\cdot)$.

For periodic on $[-\pi, \pi]$ and translation-invariant $K(\cdot,\cdot)$, K is characteristic iff $\hat{k}_\ell \neq 0 \ \forall \ell$

$$\mu_P(z) = \langle \mu_P, K(z,\cdot) \rangle$$

$$= \mathbb{E}_{x\sim P} K(z, x)$$

$$= \mathbb{E}_{x\sim P} K(z - x)$$

$$= \int_{-\pi}^{\pi} K(z-x)\, dP(x)$$

$$\Longrightarrow \hat{\mu}_{P,\ell} = \int \mu_P(z)\, e^{-i\ell z}\, dz$$

$$= \int \int_{-\pi}^{\pi} K(z-x)\, e^{-i\ell z}\, dP(x)\, dz$$

$$= \int \int_{-\pi}^{\pi} K(v)\, e^{-i\ell(v+x)}\, dP(x)\, dv$$

If $\hat{k}_\ell \neq 0\ \forall \ell$, then MMD $= 0$ iff $P = Q$

①

② MMD penalizes some freqs more than others, depending on kernel smoothness

$$= \int_{-\pi}^{\pi} K(v)\, e^{-i\ell v} \int e^{-i\ell x}\, dP(x)$$

Recalling that for periodic f and translation invariant K, $\|f\|_{\mathcal{H}}^2 = \sum_\ell \frac{|\hat{f}_\ell|^2}{\hat{k}_\ell}$

$$= \hat{k}_\ell\, \hat{\varphi}_{P,\ell}$$

Fourier coefs of the probability distribution P density function

$$\Longrightarrow \text{MMD}^2 = \|\mu_P - \mu_Q\|_{\mathcal{H}}^2 = \sum_{\ell=-\infty}^{\infty} \frac{|\hat{k}_\ell \hat{\varphi}_{P,\ell} - \hat{k}_\ell \hat{\varphi}_{Q,\ell}|^2}{\hat{k}_\ell} = \sum_{\ell=-\infty}^{\infty} \hat{k}_\ell\, |\hat{\varphi}_{P,\ell} - \hat{\varphi}_{Q,\ell}|^2$$

For any function on $\mathbb{R}^D$ we can show the following, via Bochner's theorem:

$$MMD^2 = \int_{\mathbb{R}^D} \left|\Phi_P(\ell) - \Phi_Q(\ell)\right|^2 d\Lambda(\ell)$$

which says the same thing.

Fourier transform of K

$\Rightarrow K$ characteristic iff $\mathrm{supp}(\Lambda) = \mathbb{R}^D$

↳ support = set of der that are <u>not</u> mapped to 0

$\Rightarrow$ any continuous K with Fourier transform Λ s.t. $\mathrm{supp}(\Lambda) = \mathbb{R}^D$ is characteristic

- For hypothesis testing, use $\widehat{MMD}^2 = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} K(x_i, x_j) + K(y_i, y_j) - K(x_i,$
$- K(y_i,$

(lower variance since dropped some terms, but still unbiased)

$H_1 : P \neq Q$

$\sqrt{N}\left(\widehat{MMD}^2 - MMD^2\right) \sim \mathcal{N}(0, \sigma_u^2)$

↳ assymptotically normal

$H_0 : P = Q$

$N \cdot MMD \sim \sum_{\ell=1}^{\infty} \lambda_\ell (z_\ell^2 - 2)$, $z_\ell \overset{i.i.d.}{\sim} \mathcal{N}(0, 2)$

↳ degenerate U-statistic, so need to estimate via e.g. permutation

Pearson moment matching

- Just like we can show that $\mathbb{E}_p[f(x)]$ can be expressed in feature space via the mean embedding, we can show that the cross-covariance

$$\mathbb{E}_{P_{xy}}[f(x,y)] = \mathbb{E}_{P_{xy}}\left[\langle \phi(x) \otimes \psi(y), f \rangle_{\mathcal{F}\times G}\right]$$

$$= \langle \tilde{C}_{xy}, f \rangle_{\mathcal{F}\times G}$$

for feature maps $\phi: \mathcal{X} \to \mathcal{F}$, $\psi: \mathcal{Y} \to G$, and $\underset{\sim}{\text{Hilbert-Schmidt operators}}$ $f, \tilde{C}_{xy} \in \mathcal{F}\times G$

We can again show $\tilde{C}_{xy}$ exists via Riesz representer theorem:

$$\left|\mathbb{E}_{P_{xy}}[f(x,y)]\right| \overset{\text{Jensen}}{\leq} \mathbb{E}_{P_{xy}}|f(x,y)| = \mathbb{E}_{P_{xy}}\left|\langle f, \phi(x) \otimes \psi(y) \rangle\right|$$

$$\overset{\text{Cauchy-Schwartz}}{\leq} \mathbb{E}_{P_{xy}}\left[\|f\|_{\mathcal{F}\times G} \|\phi(x) \otimes \psi(y)\|_{\mathcal{F}}\right.$$

$$\overset{\text{see HS operator norm}}{=} \|f\|_{\mathcal{F}\times G}\, \mathbb{E}_{P_{xy}}\left[\|\phi(x)\|_{\mathcal{F}} \|\psi\right.$$

$$= \|f\|_{\mathcal{F}\times G}\, \mathbb{E}_{P_{xy}}\left[\sqrt{k(x,x)\,\ell(y,y)}\right.$$

$$< \infty$$

Hence by Riesz, the bounded linear operator $\mathbb{E}_{P_{xy}}[f(x,y)]$ can be expressed as $\langle \tilde{C}_{xy}, f \rangle_{\mathcal{F}\times G}$

We can see that $\tilde{C}_{xy}$ gives us the cross covariance b/w variables in feature space by considering

$$\mathbb{E}_{P_{xy}}[k(x,X)\,\ell(y,Y)] = \mathbb{E}_{P_{xy}}\left[\langle k(x,\cdot), k(X,\cdot)\rangle_{\mathcal{F}} \langle \ell(y,\cdot), \ell(Y,\cdot)\rangle_{G}\right]$$

$$= \mathbb{E}_{P_{xy}}\left[\langle k(x,\cdot), k(X,\cdot) \otimes \ell(Y,\cdot)\,\ell(y,\cdot)\rangle_{\mathcal{F}}\right]$$

$$= \left\langle k(x,\cdot), \mathbb{E}_{P_{xy}}[\phi(X) \otimes \psi(Y)]\, l(y,\cdot) \right\rangle_{\mathcal{F}}$$

$$\overset{\text{see HS}\langle\rangle \text{ rank 1 operators}}{=} \left\langle k(x,\cdot) \otimes l(y,\cdot), \mathbb{E}_{P_{xy}} \phi(X) \otimes \psi(Y) \right\rangle_{\mathcal{F}\times\mathcal{G}}$$

$$= \left\langle k(x,\cdot) \otimes l(y,\cdot), \tilde{C}_{xy} \right\rangle_{\mathcal{F}\times\mathcal{G}}$$

where $k(x,\cdot)$, $l(y,\cdot)$ are the feature maps of two variables $x \in \mathcal{X}$, $y \in \mathcal{Y}$.

The centered cross-covariance is then

$$C_{xy} = \mathbb{E}_{P_{xy}}\left[\phi(x) \otimes \psi(y)\right] - \mathbb{E}_{P_x}[\phi(x)] \otimes \mathbb{E}_{P_y}[\psi($$

$$= \tilde{C}_{xy} - \mu_x \otimes \mu_y$$

Which we can estimate empirically by:

$$\hat{C}_{xy} := \frac{1}{N}\sum_{i=1}^{N} \phi(x_i) \otimes \psi(y_i) - \hat{\mu}_x \otimes \hat{\mu}_y, \quad \hat{\mu}_x = \frac{1}{N}\sum_i \phi(x_i)$$

We can write this in matrix notation using the centering matrix $H = I_{N\times N} - \frac{1}{N}\mathbb{1}_{N\times N}$:

$$\boxed{\hat{C}_{xy} = \frac{1}{N} X H Y^T} = \frac{1}{N} \tilde{X} Y^T = \frac{1}{N}\sum_i (\phi(x_i) - \hat{\mu}_x) \otimes \psi(y_i)$$

$$= \frac{1}{N}\sum_i \phi(x_i) \otimes \psi(y_i) - \hat{\mu}_x \otimes \frac{1}{N}\sum$$

$$X = \left[\phi(x_1) \cdots \phi(x_N)\right], \quad Y = \left[\psi(y_1) \cdots \psi(y_N)\right]$$

$$= \frac{1}{N}\sum_i \phi(x_i) \otimes \psi(y_i) - \hat{\mu}_x \hat{\mu}_y$$

- Hilbert-Schmidt operators are like matrices in $\mathcal{F} \times G$, Let $L, M \in \mathcal{F} \times G$ s.t. $L, M : G \to \mathcal{F}$

Suppose $\{f_i\}_{i \in I}$, $\{g_j\}_{j \in J}$ are bases for $\mathcal{F}$ and G, respectively

We then define the HS norm:

$$\|L\|_{HS}^2 := \sum_{j \in J} \|L g_j\|_{\mathcal{F}}^2$$

$$= \sum_{i \in I} \sum_{j \in J} \left| \langle L g_j , f_i \rangle \right|^2$$

The HS inner product is then:

$$\langle L, M \rangle_{HS} = \sum_{j \in J} \langle L g_j , M g_j \rangle_{\mathcal{F}}$$

$$= \sum_{i \in I} \sum_{j \in J} \langle L g_j , f_i \rangle_{\mathcal{F}} \langle M g_j , f_i \rangle_{\mathcal{F}}$$

For a rank 1 operator $a \otimes b$, we have:

$$\|a \otimes b\|_{HS}^2 = \sum_{j \in J} \|a \otimes b\, g_j\|_{\mathcal{F}}^2 \quad a \in \mathcal{F},\ b \in G$$

$$= \sum_{j \in J} \|a \langle b, g_j \rangle_G\|_{\mathcal{F}}^2 = \|a\|_{\mathcal{F}}^2 \sum_{j \in J} \langle b, g_j \rangle_G^2$$

$$= \|a\|_{\mathcal{F}}^2 \|b\|_G^2$$

$$\langle L, a\otimes b\rangle_{HS} = \sum_{j\in J} \langle L g_j, a\otimes b\, g_j\rangle_{F}$$

$$= \sum_{j\in J} \langle L g_j, a\rangle_{F} \langle b, g_j\rangle_{G}$$

$$= \left\langle \sum_{j\in J} L\langle b, g_j\rangle g_j, a\right\rangle_{F}$$

$$= \langle Lb, a\rangle_{F}$$

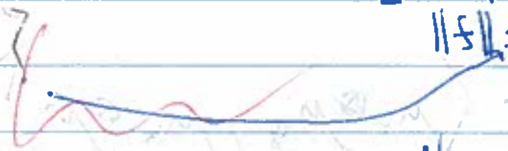
$$\langle u\otimes v, a\otimes b\rangle_{HS} = \langle (u\otimes v)b, a\rangle_{F}$$

$$= \langle v, b\rangle_{G} \langle u, a\rangle_{F}$$

- Given $X \sim P_x$, $Y \sim P_y$ we can test for independence via Hilbert-Schmidt Independence Criterion

$$HSIC(P_{xy}, P_x P_y) = MMD^2(P_{xy}, P_x P_y ; \mathcal{F} \times \mathcal{G})$$

$$= \sup_{\|f\| \le 1} \langle f, \mu_{P_{xy}} - \mu_{P_x} \mu_{P_y} \rangle^2_{\mathcal{F} \times \mathcal{G}}$$



$$= \| \mu_{P_{xy}} - \mu_{P_x} \otimes \mu_{P_y} \|^2_{\mathcal{F} \times \mathcal{G}}$$

where $f(x,y)$ is a "matrix" in $\mathcal{F} \times \mathcal{G}$ Hilbert ~~Schmidt~~ space

and

$$\mu_{P_{xy}} = \mathbb{E}_{P_{xy}} \phi(X) \otimes \psi(Y) = \tilde{C}_{xy} \in \mathcal{F} \times \mathcal{G}$$

$$\Rightarrow \mu_{P_{xy}}(x,y) = \mathbb{E}_{P_{xy}} \left[k(x,X) \, l(y,Y) \right]$$ (see above part on $\tilde{C}_{xy}$

$$\mu_{P_x P_y} = \mathbb{E}_{P_x} \mathbb{E}_{P_y} \phi(X) \otimes \psi(Y)$$

$$= \mathbb{E}_{P_x} \phi(X) \otimes \mathbb{E}_{P_y} \psi(Y)$$

$$= \mu_x \otimes \mu_y$$

Thus, we can estimate it empirically via:

$$HSIC \simeq \| \hat{C}_{xy} \|^2 = \| X H Y^T \|^2$$

$$= Tr[Y H X^T X H Y^T]$$

$$= Tr[X^T X H Y^T Y H]$$

$$= Tr[K H L H]$$

But note this estimate is biased, since we are including in our product the kernels evaluated at single points & ($K_{ii} = k(x_i, x_i)$, $L_{ii} = \ell(y_i, y_i)$):

$$\mathrm{HSIC} = \|C_{xy}\|^2 = \|\tilde{C}_{xy} - \mu_x \otimes \mu_y\|^2$$

$$= \langle \tilde{C}_{xy}, \tilde{C}_{xy} \rangle_{\mathcal{F}\times\mathcal{G}} + \langle \mu_x \otimes \mu_y, \mu_x \otimes \mu_y \rangle_{\mathcal{F}\times\mathcal{G}}$$

$$- 2 \langle \tilde{C}_{xy}, \mu_x \otimes \mu_y \rangle_{\mathcal{F}\times\mathcal{G}}$$

$$= \langle \tilde{C}_{xy}, \tilde{C}_{xy} \rangle_{\mathcal{F}\times\mathcal{G}} + \langle \mu_x, \mu_x \rangle_{\mathcal{F}} \langle \mu_y, \mu_y \rangle_{\mathcal{G}}$$

$$- 2 \mathbb{E}_{x,y\sim P_{xy}} \langle \phi(x) \otimes \psi(y), \mu_x \otimes \mu_y \rangle_{\mathcal{F}\times\mathcal{G}}$$

(these should really all be functions $k(x,\cdot)$, $\ell(y,\cdot)$)

$$= \mathbb{E}_{x,y\sim P_{xy}} \mathbb{E}_{x',y'\sim P_{xy}} \langle \phi(x) \otimes \psi(y), \phi(x') \otimes \psi(y') \rangle_{\mathcal{F}\times\mathcal{G}}$$

$$+ \mathbb{E}_{x\sim P_x} \mathbb{E}_{x'\sim P_x} \langle \phi(x), \phi(x') \rangle_{\mathcal{F}} \mathbb{E}_{y\sim P_y} \mathbb{E}_{y'\sim P_y} \langle \psi(y), \psi(y$$

$$- 2 \mathbb{E}_{x,y\sim P_{xy}} \mathbb{E}_{x'\sim P_x} \mathbb{E}_{y'\sim P_y} \langle \phi(x), \phi(x') \rangle_{\mathcal{F}} \langle \psi(y), \psi($$

$$= \mathbb{E}_{x,y\sim P_{xy}} \mathbb{E}_{x',y'\sim P_{xy}} k(x,x') \ell(y,y') + \mathbb{E}_{x\sim P_x} \mathbb{E}_{x'\sim P_x} k(x,x') \mathbb{E}_{y\sim P_y} \mathbb{E}_{y'\sim P_y} \ell(y,$$

$$- 2 \mathbb{E}_{x,y\sim P_{xy}} \mathbb{E}_{x'\sim P_x} \mathbb{E}_{y'\sim P_y} k(x,x') \ell(y,y')$$

Thus, in term 1, x and x' (and y and y') should be independent, but not x,y; in term 2 x, x', y, y' should all be independent, in term 3 (x, x') and (y, y') and (x', y') should be indepen

- We can also test for dependence by directly computing the cross-covariance using the operator C_{xy}. This is called the <u>constrained covariance</u>:

$$\mathrm{COCO}(P_{xy}; F, G) = \sup_{\|f\|_F=1,\ \|g\|_G=1} \mathrm{cov}\left[f(x), g(y)\right]$$

$$= \underset{\sim}{\sup}\ \mathbb{E}_{xy}\left[f(x) \otimes g(y)\right] - \mathbb{E}_x f(x) \otimes \mathbb{E}_y g($$

$$= \sup\ \mathbb{E}_{xy}\left[\langle f, \phi(x)\rangle_F \otimes \langle g, \psi(y)\rangle_G\right] - \mathbb{E}_x \langle f, \phi(x)\rangle_F \otimes \mathbb{E}$$

$$= \underset{\sim}{\sup} \left\langle f \otimes g, \mathbb{E}_{xy}\, \phi(x) \otimes \psi(y)\right\rangle_{F \times G} - \langle f, \mu_x\rangle_F \langle g, \mu_y\rangle_G$$

$$= \underset{\sim}{\sup} \left\langle f \otimes g, \tilde{C}_{xy}\right\rangle_{F\times G} - \left\langle f \otimes g, \mu_x \otimes$$

$$= \underset{\sim}{\sup} \left\langle f \otimes g, \tilde{C}_{xy} - \mu_x \otimes \mu_y\right\rangle_{F \times G}$$

$$= \sup_{\|f\|_F=1,\ \|g\|_G=1} \left\langle f, C_{xy}\, g\right\rangle$$

which we estimate empirically using $\hat{C}_{xy}$, $\hat{\mu}_x$, $\hat{\mu}_y$. Noting again that computing $\hat{\mu}_x$, $\hat{\mu}_y$ requires dot products between f and $\{x_i\}$ and g and $\{y_i\}$, any components of f and g orthogonal to the data space disappear and are thus irrelevant. We can therefore express f and g as follows, without loss of generality:

$$f = \sum_i \alpha_i \tilde{\phi}(x_i) = \tilde{X}\cancel{H}\alpha \qquad g = \sum_i \beta_i \tilde{\psi}(y_i) = Y\cancel{H}\beta$$

where $\bar{\phi}(x_i) = \phi(x_i) - \frac{1}{N}\sum_i \phi(x_i)$ and equivalently for $\bar{\psi}(y_i)$.

We then have the following empirical estimate of COCO:

$$\left\langle X\alpha, \frac{1}{N} X H Y^T Y \beta \right\rangle = \alpha^T H X^T X H Y^T Y H \beta$$

since $HH = H$

$$= \alpha^T H X^T X H H Y^T Y H \beta$$

$$= \alpha^T H K H H L H \beta = \boxed{\frac{1}{N} \alpha^T \tilde{K} \tilde{L} \beta}$$

We then solve for α, β by maximizing the following Lagrangian:

$$\mathcal{L}(\alpha, \beta, \lambda, \gamma) = \frac{1}{N}\alpha^T \tilde{K}\tilde{L}\beta - \frac{\lambda}{2}\left(\alpha^T \tilde{K}\alpha - 1\right) - \frac{\gamma}{2}\left(\beta^T \tilde{L}\beta - \right.$$

$$\|f\| = \|XH\alpha\| = 1$$

$$= \|XH\alpha\|^2$$

$$= \alpha^T H X^T X H \alpha$$

$$= \alpha^T \tilde{K}\alpha = 1$$

Differentiating, we get:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \alpha} = \frac{1}{N}\tilde{K}\tilde{L}\beta - \lambda \tilde{K}\alpha = 0 \\ \frac{\partial \mathcal{L}}{\partial \beta} = \frac{1}{N}\tilde{L}\tilde{K}\alpha - \gamma \tilde{L}\beta = 0 \end{cases}$$

Multiplying by α^T for eqn 1, β^T for eqn 2:

$$\Rightarrow \begin{matrix} \frac{1}{N}\alpha^T \tilde{K}\tilde{L}\beta = \lambda \alpha^T \tilde{K}\alpha \\ \frac{1}{N}\alpha^T \tilde{K}\tilde{L}\beta = \gamma \beta^T \tilde{L}\beta \end{matrix} \iff \lambda \underbrace{\alpha^T \tilde{K}\alpha}_{1} = \gamma \underbrace{\beta^T \tilde{L}\beta}_{1}$$

$$\underline{\lambda = \gamma}$$

$$\Rightarrow \begin{bmatrix} 0 & \frac{1}{N}\tilde{K}\tilde{L} \\ \frac{1}{N}\tilde{L}\tilde{K} & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \lambda \begin{bmatrix} \tilde{K} & 0 \\ 0 & \tilde{L} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

Generalized eigenvalue problem

By taking the eigenvector $\binom{\alpha}{\beta}$ with highest eigenvalue, we then have $f = X^T H\alpha$, $g = Y^T H\beta$ that maximize the covariance - but only want the first component of the eigen spectrum! It turns out that $HSIC = \sum \lambda_i^2$, so it is better than COCO b/c it computes the whole eigensp

Note that the constraints $\|f\|_{\mathcal{F}} = 1$, $\|g\|_{\mathcal{G}} = 1$ enforce smoothness in f and g, such that COCO is insensitive to high frequency dependencies (need high sample size to detect).

- Returning to HSIC, we now ask what an unbiased estimate would be and how biased $Tr(KHLH)$ is:

<u>Term 1</u>: $$\langle \tilde{C}_{xy}, \tilde{C}_{xy} \rangle_{\mathcal{F}\otimes\mathcal{G}} = \mathbb{E}_{x,y\sim P_{xy}} \mathbb{E}_{x',y'\sim P_{xy}} k(x,x')\,l(y,y')$$

$$\simeq \left(\sum_{i=1}^{N} \sum_{j\neq i} k(x_i,x_j)\,l(y_i,y_j) \right) \frac{1}{N(N-1)}$$

Difference b/w biased and unbiased estimate is the

$$\underbrace{\frac{1}{N^2}\sum_{i,j} K_{ij}L_{ij}}_{\text{biased}} - \underbrace{\frac{1}{N(N-1)}\sum_{j\neq i} K_{ij}L_{ij}}_{\text{unbiased}}$$

$$= \frac{1}{N^2}\sum_i K_{ii}L_{ii} + \frac{1}{N^2}\sum_{j\neq i} K_{ij}L_{ij} - \frac{1}{N(N-1)}\sum_{j\neq i} K$$

$$= \frac{1}{N}\left(\frac{1}{N}\sum_i K_{ii}L_{ii} - \frac{1}{N(N-1)}\sum_{j\neq i} K_{ij}L_{ij} \right)$$

$$\mathbb{E}[\text{bias}] = \frac{1}{N}\left(\mathbb{E}_x k(x,x)\,\mathbb{E}_y l(y,y) - \frac{N(\cancel{N-1})}{N(\cancel{N-1})} \mathbb{E}_x \mathbb{E}_{x'} k(x,x')\,\mathbb{E}_y \mathbb{E}_{y'} \right.$$

$$\sim \mathcal{O}\left(\frac{1}{N}\right)$$

Term 2 : $\langle \mu_x \otimes \mu_y , \mu_x \otimes \mu_y \rangle_{F \times G} = \mathbb{E}_{x \sim P_x} \mathbb{E}_{x' \sim P_x} k(x,x') \mathbb{E}_{y \sim P_y} \mathbb{E}_{y' \sim P_y} \ell(y,y')$

$$\simeq \frac{1}{N(N-1)(N-2)(N-3)} \sum_{i} \sum_{j \neq i} \sum_{q \neq i,j} \sum_{r \neq i,j,q} K_{ij} L_{qr}$$

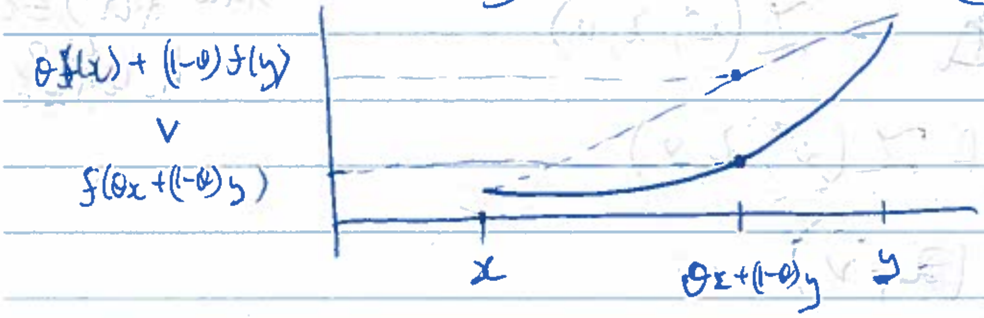
Term 3 : $\langle \tilde{C}_{xy} , \mu_x \otimes \mu_y \rangle_{F \times G} = \mathbb{E}_{x,y \sim P_{xy}} \mathbb{E}_{x' \sim P_x} k(x,x') \mathbb{E}_{y' \sim P_y} \ell(y,y')$

$$\simeq \left[\sum_{i} \sum_{j \neq i} k(x_i, x_j) \sum_{q \neq i,j} \ell(y_i, y_q) \right] \frac{1}{N(N-1)(N-}$$

Part III: SVMs and Convex Optimization

- A set C is convex if for any $x_1, x_2 \in C$, $\theta x_1 + (1-\theta) x_2 \in C$, $0 \leq \theta \leq 1$
- A function $f(x)$ is convex if its domain $\operatorname{dom} f$ is a convex set and for any $0 \leq \theta \leq 1$,

$$f(\theta x + (1-\theta) y) \leq \theta f(x) + (1-\theta) f(y)$$



- Suppose we want to solve the following optimization problem: minimize $f_0(x)$
 subject to $f_i(x) \leq 0 \quad i = 1, \dots, m$
 $h_i(x) = 0 \quad i = 1, \dots, p$

Calling us the optimum $f_0(x^*)$.
It turns out we can solve this by solving a different easier — convex — optimization problem, called the Lagrange dual problem:

$$\begin{aligned} &\text{maximize} && g(\lambda, \nu) \\ &\text{subject to} && \lambda \succeq 0 \qquad (\lambda_i \geq 0 \;\; \forall i) \end{aligned}$$

where $g(\lambda, \nu) = \inf_x \mathcal{L}(x, \lambda, \nu)$

$$= \inf_x \underbrace{\left[f_0(x) + \sum_i \lambda_i f_i(x) + \sum_j \nu_j h_j(x) \right]}_{\text{Lagrange dual function}}$$

Lagrange multipliers

dual vars

$$= \sup_{\lambda \succeq 0} \mathcal{L}(x^*, \lambda, \nu) \quad \left(\text{i.e. } \lambda_i = 0 \text{ since } f_i(x^*) \le 0 \right.$$

$$= \inf_x \sup_{\lambda \succeq 0} \mathcal{L}(x, \lambda, \nu)$$

We can see this Lagrange dual problem is equivalent to our original minimization problem by noting that $g(\lambda, \nu)$ is upper bounded by $f_0(x^*$

$$f_0(x^*) \ge f_0(x^*) + \sum_i \lambda_i f_i(x^*) + \sum_j \nu_j h_i(x^*) \quad \text{since } f_i(x^*) \le \ \lambda \succeq 0 \ \ h_i(x^*) = 0$$

$$= \mathcal{L}(x^*, \lambda, \nu)$$

$$\ge \inf_x \mathcal{L}(x^*, \lambda, \nu)$$

$$\ge g(\lambda, \nu)$$

We call the solution to the Lagrange dual problem ~~λ* dual optimal~~ (λ^*, ν^*) dual optimal.

We know the dual problem is convex, since $g(\lambda, \nu)$ is concave and the constraint set is convex.
$\hookrightarrow \lambda \succeq 0, \ g(\lambda, \nu) > -\infty$

(λ, ν) is dual feasible

Two cases are possible when replacing the original opt. problem w/ the dual problem:

$$g(\lambda^*, \nu^*) \le f_0(x^*) \qquad \text{weak duality}$$

$$g(\lambda^*, \nu^*) = f_0(x^*) \qquad \text{strong duality}$$

The conditions under which strong duality holds are called constraint qualifications

(put simply, there exists an $\tilde{x}$ that satisfies all the constraints)
I think!

(sufficient, not necessary conditions for strong duality)

One example is:

- primal problem is convex: i.e. $f_i(x)$ convex and $h(x) = Ax + b$
- Slater's condition holds: there exists some strictly feasible point $\tilde{x}$ s.t. $f_i(\tilde{x}) < 0$ $\forall i$ and $h(\tilde{x})$:

Note that when strong duality holds (i.e. the dual problem is equal to the primal),

$$f_0(x^*) = g(\lambda^*, \nu^*)$$

$$= \inf_x f_0(x) + \sum_i \lambda_i^* f_i(x) + \sum_j \nu_j^* h_i(x)$$

$$\leq f_0(x^*) + \sum_i \lambda_i^* f_i(x^*) + \cancelto{0}{\sum_j \nu_j^* h_i(x^*)}$$

$$\leq f_0(x^*) \Rightarrow \sum_i \lambda_i^* f_i(x^*) = 0$$

This implies complementary slackness:

$$\lambda_i^* > 0 \Rightarrow f_i(x^*) = 0$$

$$f_i(x^*) < 0 \Rightarrow \lambda_i^* = 0$$

Lastly, if f_0, $\{f_i\}$, and $\{h_i\}$ are all differentiable, then

$$\nabla f_0(x^*) + \sum_i \lambda_i^* \nabla f_i(x^*) + \sum_j \nu_j \nabla h_j(x^*) = 0$$

Putting all the above conditions together gives the KKT conditions!

$$f_i(x) \leq 0 \qquad \lambda_i f_i(x) = 0$$

$$h_i(x) = 0 \qquad \nabla f_0(x) + \sum_i \lambda_i \nabla f_i(x) + \sum_i \nu_i \nabla h_i(x) = 0$$

$$\lambda_i \geq 0$$

~~When strong duality holds, Slater's condition holds, valid~~

When an optimization problem is convex and Slater's condition holds (i.e. strong duality holds) if $f_0, \{f_i\}, \{h_i\}$ are differentiable then the KKT conditions are necessary and sufficient for global optimality, i.e. a solution x^* that satisfies the KKT conditions is a global optimum.

- The representer theorem says that ~~an~~ the solution to ~~an~~ arbitrary optimization problem

$$f^* = \underset{f \in \mathcal{H}}{\operatorname{argmin}}\ L_y(f(x_1), \dots, f(x_N)) + \Omega(\|f\|^2_{\mathcal{H}})$$

where L_y is parametrized by $\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$ and $\Omega(\cdot)$ is non-decreasing, takes the form:

$$f^* = \sum_{i=1}^{N} \alpha_i K(x_i, \cdot)$$

where K is the kernel corresponding to the RKHS $\mathcal{H}$ where f lives.

Pf. Let $f = f_s + f_\perp$, where f_s is the projection of f onto the subspace spanned by $\{K(x_i, \cdot)\}_{i=1}^{N}$ and $f_\perp$ is the orthogonal error.

By reproducing property of RKHS $\mathcal{H}$,

$$f(x_i) = \langle f, x_i \rangle_{\mathcal{H}} = \langle f_s, x_i \rangle_{\mathcal{H}} + \langle f_\perp, x_i \rangle_{\mathcal{H}}$$

$$= \langle f_s, x_i \rangle_{\mathcal{H}}$$

so $L_y(f(x_1), \dots, f(x_N)) = L_y(f_s(x_1), \dots, f_s(x_N))$

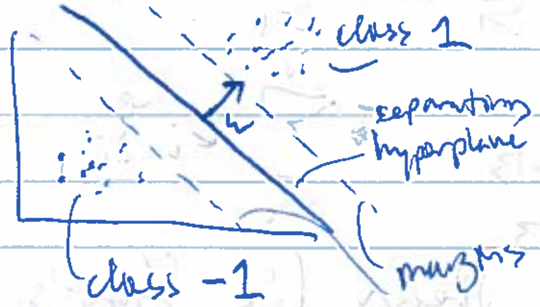
$$\|f\|^2_{\mathcal{H}} = \langle f, f\rangle_{\mathcal{H}} = \langle f_s, f_s\rangle_{\mathcal{H}} + \langle f_\perp, f_\perp\rangle_{\mathcal{H}} - 2\cancel{\langle f_s, f_\perp\rangle_{\mathcal{H}}}$$

$$= \|f_s\|^2_{\mathcal{H}} + \|f_\perp\|_{\mathcal{H}}$$

If $\Omega(\cdot)$ is strictly increasing, then its minimum is achieved when $\|f_\perp\|^2_{\mathcal{H}} = 0$ and $\|f_s\|^2_{\mathcal{H}}$ is minimized, leaving the optimization problem

$$f^* = \underset{f_s = \sum_{i=1}^{N} \alpha_i K(x_i, \cdot)}{\arg\min} L_y(f_s(x_1), \ldots, f_s(x_N)) + \Omega(\|f_s\|^2_{\mathcal{H}})$$

- In support vector classification we want to find a hyperplane that separates data of two different classes within the data space itself.



The best such hyperplane is the one that maximizes the distance b/w the margins while enforcing perfect classification.

Let w = vector perpendicular to hyperplane.
We want:

① $w^T x_i + b \geq 1 \quad \forall i : y_i = 1$

$w^T x_i + b \leq -1 \quad \forall i : y_i = -1$

$\Rightarrow y_i(w^T x_i + b) \geq 1$

② For x^+, x^- on opposite margins,

maximize $\frac{x^{+T} w}{\|w\|} - \frac{x^{-T} w}{\|w\|} = \frac{1}{\|w\|} - \frac{-1}{\|w\|} = \frac{2}{\|w\|}$

So, we want to solve

maximize $\frac{2}{\|w\|}$ / minimize $\frac{1}{2}\|w\|^2$

subject to $y_i(w^T x_i + b) \geq 1 \quad \forall i$

However, there will rarely exist a hyperplane that exactly separates the data (i.e. impossible to get $y_i(w^Tx_i+b)\geq$ for all x_i), so we soften the constraints with an error term ξ_i, which we minimize as well:

$$\min_{w,b}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i\right)$$

(controls trade-off b/w flexibility and accuracy)

$$\text{subject to}\quad y_i(w^Tx_i+b)\geq 1-\xi_i$$

$$\xi_i\geq 0$$

Giving us the following Lagrangian dual function:

$$\mathcal{L}(w,b,\xi,\lambda,\nu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\lambda_i(1-y_i(w^Tx_i+b)-$$

$$-\sum_{i=1}^{N}\nu_i\xi_i$$

- Noting that each of our constraints are convex and there always exists a set of $\{w, b, \{\xi_i\}\}$ that satisfies them (i.e. Slater's condition holds), since the objective and constraint functions are differentiable we need only solve for the KKT conditions to ~~find~~ the global optimum

(prove that strong duality holds and we can solve for the KKT conditions and optimize the dual Lagr)

$$\lambda_i\geq 0,\quad \nu_i\geq 0$$

$$\frac{\partial\mathcal{L}}{\partial w} = w - \sum_{i=1}^{N}\lambda_i y_i x_i = 0 \iff w = \sum_{i=1}^{N}\lambda_i y_i x_i$$

$$\frac{\partial\mathcal{L}}{\partial b} = -\sum\lambda_i y_i = 0 \qquad \frac{\partial\mathcal{L}}{\partial\xi_i} = C-\lambda_i-\nu_i = 0$$

$$\Rightarrow \lambda_i = C-\nu_i$$

Since $\lambda_i, v_i \geq 0$, we have $0 \leq \lambda_i \leq C$

Then, by complimentary slackness, we can work out the following three possible cases:

For $\lambda_i = C$,

$v_i = 0 \Rightarrow \xi_i \geq 0$ i.e. x_i lies inside the margin

$y_i(w^T x_i + b) = 1 - \xi_i \leq 1$

For $0 < \lambda_i < C$,

$v_i \geq 0 \Rightarrow \xi_i = 0$ i.e. x_i lies on the margin

$y_i(w^T x_i + b) = 1$

For $\lambda_i = 0$

$v_i = C \Rightarrow \xi_i = 0$ i.e. x_i correctly outside the margins

$y_i(w^T x_i + b) \geq 1$

In sum, ~~our solution is~~ only points on or inside the margin with $\lambda_i > 0$ contribute to ~~the support vector~~ $w = \sum \lambda_i y_i x_i$ their contribution bounded by C. These points are thus called the support vector

↳ note also that this is a sparse solution, for since most x_i will be outside the margin so $\lambda_i = 0$

Given that strong duality holds, we can find λ_i's by maximizing the Lagrangian dual function, plugging in our KKT conditions to simplify:

$$g(\lambda, \nu) = \frac{1}{2}\sum_{i,j} \lambda_i \lambda_j y_i y_j x_i^T x_j + C\sum_{i=1}^{N} \xi_i$$

$$+ \sum_i \lambda_i \left(1 - y_i\left(\sum_j \lambda_j y_j x_j^T x_i + b\right) - \xi_i\right)$$

$$- \sum_i \nu_i \xi_i$$

$$= \frac{1}{2}\sum_{i,j} \lambda_i \lambda_j y_i y_j x_i^T x_j + C\sum_i \xi_i$$

$$- \sum \lambda_i \xi_i + \sum_i \lambda_i - \sum_{i,j} \lambda_i \lambda_j y_i y_j x_i^T x_j$$

$$\cancel{-b\sum_i \lambda_i y_i} - \sum_i \nu_i \xi_i$$

$$= -\frac{1}{2}\sum_{i,j} \lambda_i \lambda_j y_i y_j x_i^T x_j + \cancel{\sum_i \overbrace{(C - \lambda_i)}^{\nu_i} \xi_i - \sum_i \nu_i \xi_i} + \sum_i$$

$$= -\frac{1}{2}\sum_{i,j} \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_i \lambda_i = g(\lambda)$$

Thus, to find the support vector, we solve

maximize $g(\lambda) = -\frac{1}{2}\|w\|^2 + \sum \lambda_i$

subject to $0 \le \lambda_i \le C$

We can the estimate b by solving $y_i(w^T x_i + b) \ge 1$ for an x_i on the margin (or averaging over all x_i on the m

- Since C is hard to interpret, we can reparametrize with a new parameter ν. This is called ν-SVM:

$$\min_{w,\rho,\xi} \frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{N}\sum_{i=1}^{N} \xi_i$$

subject to $y_i(w^T x_i + b) \geq \rho - \xi_i$

$$\xi_i, \rho \geq 0$$

Again Slater's condition holds and our constraints are convex, so we go on to write out the KKT conditions and then optimize the Lagrangian!

$\lambda_i \geq 0 \quad \xi_i \geq 0 \quad \frac{\partial \mathcal{L}}{\partial w} = w - \sum_i \lambda_i y_i x_i = 0$

$\alpha_i \geq 0 \quad y_i(w^T x_i + b) \geq \rho - \xi_i \quad \Rightarrow w = \sum_i \lambda_i y_i x_i$

$\gamma \geq 0 \quad \rho \geq 0 \quad \frac{\partial \mathcal{L}}{\partial b} = -\sum_i \lambda_i y_i = 0$

$\mathcal{L}(w, \rho, \xi, \lambda, \alpha, \gamma) \quad \frac{\partial \mathcal{L}}{\partial \xi_i} = \frac{1}{N} - \lambda_i - \alpha_i = 0$

$\Rightarrow \lambda_i + \alpha_i = \frac{1}{N}$

$$= \frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{N}\sum_i \xi_i + \sum_i \lambda_i\left(\rho - y_i(w^T x_i + b) - \xi_i\right)$$

$$- \sum_i \alpha_i \xi_i - \gamma\rho$$

$$\frac{\partial \mathcal{L}}{\partial \rho} = -\nu + \sum_i \lambda_i - \gamma = 0$$

$$\Rightarrow \nu = \sum_i \lambda_i - \gamma$$

$g(\lambda, \alpha, \gamma)$

~~$= -\frac{1}{2}\sum_{i,j} \lambda_i \lambda_j y_i y_j x_i^T x_j - \left(\sum_i \lambda_i\right)\rho + \nu\rho$~~

~~$+ \frac{1}{N}\sum_i \xi_i + \sum_i \lambda_i \rho - \sum_i \lambda_i \xi_i$~~

~~$- \sum_i \alpha_i \xi_i - \gamma\rho$~~

$$\boxed{= -\frac{1}{2}\sum_{i,j} \lambda_i \lambda_j y_i y_j x_i^T x_j = g(\lambda)}$$

So, ~~Lagrange dual problem~~

~~becomes:~~

~~minimize $g(\lambda) = \frac{1}{2}\sum_{i,j} \lambda_i \lambda_j y_i y_j x_i^T x_j$~~

~~subject to~~

~~By complimentary slackness,~~

So, Lagrange dual problem becomes:

minimize $g(\lambda)$

subject to $0 \le \lambda \le \frac{1}{N}$

How do we interpret ν?

- assume $\rho > 0 \Rightarrow \gamma = 0 \Rightarrow \nu = \sum_i \lambda_i$
- By complimentary slackness,

If $\xi_i > 0 \Rightarrow \alpha_i = 0 \Rightarrow \lambda_i = \frac{1}{N}$

If $\xi_i = 0 \Rightarrow \alpha_i \ge 0 \Rightarrow \lambda_i \le \frac{1}{N}$

$\Rightarrow$ For $N(\lambda) = \{\lambda_i = \frac{1}{N}\}$, $M(\lambda) = \{0 < \lambda < \frac{1}{N}\}$ $\sum_{i \in N(\lambda)} \lambda_i + \sum_{j \in M(\lambda)} \lambda_j < \frac{|N(\lambda)| + |M(\lambda)|}{N} \le \sum_i \lambda_i = \nu$

ν is thus a upper bound the # support vectors (the data p with non-z weight) $\geq$

RKH

- We can "kernelize" SVM by moving into a feature space $\mathcal{H}$: $K(x_i, \cdot)$, $w \in \mathcal{H}$. Writing down the objective as follows:

$$w^* = \underset{w \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{2}\|w\|^2_{\mathcal{H}} + C \sum_i \xi_i$$

$$= \underset{w \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{2}\|w\|^2_{\mathcal{H}} + C \sum_i [1 - y_i (\langle w, K(x_i, \cdot)\rangle_{\mathcal{H}} +$$

rectification

$$= \underset{w \in \mathcal{H}}{\operatorname{argmin}} \Omega(\|w\|_{\mathcal{H}}) + F_y(\langle w, K(x_1, \cdot)\rangle, \dots, \langle w, K(x_N, \cdot)\rangle)$$

we recognize that $\Omega(\cdot) = \frac{1}{2}(\cdot)^2$ is non decreasin so the representer theorem applies, giving us

$$w^* = X\beta$$

Thus the minimization problem becomes

$$\text{minimize} \quad \frac{1}{2}\beta^T K \beta + C \sum_i \xi_i$$

$$\text{subject to} \quad \xi_i \geq 0$$

$$y_i\left(\sum_j \beta_j K(x_i, x_j) + b\right) \geq 1 - \xi_i$$

Since K is positive definite, the objective function and constraints are convex so Slater's condition gives us strong duality

We can thus instead optimize the dual, as before:

minimize $g(\lambda) = -\frac{1}{2}\sum_{i,j} \lambda_i \lambda_j y_i y_j K(x_i, x_j) + \sum_i \lambda_i$

subject to $0 \le \lambda_i \le C$

- My convex optimization recipe!

① write down objective and constraints

② Check constraints don't conflict
if not → Slater's condition holds

③ Check if constraints and objective are conve

if so → strong duality holds

④ Write down Lagrangian

⑤ write down KKT conditions ~~[illegible]~~ (need to differentiate Lagrangian)

⑥ Expand Lagrangian and simplify using KKT conditions, plugging in where possible

⑦ Rewrite optimization problem in terms of the Lagrangian dual subject to constraints implied by KKT conditions

⑧ Use complimentary slackness conditions to interpret different parameter values