Kernel methods

prediction functions: $f : \mathcal{X} \to \mathbb{R}$

$f(x) = \varphi(x)^T \theta \qquad \theta \in \mathbb{R}^d$

$\varphi : \mathcal{X} \to \mathbb{R}^d$ feature map

Estimation error = optimization error + uniform deviations

$$\frac{GRD}{\sqrt{n}}$$

Lipschitz constant of the loss

bound on $\|\theta\|_2$

bound on $\|\varphi(x)\|_2$

$f(x) = \langle \varphi(x), \theta \rangle$

$\varphi(x) : \mathcal{X} \to \mathcal{H}$ Hilbert space

Training data: $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$ $\quad i = 1, \dots, n$

Model: $f_\theta(x) = \langle \varphi(x), \theta \rangle$ $\qquad \varphi : \mathcal{X} \to \mathcal{H}$

Training objective:

$$\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle \varphi(x_i), \theta \rangle) + \frac{\lambda}{2} \|\theta\|^2$$

Representer theorem: The solution is of the form

(if $\lambda > 0$) $\quad \theta = \sum_{i=1}^{n} \alpha_i \varphi(x_i) \quad \Rightarrow \alpha \in \mathbb{R}^n$

Proof: $\left\{ \sum_{i=1}^{n} \alpha_i \varphi(x_i), \alpha \in \mathbb{R}^n \right\} = \mathcal{H}_D \subset \mathcal{H}$

$$\mathcal{H} = \mathcal{H}_D \oplus \mathcal{H}_D^{\perp} \Rightarrow \theta = \theta_{\mathcal{H}_D} + \theta_{\mathcal{H}_D^{\perp}}, \quad \overbrace{\|\theta\|^2 = \|\theta_{\mathcal{H}_D}\|^2 + \|\theta_{\mathcal{H}_D^{\perp}}\|^2}^{\text{Pythagore}}$$

$$\langle \varphi(x_i), \theta \rangle = \langle \varphi(x_i), \theta_{\mathcal{H}_D} \rangle + \underbrace{\langle \varphi(x_i), \theta_{\mathcal{H}_D^{\perp}} \rangle}_{0}$$

$$\Rightarrow \boxed{\theta_{\mathcal{H}_D^{\perp}} = 0}$$

$$\min_{\theta \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \langle \varphi(x_i), \theta\rangle) + \frac{\lambda}{2}\|\theta\|^2 \qquad (E)$$

$$\boxed{\varphi : \mathcal{X} \to \mathcal{H}}$$

$$\theta = \sum_{j=1}^{n} \alpha_j \varphi(x_j) \Rightarrow \bullet\ \|\theta\|^2 = \Big\|\sum_{i=1}^{n} \alpha_i \varphi(x_i)\Big\|^2$$

$$= \sum_{i,j=1}^{n} \alpha_i \alpha_j \underbrace{\langle \varphi(x_i), \varphi(x_j)\rangle}_{k(x_i,x_j)\ \text{kernel function}}$$

$$= \alpha^T K \alpha$$

$\llcorner\ K \in \mathbb{R}^{n\times n}$ $K_{ij} = k(x_i, x_j)$ kernel matrix

$$\bullet\ \langle \varphi(x_i), \theta\rangle = \sum_{j=1}^{n} \alpha_j \langle \varphi(x_i), \varphi(x_j)\rangle$$

$$= (K\alpha)_i$$

$$(E) \Leftrightarrow \boxed{\min_{\alpha \in \mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2}\alpha^T K \alpha}$$

prediction function $f(x) = \langle \theta, \varphi(x)\rangle = \sum_{i=1}^{n} \alpha_i k(x, x_i)$

KERNEL TRICK

Definition (positive definite kernel): $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
p.d.

k is a p.d kernel $\iff$ $\forall x_1, \dots, x_n \in \mathcal{X}$
the kernel matrix $K_{ij} = k(x_i, x_j)$
$i, j = 1, \dots, n$
is positive semi-definite (PSD)
$(K \succeq 0)$
$(\forall \alpha \in \mathbb{R}^n, \ \alpha^T K \alpha \geq 0)$

Theorem (Aronszajn, 1950)
k is p.d $\iff$ $\exists \varphi : \mathcal{X} \to \mathcal{H}$ Hilbert space (feature space)
such that $k(x, x') = \langle \varphi(x), \varphi(x') \rangle$

Proof $\Leftarrow$ $\alpha^T K \alpha = \sum_{i,j} \alpha_i \alpha_j \langle \varphi(x_i), \varphi(x_j) \rangle$
$= \| \sum_i \alpha_i \varphi(x_i) \|^2 \geq 0$

<u>Linear kernel</u>: $\mathcal{X} = \mathbb{R}^d$ $\quad k(x,y) = x^\top y$

$\varphi(x) = x \quad \mathcal{H} = \mathcal{X}$

$f_\theta(x) = \theta^\top \varphi(x) = \theta^\top x$

if $d > n$, solving for α is more efficient.

if sparse x, x', $\quad x^\top x'$ can be computed in time $o(d)$

<u>polynomial kernel</u>: $\mathcal{X} = \mathbb{R}^d$, $s \in \mathbb{N}$, $s \geq 1$

$k(x,x') = (x^\top x')^s$ is p.d

$$= \left(\sum_{i=1}^d x_i x_i'\right)^s = \sum_{\alpha_1 + \cdots + \alpha_d = s} \binom{s}{\alpha_1, \ldots, \alpha_d} \underbrace{(x_1 x_1')^{\alpha_1} \cdots (x_d x_d')^{\alpha_d}}_{x_1^{\alpha_1} \cdots x_d^{\alpha_d} \cdot x_1'^{\alpha_1} \cdots x_d'^{\alpha_d}}$$

$$= \varphi(x)^\top \varphi(x')$$

with $\varphi(x)_{\alpha_1, \ldots, \alpha_d} = \binom{s}{\alpha_1 \cdots \alpha_d}^{1/2} x_1^{\alpha_1} \cdots x_d^{\alpha_d}$

dimension $\binom{d+s-1}{s} \sim d^s$

$f_\theta(x) = \langle \theta, \varphi(x) \rangle$

polynomial homogeneous

<u>Mercer</u>: $\mathcal{X}$ compact

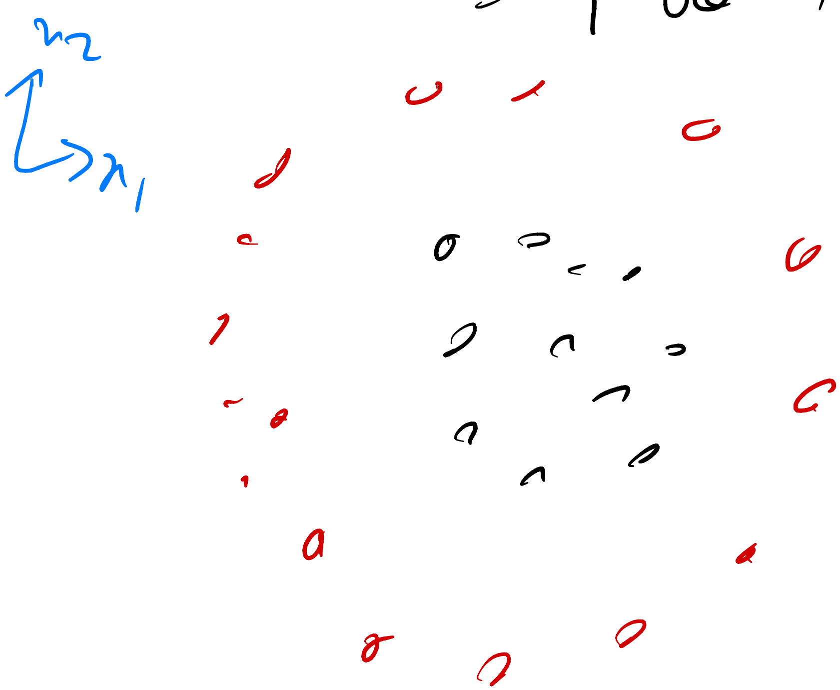
$$k(x,x') = \sum_{i \geq 0} \lambda_i \, e(x)_i \, e(x')_i \qquad \varphi(x)_i = \lambda_i^{1/2} e_i(x)$$

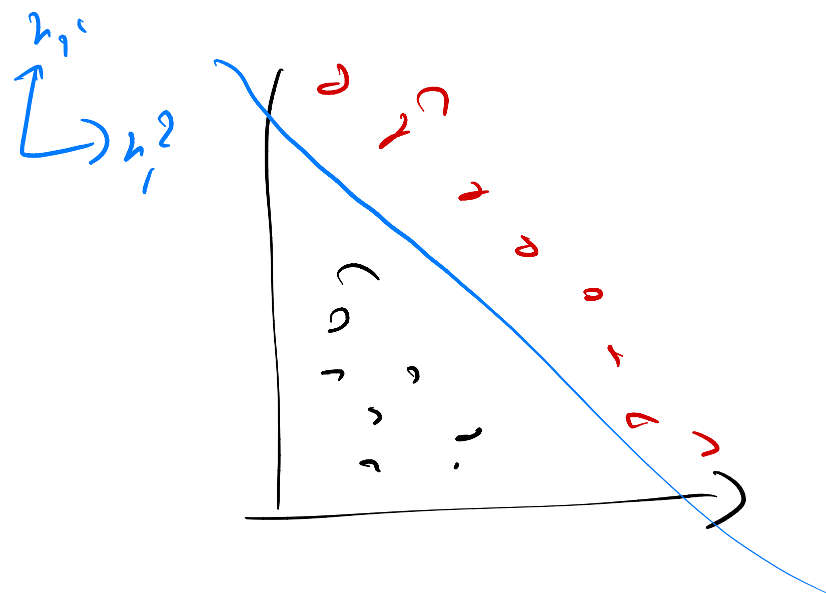
$\|e_i\|_\infty$ <u>bounded</u>

<u>RKHS</u> $= \varphi : \mathcal{X} \to \mathcal{H}$.

$\mathcal{H}$ space of functions from $\mathcal{X} \to \mathbb{R}$

$= \{ f_\theta(x) = \langle \theta, \varphi(x) \rangle \}$ with norm $\|\theta\|^2$





Translation invariant kernels: $\mathcal{X} = \mathbb{R}^d$

$k(x, x') = q(x - x')$. $q: \mathbb{R}^d \to \mathbb{R}$

Bochner Theorem: k is p.d $\iff$ q is the Fourier transform of a non negative Borel measure

Simplified version: if $\hat{q} \geq 0$ then k is p.d.

$$\hat{q}(\omega) = \int q(x) e^{-i\omega^T x} dx$$ Fourier transform

$$q(x) = \frac{1}{(2\pi)^d} \int \hat{q}(\omega) e^{i\omega^T x} d\omega$$ inversion formula

Parseval's identity: $\int |q(x)|^2 dx = \frac{1}{(2\pi)^d} \int |\hat{q}(\omega)|^2 d\omega$

Proof: $\alpha^T K \alpha = \sum_{s,j=1}^{n} \alpha_s \alpha_j \frac{1}{(2\pi)^d} \int \hat{q}(\omega) e^{i\omega^T (x_s - x_j)} d\omega$

$$= \int \frac{\hat{q}(\omega)}{(2\pi)^d} \sum_{s,j} \alpha_s \alpha_j e^{i\omega^T x_s} e^{-i\omega^T x_j} d\omega = \int \frac{\hat{q}(\omega)}{(2\pi)^d} \left| \sum_s \alpha_s e^{i\omega^T x_s} \right|^2 d\omega \geq 0$$

Example: ① $h(x,x') = \exp\left(-\frac{\|x-x'\|_2^2}{s^2}\right)$ Gaussian kernel

② $h(x,x') = \exp\left(-\frac{\|x-x'\|_2}{s}\right)$ Exponential kernel

$\to$ FT $\propto \frac{1}{\left(1+\|\omega\|^2\right)^{\frac{d+1}{2}}}$

③ Matern kernels

FT $\propto \frac{1}{\left(1+\|\omega\|^2\right)^s}$ $\qquad \underline{s > \frac{d}{2}}$

$f_\Theta(x) = \langle \Theta, \varphi(x) \rangle$ $\qquad \|\Theta\|^2$

$h(x,x') = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \hat{q}(\omega)\, e^{i\omega^T(x-x')} = \frac{1}{(2\pi)^d}\int \hat{q}(\omega)^{1/2} e^{i\omega^T x}\left(\hat{q}(\omega)^{1/2} e^{i\omega^T x'}\right)^* d\omega$

$= \int \varphi(x)_\omega \varphi(x')_\omega \, d\omega$ if $\varphi(x) = \frac{1}{(2\pi)^{d/2}}\hat{q}(\omega)^{1/2} e^{i\omega^T x}$

$= \langle \varphi(x), \varphi(x') \rangle$

$\to \underbrace{\langle \Theta, \varphi(x) \rangle}_{f_\Theta(x)} = \frac{1}{(2\pi)^{d/2}}\int \hat{q}(\omega)\,\Theta_\omega\, e^{i\omega^T x}\, d\omega \Rightarrow \hat{f}_\Theta(\omega) = (2\pi)^{d/2}\,\Theta_\omega\, \hat{q}(\omega)^{1/2}$

$\|\Theta\|^2 = \int |\Theta_\omega|^2 = \boxed{\frac{1}{(2\pi)^d}\int \frac{|\hat{f}_\Theta(\omega)|^2}{\hat{q}(\omega)}\, d\omega}$

Derivatives and Fourier transforms

$$f(x) = \frac{1}{(2\pi)^d} \int \hat{f}(\omega)\, e^{i\omega^T x}\, d\omega$$

$$\|{}^{H}\| \to \|{}^{\prime\prime}33$$

$$f'(x) = \frac{1}{(2\pi)^d} \int \hat{f}(\omega)\, i\omega\, e^{i\omega^T x}\, d\omega$$

$$\Rightarrow \quad FT(f') = i\omega \cdot FT(f)$$

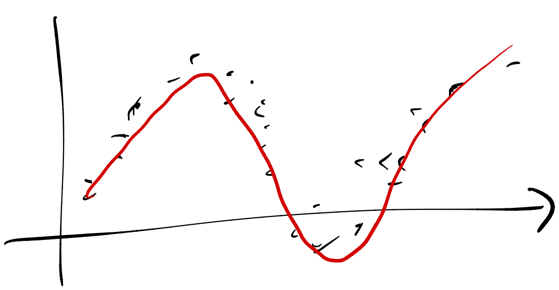
$$\frac{1}{(2\pi)^d} \int \underbrace{|\hat{f}(\omega)|^2 \cdot \|\omega\|_2^2}_{\|\hat{f}(\omega)\, i\omega\|_2^2} = \int \|f'(x)\|^2\, dx$$

Ex: $d=1$. $q(x) = e^{-|x|} \Rightarrow k(x,x') = e^{-|x-x'|}$

$$\hat{q}(\omega) = \frac{1}{1+\omega^2} \Rightarrow \int \frac{|\hat{f}(\omega)|^2}{\hat{q}(\omega)} = \int |\hat{f}(\omega)|^2 (1+\omega^2)$$

$$= \int |\hat{f}(\omega)|^2 + \int \omega^2 |\hat{f}(\omega)|^2$$

$$= \int |f|^2 + \int |f'|^2$$



$\boxed{\text{Sobolev norm}}$

Kernels for everyone

Algo: $\min_\theta \frac{1}{n}\sum_i \ell(y_i, f_\theta(x_i)) + \frac{\lambda}{2}\|\theta\|^2$, with $f_\theta(x_i) = \langle\theta, \varphi(x_i)\rangle$

$\min \frac{1}{n}\sum_i \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2}\alpha^T K \alpha$

square loss: $\min_\alpha \frac{1}{2n}\|y - K\alpha\|_2^2 + \frac{\lambda}{2}\alpha^T K\alpha$

Hessian: $\frac{K^2}{n} + \lambda K \Rightarrow$ slow convergence

$K = \Phi\Phi^T \quad \Phi \in \mathbb{R}^{n\times r} \quad r = \text{rank of } K$

$\beta = \Phi^T\alpha \qquad \min_\beta \frac{1}{2n}\|y - \Phi\beta\|_2^2 + \frac{\lambda}{2}\|\beta\|^2$

Hessian: $\frac{\Phi^T\Phi}{n} + \lambda I \Rightarrow$ faster convergence

empirical feature map

$\Rightarrow$ source of approximation methods

r small: $O(nr^2 + r^3)$ from $O(n^3)$

+ Duality

+ random features

$\Rightarrow$ Section 7.4

SGD: $F(\theta) = \mathbb{E}[\ell(y, \langle \theta, \phi(x) \rangle)]$ test error

$\theta_0 = 0$

$$\theta_t = \theta_{t-1} - \gamma \frac{\partial \ell(y_t, \langle \theta, \varphi(x_t) \rangle)}{\partial \theta} \Big|_{\theta = \theta_{t-1}}$$

$$\theta_t = \theta_{t-1} - \gamma \underbrace{\ell'(y_t, \langle \theta_{t-1}, y_t \rangle)}_{\alpha_t \in \mathbb{R}} \varphi(x_t) \in \mathcal{F}$$

$$\theta_t = \underbrace{\theta_0}_{0} - \gamma \sum_{u=1}^{t} \alpha_t \varphi(x_t)$$

$$\alpha_t = \ell'(y_t, -\gamma \sum_{u=1}^{t-1} \alpha_{t-1} \underbrace{\langle \phi(x_u), \phi(x_t) \rangle}_{h(x_u, x_t)}$$

single pass

after n steps

$O(n^2)$ running time

$$\mathbb{E}[F(\bar{\theta}_n)] - F(\theta) \le \frac{\|\theta\|_2^2}{2\gamma n} + \frac{\gamma}{2} G^2 R^2$$

$$\mathbb{E}[F(\bar{\theta}_n)] \le \inf_{\theta} \left(F(\theta) + \frac{\|\theta\|^2}{2\gamma n} \right) + \frac{\gamma}{2} R^2 G^2$$

cf section 7.5.1.1

$$\mathbb{E}\left[F(\bar{\theta}_n)\right] \leq \inf_{\theta} \left(F(\theta) + \frac{\|\theta\|^2}{2\gamma n} \right) + \frac{\gamma R^2 G^2}{2}$$

Assume: the set $\{ \langle \theta, \varphi(x) \rangle, \ \theta \in H \}$ is dense in $L^2(\mathcal{X})$

ex: translation invariant kernels in $\mathbb{R}^d$

then $\inf_{\theta \in H} F(\theta) = \inf_{\theta \in H} \mathbb{E}\, \ell(y, \langle \theta, \varphi(x) \rangle)$

$= \text{Bayes risk}$

typically not attained unless $f^*(x) = \langle \varphi(x), \theta_* \rangle$

① f^* is in my model class.

$$\mathbb{E}\, F(\bar{\theta}_n) \leq F(\theta_*) + \underbrace{\frac{\|\theta_*\|^2}{2\gamma n} + \frac{\gamma R^2 G^2}{2}}_{}$$

$\inf_{\gamma} \Rightarrow \gamma \propto \frac{1}{\sqrt{n}}$

$$\boxed{\mathbb{E}\, F(\bar{\theta}_n) \leq F(\theta_*) + \frac{\|\theta_*\| R G}{\sqrt{n}}}$$

② f^* not in the model class

- Matern kernel : $\hat{q}(\omega) \propto \frac{1}{(1+\|\omega\|^2)^s}$ $s > \frac{d}{2}$
- f is in the model class
  $\Leftrightarrow$ all derivatives of order s are in L^2
- f has only order t derivatives in L^2, $t < s$

$$\Rightarrow \boxed{\inf F(\theta) + \lambda\|\theta\|^2 \le F^* + O(\lambda^{t/s})}$$

$$E[F(\bar{\theta}_n)] \le \inf_\theta \underbrace{\left(F(\theta) + \frac{\|\theta\|^2}{2\gamma n}\right)} + \frac{\gamma R^2 G^2}{2}$$

$t = s \Rightarrow \frac{1}{\sqrt{n}}$

$t = 1 \Rightarrow \frac{1}{n^{1/s+1}}$

$s > \frac{d}{2} \Rightarrow \boxed{\frac{1}{n^{1/d}}}$

$$F(\theta_\gamma) + \frac{1}{(\gamma n)^{t/s}} + \gamma$$

$$\le \frac{1}{n^{t/s+t}} + F(\theta_*)$$

tradeoff when

$\gamma^{1+t/s} = n^{-t/s}$

$\gamma = n^{\frac{-t/s}{1+t/s}}$

$\gamma = n^{-t/s+t}$