Module 2: Connections between NNs & Kernels.

Review: (1) Kernel Regression

$x \mapsto \psi(x)$

$\mathbb{R}^d \quad \mathbb{R}^p$

$\{(\psi(x^{(i)}), y^{(i)})\}_{i=1}^n$

$\omega \; \psi(x^{(i)}) = y^{(i)}$

$K(x, \tilde{x}) \in \mathbb{R} \rightarrow \langle \psi(x), \psi(\tilde{x}) \rangle$.

Algorithm: Given data (X, y), kernel K

(1) $K_{train} = K(X, X)$

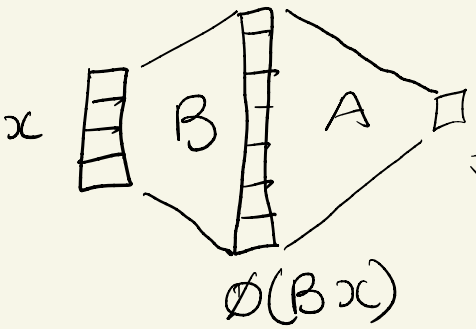
(2) $\alpha = y K_{train}^{-1}$.

---

(2) Neural Networks. $f_w(x) = A \phi(Bx)$

1-hidden layer FC network; $A \in \mathbb{R}^{1 \times k}$, $B \in \mathbb{R}^{k \times d}$,

$\phi : \mathbb{R} \rightarrow \mathbb{R}$; $\phi(z) = \max(z, 0)$ ReLU.



$f(x) = A \phi(Bx)$

$\psi(x)$ psi

k : width

$k \rightarrow \infty$

$f(x) = \underbrace{A \frac{1}{\sqrt{K}} \phi(Bx)}_{\psi(x)}$ $\quad \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$

$\underbrace{\Sigma(x, \tilde{x})}_{\text{Kernel}} := \langle \psi(x), \psi(\tilde{x}) \rangle = \left\langle \underbrace{\frac{1}{\sqrt{K}} \phi(Bx)}_{\psi(x)}, \underbrace{\frac{1}{\sqrt{K}} \phi(B\tilde{x})}_{\psi(\tilde{x})} \right\rangle$

$\phi(Bx) \in \mathbb{R}^K$, $\phi(B\tilde{x}) \in \mathbb{R}^K$

$= \frac{1}{K} \sum_{i=1}^{K} \phi(Bx)_i \, \phi(B\tilde{x})_i$

$= \frac{1}{K} \sum_{i=1}^{K} \phi(B_{i,:} x) \, \phi(B_{i,:} \tilde{x})$

$K \begin{bmatrix} B_{1,:} \\ B_{2,:} \\ \vdots \\ B_{K,:} \end{bmatrix} = B \in \mathbb{R}^{K \times d}$

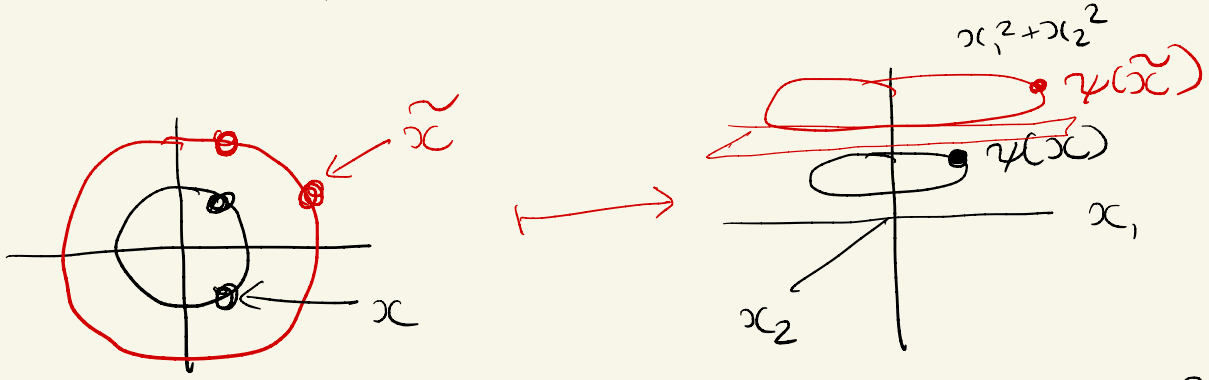
As $K \to \infty$ $\quad \to \mathbb{E}_{\omega}[\phi(\omega^T x)\phi(\omega^T \tilde{x})]$

$B_{ij} \sim N(0,1)$

$\omega_i \sim N(0,1)$

$B_{ij} \sim N(0, \frac{1}{K})$ Aside

Bochner's Thm.



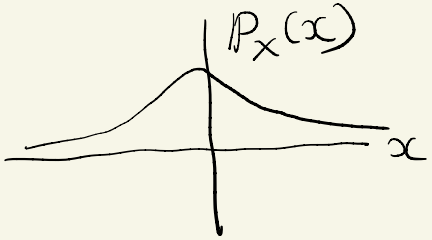
$\psi(x) = (x_1, x_2, x_1^2 + x_2^2)$ $\quad \psi(\tilde{x}) = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_1^2 + \tilde{x}_2^2)$

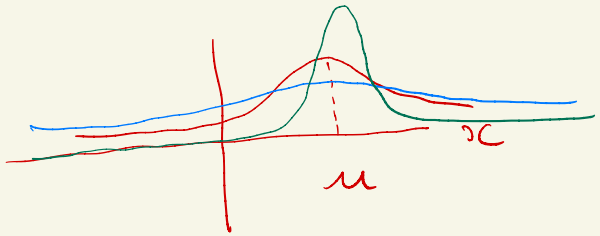
Review of Multivariate Gaussians

$x \in \mathbb{R}$; $P_x(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} x^2 \right\}$ $\quad x \sim \mathcal{N}(0,1)$ (Standard Gaussian)

$P_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}$ $\quad x \sim \mathcal{N}(\mu, \sigma^2)$

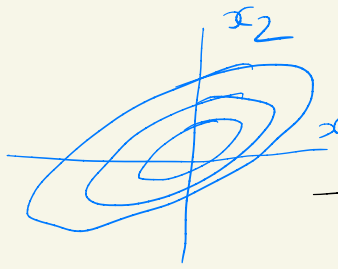
mean, variance

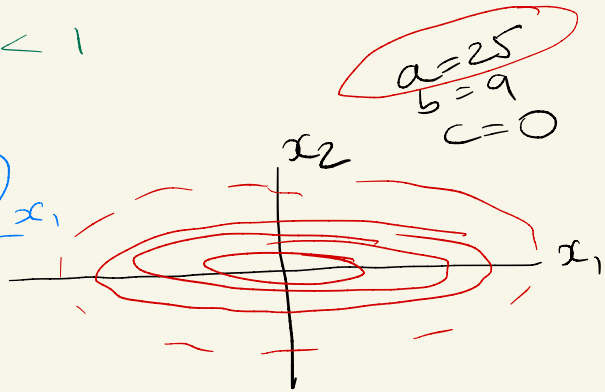




$\sigma >> 1$

$\sigma << 1$





2D - Gaussian Distribution

$x \in \mathbb{R}^2$ $\quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$x \sim \mathcal{N}\left(\underbrace{\begin{bmatrix} 0 \\ 0 \end{bmatrix}}_{\mu}, \Lambda \right)$ $\quad \Lambda \in \mathbb{R}^{2\times 2}$ symmetric.

$$P_x(x) = \frac{1}{2\pi \det(\Lambda)^{1/2}} \exp\left\{ -\frac{1}{2} (x-\mu)^T \Lambda^{-1} (x-\mu) \right\}$$

$$\Lambda = \begin{bmatrix} a & c \\ c & b \end{bmatrix} \quad ; \quad \det(\Lambda) = ab - c^2 \quad ; \quad \Lambda^{-1} = \frac{1}{ab-c^2} \begin{bmatrix} b & -c \\ -c & a \end{bmatrix}$$

$$f(x) = A \frac{1}{\sqrt{K}} \underbrace{\phi(Bx)}_{\psi(x)}$$

$$\Sigma(x, \tilde{x}) := \frac{1}{K} \sum_{i=1}^{K} \phi(B_{i,:} x)\ \phi(B_{i,:} \tilde{x})$$

$B_{ij} \sim \mathcal{N}(0,1)$

$$\omega^T : [\omega_1, \omega_2, \dots, \omega_d] \; ; \; \omega_i \sim \mathcal{N}(0,1)$$

Law of Large Numbers $\Rightarrow$ As $K \to \infty$:

$$\Sigma(x, \tilde{x}) = \mathbb{E}_{\omega} [\phi(\omega^T x)\ \phi(\omega^T \tilde{x})]$$

$u := \omega^T x \; ; \; v := \omega^T \tilde{x}$

$u \sim \mathcal{N}(0, \|x\|_2^2)$

$v \sim \mathcal{N}(0, \|\tilde{x}\|_2^2)$

$uv \sim \mathcal{N}(0, x^T \tilde{x})$

$$\Sigma(x, \tilde{x}) = \mathbb{E}_{(u,v)} [\phi(u)\ \phi(v)]$$

$$(u,v) \sim \mathcal{N}\left(0, \begin{bmatrix} \|x\|_2^2 & x^T \tilde{x} \\ x^T \tilde{x} & \|\tilde{x}\|_2^2 \end{bmatrix}\right)$$

$$u = \sum_{i=1}^{d} \omega_i x_i$$

$$\mathbb{E}[u] = \sum_{i=1}^{K} \mathbb{E}[\omega_i x_i] = 0.$$

$$\mathbb{E}[u^2] = \mathbb{E}\left[\left(\sum_{i=1}^{K} \omega_i x_i\right)^2\right]$$

$$= \sum_{i=1}^{d} \mathbb{E}[\omega_i^2 x_i^2] + 0$$

$$= \sum_{i=1}^{d} x_i^2 = \|x\|_2^2$$

$$\Sigma(x, \tilde{x}) = \mathop{\mathbb{E}}_{(u,v)} [\phi(u)\,\phi(v)]$$

$$(u, v) \sim \mathcal{N}\left(0, \begin{bmatrix} \|x\|_2^2 & x^T\tilde{x} \\ x^T\tilde{x} & \|\tilde{x}\|_2^2 \end{bmatrix}\right)$$

Neural Net Gaussian Process

NNGP

$$\|x\|_2 = \|\tilde{x}\|_2 = 1.$$

$$\Sigma(x, \tilde{x}) = \check{\phi}(x^T\tilde{x})$$

"phi-hat" : dual activation

$$\phi \longrightarrow \check{\phi}$$

$$\phi(z) = z \Rightarrow \check{\phi}(\xi) = \xi$$

$$\phi(z) = h_i(z) \Rightarrow \check{\phi}(\xi) = \xi^i$$

Hermite polynomials

$$\phi(z) = \max(z, 0) \Rightarrow \check{\phi}(\xi) = \frac{1}{\pi}\left(\xi\,(\pi - \cos^{-1}(\xi)) + \sqrt{1-\xi^2}\right)$$

arc cosine kernel.

Module 2: Connecting NNs and Kernel Machines

Kernel Machines

Kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

Training:

(1) $K_{train} = K(X, X)$ } $x \mapsto \psi(x)$, $\langle \psi(x), \psi(z) \rangle = K(x, z)$

(2) $\alpha = y K_{train}^{-1}$

Neural Networks

$f_w(x) = A \frac{1}{\sqrt{K}} \phi(Bx)$

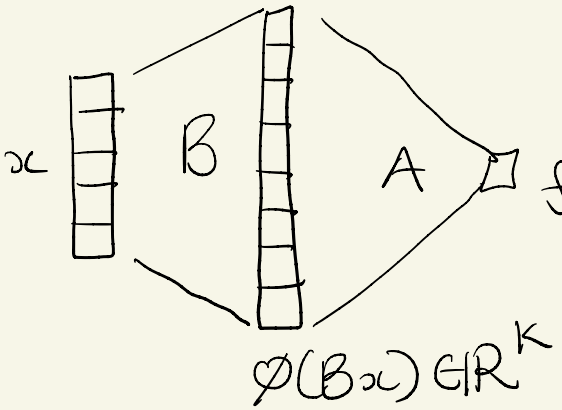
$A \in \mathbb{R}^{1 \times K}$ $B \in \mathbb{R}^{K \times d}$

$\phi : \mathbb{R} \to \mathbb{R}$ elementwise nonlinearity

Training: Gradient Descent on

$$\mathcal{L}(w) = \frac{1}{2} \sum_{p=1}^{n} (y^{(p)} - f_w(x^{(p)}))^2$$

Last time: NNGP (Neural Net Gaussian Process)



$\phi(Bx) \in \mathbb{R}^K$

linear regression.

$f_w(x) = A \frac{1}{\sqrt{K}} \phi(Bx)$

$\psi(x)$ feature map

$K(x, \tilde{x}) = \langle \psi(x), \psi(\tilde{x}) \rangle$

$\Sigma(x, \tilde{x})$: NNGP

$$f_W(x) = A \frac{1}{\sqrt{K}} \phi(Bx) \qquad W = \begin{bmatrix} A \\ B_{K\times d} \end{bmatrix}$$

$$B_{ij} \sim N(0,1)$$

$$\Sigma(x,\tilde{x}) = \lim_{K\to\infty} \left\langle \frac{1}{\sqrt{K}} \phi(Bx), \frac{1}{\sqrt{K}} \phi(B\tilde{x}) \right\rangle$$

$$= \lim_{K\to\infty} \frac{1}{K} \sum_{i=1}^{K} \phi(B_{i,:}x)\, \phi(B_{i,:}\tilde{x})$$

random variable.

$B_{i,:} = [B_{i1}, \dots, B_{id}]$, $B_{ij} \sim N(0,1)$, $W \sim N(0, I_d)$

$$= \mathbb{E}_W [\phi(w^T x)\, \phi(w^T \tilde{x})]$$

$u = w^T x$, $v = w^T \tilde{x}$

$$= \mathbb{E}_{(u,v)\sim N(0,\Lambda)} [\phi(u)\phi(v)] \; ; \quad \Lambda = \begin{bmatrix} \|x\|_2^2 & x^T\tilde{x} \\ x^T\tilde{x} & \|\tilde{x}\|_2^2 \end{bmatrix}$$

$$= \frac{1}{2\pi(\det\Lambda)^{1/2}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \phi(u)\phi(v) \exp\left\{-\frac{1}{2}[u,v]\Lambda^{-1}\begin{bmatrix} u \\ v \end{bmatrix}\right\} du\,dv.$$

If $\|x\|_2 = \|\tilde{x}\|_2 = 1$, $\Sigma(x,\tilde{x}) = \underbrace{\check{\phi}(x^T\tilde{x})}_{\text{dual activation}} = \check{\phi}(\xi)$

$\xi = x^T\tilde{x}$

$$\Sigma(x, \tilde{x}) = \check{\phi}(x^T \tilde{x}) := \check{\phi}(\xi) \quad ; \quad \xi = x^T \tilde{x}.$$

Examples: $\phi(x) = x \Rightarrow \check{\phi}(\xi) = \xi$ ← $h_1(x)$

$$f_w(x) = \underbrace{A \frac{1}{\sqrt{k}} B}\, x$$

ReLU: $\phi(x) = \max(x, 0) \Rightarrow \check{\phi}(\xi) = \frac{1}{\pi}\left(\xi(\pi - \cos^{-1}(\xi)) + \sqrt{1-\xi^2}\right)$

RBF: $\phi(x) = e^{ix} \Rightarrow \Sigma(x, \tilde{x}) = \exp\{-L\|x - \tilde{x}\|_2^2\}$

$\hookrightarrow \check{\phi}(\xi) = \exp\{-L(2 - 2\xi)\}$

Hermite polynomials: $\phi(x) = h_i(x) \Rightarrow \check{\phi}(\xi) = \xi^i$

$h_0(x) = 1,$
$h_1(x) = x$
$h_2(x) = \frac{x^2 - 1}{\sqrt{2}}$

Properties: (1) $\phi(x) \Rightarrow \check{\phi}(\xi)$

$\alpha\phi(x) \Rightarrow \alpha^2 \check{\phi}(\xi)$

(2) $\phi(x) \Rightarrow \check{\phi}(\xi)$

$\phi'(x) \Rightarrow \check{\phi}'(\xi)$

Ex. $\phi(x) = \frac{x^2 - 1}{\sqrt{2}}$

$\check{\phi}(\xi) = \xi^2$

$\phi'(x) = \frac{2x}{\sqrt{2}}$

$\check{\phi}'(\xi) = 2\xi$

Neural Tangent Kernel (NTK)

changing

$$f_w(x) = A\frac{1}{\sqrt{K}}\phi(Bx) \iff \boxed{f_x(w)} = A\frac{1}{\sqrt{K}}\phi(Bx)$$

fixed

Taylor Series Expansion around initial weights w_0.

$f_x(w)$; $w \in \mathbb{R}$, w_0: $f_x(w) = f_x(w_0) + f_x'(w_0)(w - w_0)$

$w \in \mathbb{R}^p$, w_0: $f_x(w) = f_x(w_0) + \underbrace{\nabla f_x(w_0)^T}_{\mathbb{R}^{1\times p}}\underbrace{(w - w_0)}_{\mathbb{R}^p}$

$$f_x(w) \approx f_x(w_0) + \nabla f_x(w_0)^T (w - w_0)$$

$$\nabla f_x(w_0)^T w$$

$$\mathcal{L}(w) = \frac{1}{2}\sum_{i=1}^{n}\left(y^{(i)} - \underbrace{\nabla f_{x^{(i)}}(w_0)^T w}_{\psi(x^{(i)})}\right)^2$$

Kernel Regression

$\psi : x \longmapsto \nabla f_x(w_0)$

$\langle \psi(x), \psi(\tilde{x}) \rangle = \boxed{\langle \nabla f_x(w_0), \nabla f_{\tilde{x}}(w_0) \rangle}$ NTK.

Example: $f_x(w) = A \frac{1}{\sqrt{K}} \phi(Bx) = \frac{1}{\sqrt{K}} \sum_{i=1}^{K} A_i \phi(B_{i,:} x)$

$A \in \mathbb{R}^{1 \times K}$, $B \in \mathbb{R}^{K \times d}$, $\phi : \mathbb{R} \to \mathbb{R}$.

$$\nabla f_x(w_0) = \begin{bmatrix} \frac{\partial f}{\partial A_1} \\ \vdots \\ \frac{\partial f}{\partial A_K} \\ \frac{\partial f}{\partial B_{11}} \\ \vdots \\ \frac{\partial f}{\partial B_{Kd}} \end{bmatrix}_{w = w_0}$$

$$\frac{\partial f}{\partial A_i} = \frac{1}{\sqrt{K}} \phi(B_{i,:} x)$$

$$\frac{\partial f}{\partial B_{ij}} = \frac{1}{\sqrt{K}} A_i \phi'(B_{i,:} x) x_j$$

NTK

$$K(x, \tilde{x}) = \langle \nabla f_x(w_0), \nabla f_{\tilde{x}}(w_0) \rangle = \sum_{i=1}^{K} \frac{\partial f_x}{\partial A_i} \frac{\partial f_{\tilde{x}}}{\partial A_i} + \sum_{i=1}^{K} \sum_{j=1}^{d} \frac{\partial f_x}{\partial B_{ij}} \frac{\partial f_{\tilde{x}}}{\partial B_{ij}}$$

NTK

$$K(x,\tilde{x}) = \langle \nabla f_x(\omega_0), \nabla f_{\tilde{x}}(\omega_0) \rangle = \sum_{i=1}^{K} \frac{\partial f_x}{\partial A_i} \frac{\partial f_{\tilde{x}}}{\partial A_i} + \sum_{i=1}^{K} \sum_{j=1}^{d} \frac{\partial f_x}{\partial B_{ij}} \frac{\partial f_{\tilde{x}}}{\partial B_{ij}}$$

$$\frac{\partial f}{\partial A_i} = \frac{1}{\sqrt{K}} \phi(B_{i,:} x)$$

$$\frac{\partial f}{\partial B_{ij}} = \frac{1}{\sqrt{K}} A_i \phi'(B_{i,:} x)\, x_j$$

$$K(x,\tilde{x}) = \frac{1}{K} \sum_{i=1}^{K} \phi(B_{i,:} x)\, \phi(B_{i,:} \tilde{x}) \quad \} \text{ NNGP.} \qquad \phi(\xi)$$

$$+ \frac{1}{K} \sum_{i=1}^{K} \sum_{j=1}^{d} A_i^2\, \phi'(B_{i,:} x)\, \phi'(B_{i,:} \tilde{x})\, x_j \tilde{x}_j$$

Red: $\frac{1}{K} \sum_{i=1}^{K} A_i^2 \phi'(B_{i,:} x) \phi'(B_{i,:} \tilde{x}) \simeq \check{\phi}'(\xi)$ $\sum_{j=1}^{d} x_j \tilde{x}_j = x^\top \tilde{x}$

$$K(x,\tilde{x}) = \frac{1}{K}\sum_{i=1}^{K} \phi(B_{i,:}\, x)\, \phi(B_{i,:}\, \tilde{x})$$

$$+ \frac{1}{K}\sum_{i=1}^{K} \sum_{j=1}^{d} A_i^2\, \phi'(B_{i,:}\, x)\, \phi'(B_{i,:}\, \tilde{x})\, x_j\, \tilde{x}_j$$

$$\lim_{K\to\infty} K(x,\tilde{x}) = \check{\phi}(\xi) + \lim_{K\to\infty} \frac{1}{K}\sum_{i=1}^{K} A_i^2\, \check{\phi}'(\xi)\, x^T\tilde{x}$$

$$\mathbb{E}[A_i^2] = 1 \quad A_i \sim N(0,1)$$

$A_i \sim N(0,1)$

$B_{ij} \sim N(0,1)$

$$= \check{\phi}(\xi) + \check{\phi}'(\xi)\,\xi$$

$\check{\phi}(\xi)$ → NNGP

$\check{\phi}'(\xi)\,\xi$ → Correction

Depth L net:

$$K_L(x,\tilde{x}) = \check{\phi}^{(L)}(\xi) + \check{\phi}'\left(\check{\phi}^{(L-1)}(\xi)\right) K_{L-1}(x,\tilde{x})$$

$\check{\phi}^{(L)}(\xi)$ → NNGP

$\check{\phi}'\left(\check{\phi}^{(L-1)}(\xi)\right) K_{L-1}(x,\tilde{x})$ → correction.

$$\check{\phi}^{(L)} = \check{\phi}(\check{\phi}(\cdots \check{\phi}(\xi)\cdots))$$

$L \to \infty$.

$\xi = x^T\tilde{x}$.

Neural Tangents

$$\frac{1}{K}\sum_{i=1}^{K} \phi'(B_{i,:}x)\,\phi'(B_{i,:}\tilde{x}) \xrightarrow{K\to\infty} \check{\phi}'(\xi)$$

$$f_x(w) \approx f_x(w_0) + \nabla f_x(w_0)^T (w - w_0)$$

$$w_0 \sim \mathcal{N}(0, I_p) \Big\}$$

$$f_{x^{(i)}}(w_0) + \nabla f_{x^{(i)}}(w_0)^T w = y^{(i)}$$

$$\nabla f_{x^{(i)}}(w_0)^T w = y^{(i)} - f_{x^{(i)}}(w_0) \Big\}$$

$$f_x(w) = A \frac{1}{\sqrt{K}} \phi(Bx)$$

Review:

Kernel Regression

$x \mapsto \psi(x)$ feature map

(1) $K_T = K(X, X)$

(2) $\alpha = y K_T^{-1}$

FC-NN

$f_W(x) = A \frac{1}{\sqrt{K}} \phi(Bx)$

$A \in \mathbb{R}^{1 \times K}$; $B \in \mathbb{R}^{K \times d}$

$\phi : \mathbb{R} \to \mathbb{R}$ elementwise activation

Ex: $\phi(x) = \max(x, 0)$ ReLU

Training: G.D. on $\mathcal{L}(W) = \frac{1}{2} \| y - f_W(X) \|_2^2$

NTK

$A_i, B_{ij} \sim N(0,1)$

NTK: $K(x, \tilde{x}) = \langle \nabla f_x(W_0), \nabla f_{\tilde{x}}(W_0) \rangle$; $\|x\|_2 = \|\tilde{x}\|_2 = 1$

$\lim_{K \to \infty} K(x, \tilde{x}) = \check{\phi}(x^T \tilde{x}) + \check{\phi}'(x^T \tilde{x}) \, x^T \tilde{x}$

$\check{\phi}(x^T \tilde{x})$: NNGP $\Sigma(x, \tilde{x})$

$\check{\phi}'(x^T \tilde{x}) \, x^T \tilde{x}$: correction term

$\check{\phi}$: dual activation; closed form for many activations ϕ.

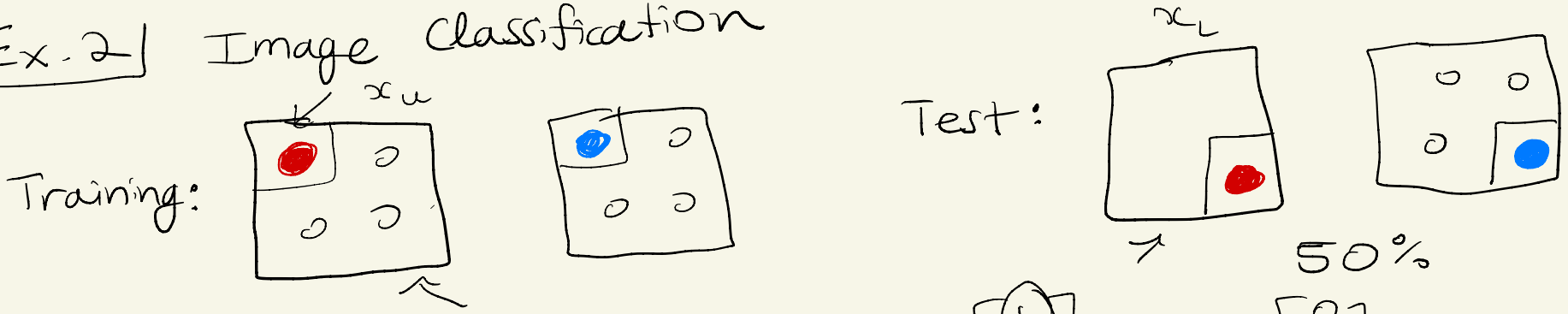
Motivating Problems

Tabular Datasets;

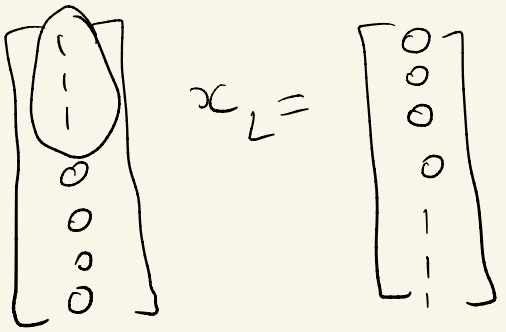
Ex. 1] Housing Price Prediction

$$x \begin{bmatrix} \#\text{ beds} \\ \#\text{ baths} \\ \text{sq. ft} \\ \vdots \end{bmatrix} \longrightarrow \$ \text{ House} \quad y$$

Ex. 2] Image Classification



NNGP: $x^T \tilde{x} = 0.$



$$x_u^T x_L = 0.$$

$$[1,1,1,0,0,0,0] \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} = 0.$$

1D - Convolutional Nets.

Ex. 1

	Training	Test
x_R	$[1\ 0\ 0\ 0]$	$[0\ 0\ 0\ 1]\ \tilde{x}_R$
x_B	$[-1\ 0\ 0\ 0]$	$[0\ 0\ 0\ -1]\ \tilde{x}_B$

$$\sum_{i=1}^{4} (A * x_R)_i = 1/2$$

$$\sum_{i=1}^{4} (A * \tilde{x}_R)_i = \frac{0+0+1+1}{4} = \frac{1}{2}$$

x_R: $[0\ 1\ 0\ 0\ 0\ 0]$

Filter: $[1\ \underline{1}\ 1] : A$

$$(A * x_R) = \Big[\underbrace{[1\ 1\ 1]\begin{bmatrix}0\\1\\0\end{bmatrix}}_{1},\ \underbrace{[1\ 1\ 1]\begin{bmatrix}1\\0\\0\end{bmatrix}}_{1},\ \underbrace{[1\ 1\ 1]\begin{bmatrix}0\\0\\0\end{bmatrix}}_{0},\ \underbrace{[1\ 1\ 1]\begin{bmatrix}0\\0\\0\end{bmatrix}}_{0}\Big]$$

$$= [1, 1, 0, 0]$$

$$(A * x_B) = [-1, -1, 0, 0]$$

$$(A * x)_i = \sum_{\alpha=-\frac{q-1}{2}}^{\frac{q-1}{2}} A_\alpha x_{i+\alpha}$$

$$A = [A_{-1}, A_0, A_1]$$

$$q = 3$$

$$(A * x_R)_i = \sum_{\alpha=-1}^{1} A_\alpha x_{i+\alpha} = A_{-1} x_{i-1} + A_0 x_i + A_1 x_{i+1}$$

$$i = 0 \Rightarrow A_{-1} \cdot 0 + A_0 \cdot 1 + A_1 \cdot 0 = A_0 = 1.$$

Def: 1D conv. $*$; $A \in \mathbb{R}^q$ $\quad A = [A_{-\frac{q-1}{2}} \cdots A_0 \cdots A_{\frac{q-1}{2}}]$

$x \in \mathbb{R}^d \Rightarrow (A * x) \in \mathbb{R}^d$; $(A * x)_i := \sum_{\alpha=-\frac{q-1}{2}}^{\frac{q+1}{2}} A_\alpha x_{i+\alpha}$.

Ex: $A = [1\ 1\ 1]$; $x = [0\ 0\ 0\ 1]$, $q = 3$, $\alpha \in \{-1, 0, 1\}$

$$(A*x)_0 = A_{-1} \cdot \underbrace{x_{-1}}_{=0} + A_0 x_0 + A_1 x_1$$

$$= 1 \cdot 0 + 1 \cdot 0 + 1 \cdot 0 = 0.$$

$$(A*x) = [0\ 0\ 1\ 1].$$

1D-CNN : (1) $f_\omega(x) = \frac{1}{d} \sum_{i=1}^{d} (A * x)_i$; Linear 1-filter CNN

$\rightarrow$ (2) $f_\omega(x) = \frac{1}{d} \sum_{i=1}^{d} \phi(A * x)_i$; Nonlinear 1-filter CNN.

$\rightarrow$ (3) $f_\omega(x) = \frac{1}{d} \sum_{i=1}^{d} \sum_{l=1}^{k} \phi(A^{(l)} * x)_i$; k-filter Nonlinear CNN.

(4) $f_\omega(x) = \sum_{l=1}^{k} \underbrace{B^{(l)}}_{\mathbb{R}^{1\times d}} \underbrace{\phi(A^{(l)} * x)}_{\mathbb{R}^d}$; k-filter Nonlinear FC NN

1D-CNN : (1) $f_w(x) = \frac{1}{d}\sum_{i=1}^{d} (A * x)_i$; Linear 1-filter CNN

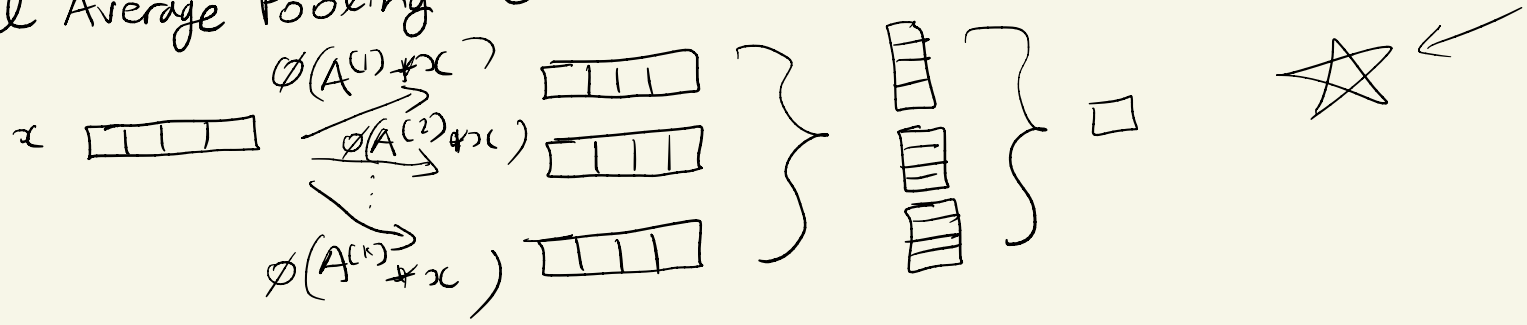
→ (2) $f_w(x) = \frac{1}{d}\sum_{i=1}^{d} \phi(A * x)_i$; Nonlinear 1-filter CNN.

→ { → (3) $f_w(x) = \frac{1}{d}\sum_{i=1}^{d}\sum_{l=1}^{k} \phi(A^{(l)} * x)_i$; k-filter Nonlinear CNN.

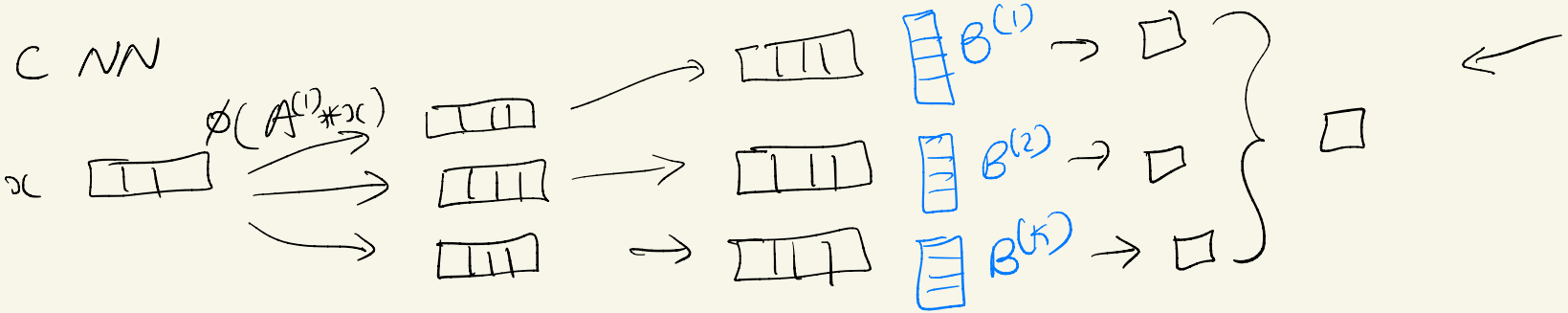
→ { (4) $f_w(x) = \sum_{l=1}^{k} \underbrace{B^{(l)}}_{\mathbb{R}^{1\times d}} \underbrace{\phi(A^{(l)} * x)}_{\mathbb{R}^{d}}$; k-filter Nonlinear FC NN

Diagrams:

(1) Global Average Pooling CNN : GAP-CNN.



(2) FC NN



$(A * x) \in \mathbb{R}^d$, $x \in \mathbb{R}^d$

Ex: $A = [A_{-1} \quad A_0 \quad A_1]$ $\quad x \in \mathbb{R}^4$

$$\begin{bmatrix} A_0 & A_1 & 0 & 0 \\ A_{-1} & A_0 & A_1 & 0 \\ 0 & A_{-1} & A_0 & A_1 \\ 0 & 0 & A_{-1} & A_0 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} = A * x$$

Toeplitz.

FC.

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ \vdots & & & \vdots \\ A_{41} & \cdots & & A_{44} \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

CNTK:

GAP-CNN $= f_w(x) = \frac{1}{d}\sum_{i=1}^{d}\sum_{l=1}^{K}\frac{1}{\sqrt{K}}\phi(A^{(l)} * x)_i$

$$= \frac{1}{d}\sum_{i=1}^{d}\sum_{l=1}^{K}\frac{1}{\sqrt{K}}\phi\left(\sum_{\alpha} A_{\alpha}^{(l)} x_{i+\alpha}\right) \quad \star$$

$$\nabla f_x(w_0) = \left[\frac{\partial f(w_0)}{\partial A^{(1)}_{-\frac{q-1}{2}}}, \ldots, \frac{\partial f(w_0)}{\partial A^{(K)}_{\frac{q-1}{2}}}\right] \qquad A_\alpha^{(l)} \sim N(0,1)$$

$$\frac{\partial f(w_0)}{\partial A^{(l)}_{j}} = \frac{1}{d}\sum_{i=1}^{d}\frac{1}{\sqrt{K}}\phi'\left(\sum_{\alpha} A_{\alpha}^{(l)} x_{i+\alpha}\right) x_{i+j}$$

$$K(x,\tilde{x}) = \sum_{l=1}^{K}\sum_{j=-\frac{q-1}{2}}^{\frac{q+1}{2}} \frac{\partial f_x(w_0)}{\partial A^{(l)}_j}\,\frac{\partial f_{\tilde{x}}(w_0)}{\partial A^{(l)}_j}$$

$$= \sum_{l=1}^{K}\sum_{j} \frac{1}{d^2}\frac{1}{K}\sum_{i=1}^{d}\sum_{i'=1}^{d}\phi'(A^{(l)} * x)_i\, x_{i+j}\cdot\phi'(A^{(l)} * \tilde{x})_{i'}\,\tilde{x}_{i'+j}$$

As $K \to \infty$ $= \frac{1}{d^2}\sum_{i=1}^{d}\sum_{i'=1}^{d}\{\phi'(x[i], \tilde{x}[i'])\langle x[i], \tilde{x}[i']\rangle\}$

$x[i] : [x_{i-1}, x_i, x_{i+1}]$

$$K(x,\tilde{x}) = \sum_{l=1}^{K} \sum_{j=-\frac{q-1}{2}}^{\frac{q+1}{2}} \frac{\partial f_x(\omega_0)}{\partial A_{ij}^{(l)}} \frac{\partial f_{\tilde{x}}(\omega_0)}{\partial A_{ij}^{(l)}}$$

$$= \left(\sum_{l=1}^{K}\right) \sum_{j} \frac{1}{d^2}\frac{1}{K} \sum_{i=1}^{d}\sum_{i'=1}^{d} \phi'(A^{(l)} * x)_{i,i+j} \cdot \phi'(A^{(l)} * \tilde{x})_{i',i'+j}$$

$$= \frac{1}{K}\sum_{l=1}^{K} \frac{1}{d^2} \sum_{i=1}^{d}\sum_{i'=1}^{d} \phi'(A^{(l)} * x)_{i,} \phi'(A^{(l)} * \tilde{x})_{i',} \sum_{j=-\frac{q-1}{2}}^{\frac{q+1}{2}} x_{i,i+j}\, \tilde{x}_{i',i'+j}$$

$$x[i,] = [x_{i,-\frac{q-1}{2}}, \ldots, x_{i,+\frac{q-1}{2}}]$$

$$\tilde{x}[i',] = \begin{bmatrix} \tilde{x}_{i',-\frac{q-1}{2}} \\ \vdots \\ \tilde{x}_{i',+\frac{q-1}{2}} \end{bmatrix}$$

$$= \frac{1}{d^2} \sum_{i=1}^{d}\sum_{i'=1}^{d} \left(\frac{1}{K}\sum_{l=1}^{K} \phi'(A^{(l)} * x)_{i,} \phi'(A^{(l)} * \tilde{x})_{i',}\right) < x[i,], \tilde{x}[i',] >$$

As $K \to \infty$, $\phi'(x[i,], \tilde{x}[i',])$

$$= \frac{1}{d^2} \sum_{i=1}^{d}\sum_{i'=1}^{d} \phi''(x[i,], \tilde{x}[i',]) < x[i,], \tilde{x}[i',] >.$$

$$K_{GAP}(x, \tilde{x}) = \frac{1}{d^2} \sum_{i=1}^{d} \sum_{i'=1}^{d} \dot{\phi}'(x[i], \tilde{x}[i']) \, \langle x[i], \tilde{x}[i'] \rangle.$$

$$K_{FC}(x, \tilde{x}) = \frac{1}{d} \sum_{i=1}^{d} \dot{\phi}'(x[i], \tilde{x}[i]) \, \langle x[i], \tilde{x}[i] \rangle. \}$$

Role of Increasing Depth in CNNs

$$f_w(x) = W_1 * (W_2 * (W_3 * x))$$

$$W_3 = \begin{bmatrix} \cdot & \cdot & & & 0 \\ \cdot & \cdot & \cdot & & \\ & \cdot & \cdot & \cdot & \\ 0 & & \cdot & \cdot & \cdot \end{bmatrix}$$

$$\underbrace{W_1 W_2 W_3} = \underbrace{\begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}}$$

Review:

(1) Linear Regression: $\omega X = y$ $\quad \|\omega X - y\|_2^2$ $\quad \omega = yX^{\dagger}$

(2) Kernel Regression: $x \mapsto \psi(x)$, $x \in \mathbb{R}^d$, $\psi(x) \in \mathbb{R}^p$

(1) $K_T = K(X, X)$

$K(x, z) = \langle \psi(x), \psi(z) \rangle$

(2) $\alpha = y K_T^{-1}$

$\|y - f_\omega(x)\|_2^2$ using G.D.

(3) Neural Nets: $f_\omega(x) = A\,\phi(Bx)$

---

(4) NNGP & NTK connected inf. wide neural nets with kernel methods.

NTK $\Rightarrow$ $K(x, z) = \langle \nabla f_x(\omega_0), \nabla f_z(\omega_0) \rangle$

$\|x\|_2 = \|z\|_2 = 1 \Rightarrow K(x, z) = \check{\phi}(\xi) + \xi \check{\phi}'(\xi)$

where $\xi = x^T z$.

Course challenge:

	Test R^2
NTK: .2, .3	.98
Laplace: < .3	

$f^*(x) = x_1 x_2$

$M = W_1^T W_1 =$ $\begin{bmatrix} 1 & & 0 \\ & 1 & \\ 0 & & 0 \end{bmatrix}$ (100 × 100)

Expected Gradient Outer Product

$$\nabla f^*(x) = \begin{bmatrix} \frac{\partial f^*}{\partial x_1} \\ \frac{\partial f^*}{\partial x_2} \\ 0 \\ \vdots \end{bmatrix} = \begin{bmatrix} x_2 \\ x_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\nabla f^*(x) \; \nabla f^*(x)^T = \begin{bmatrix} x_2 \\ x_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} x_2 & x_1 & 0 & \cdots & 0 \end{bmatrix} = \begin{bmatrix} x_2^2 & x_2 x_1 & & \\ x_1 x_2 & x_1^2 & & \\ & & 0 & \\ & & & 0 \end{bmatrix}$$ (100 × 100)

$x_i \sim \mathcal{N}(0,1)$

$$G(f^*) = \underbrace{\mathbb{E}_x \left[\nabla f^*(x) \; \nabla f^*(x)^T \right]}_{} = \begin{bmatrix} 1 & 0 & & \\ 0 & 1 & & \\ & & 0 & \\ & & & 0 \end{bmatrix}$$

$f^* \approx \hat{f} \; ; \; G(f^*) \approx G(\hat{f})$

Goal: $f(x) = A \frac{1}{\sqrt{K}} Bx$; $A \in \mathbb{R}^{1 \times K}$, $B \in \mathbb{R}^{K \times d}$

$$B^{(1)T} B^{(1)} = \nabla f^{(1)}(x) \nabla f^{(1)}(x)^T$$

$$A^{(1)} = A^{(0)} + \eta \left(y - A^{(0)} \frac{1}{\sqrt{K}} B^{(0)} x\right) x^T B^{(0)T} \frac{1}{\sqrt{K}} \quad (1)$$

$$B^{(1)} = B^{(0)} + \eta A^{(0)T} \left(y - A^{(0)} \frac{1}{\sqrt{K}} B^{(0)} x\right) x^T \frac{1}{\sqrt{K}} \quad (2)$$

$B^{(0)} = 0$, $A_i^{(0)} \sim N(0,1)$

(1) $A^{(1)} = A^{(0)}$

(2) $B^{(1)} = \eta A^{(0)T} y x^T \frac{1}{\sqrt{K}}$

$$B^{(1)T} B^{(1)} = \eta x y^T A^{(0)} \frac{1}{\sqrt{K}} \cdot \frac{1}{\sqrt{K}} \eta A^{(0)T} y x^T$$

$$= \eta^2 x y^T \frac{1}{K} A^{(0)} A^{(0)T} y x^T$$

$$\frac{1}{K} A^{(0)} A^{(0)T} \hookrightarrow \frac{1}{K} \sum_{i=1}^{K} A_i^{(0)2}$$

$$\lim_{K \to \infty} B^{(1)T} B^{(1)} = \eta^2 x y^T y x^T$$

① $A^{(1)} = A^{(0)}$ ② $B^{(1)} = \eta A^{(0)T} y x^T \frac{1}{\sqrt{K}}$

$$\lim_{K\to\infty} B^{(1)T} B^{(1)} = \eta^2 x y^T y x^T$$

$$f^{(1)}(x) = A^{(1)} \frac{1}{\sqrt{K}} B^{(1)} x \; ; \; \nabla f^{(1)}(x) = B^{(1)T} \frac{1}{\sqrt{K}} A^{(1)T}$$

$$\nabla f^{(1)}(x) \nabla f^{(1)}(x)^T = B^{(1)T} \frac{1}{\sqrt{K}} A^{(1)T} A^{(1)} \frac{1}{\sqrt{K}} B^{(1)}$$

$$= \underbrace{\eta x y^T A^{(0)} \frac{1}{\sqrt{K}}}_{B^{(1)T}} \cdot \frac{1}{\sqrt{K}} A^{(0)T} \cdot A^{(0)} \frac{1}{\sqrt{K}} \underbrace{\frac{1}{\sqrt{K}} \eta A^{(0)T} y x^T}_{B^{(1)}}$$

$$= \eta^2 x y^T \underbrace{\left(\frac{1}{K} A^{(0)} A^{(0)T}\right)}_{\to 1 \text{ as } K\to\infty} \underbrace{\left(\frac{1}{K} A^{(0)} A^{(0)T}\right)}_{\to 1 \text{ as } K\to\infty} y x^T$$

$$\lim_{K\to\infty} \nabla f^{(1)}(x) \nabla f^{(1)}(x)^T = \eta^2 x y^T y x^T$$

After 1-step: $f^{(1)}(x) \leftarrow$

$$B^{(1)\top} B^{(1)} \equiv \nabla f^{(1)}(x)\, \nabla f^{(1)}(x)^\top$$

Recursive Feature Machines (RFMs)

(1) $\hat{f}(M_t^{1/2} x) \approx f^*$ ← kernels for $\hat{f}$ $\quad M^{(0)} = I$.

(2) $M_t = \mathbb{E}_x\left[\nabla \hat{f}(x)\, \nabla \hat{f}(x)^\top\right]$

Algorithm: $K(x, z) = \exp\left\{-L \underbrace{\|x - z\|_2}_{\text{Euclidean dist.}}\right\}$

$$K_M(x, z) = K(M^{1/2} x, M^{1/2} z)$$

$$= \exp\left\{-L \|M^{1/2} x - M^{1/2} z\|_2\right\}$$

$$= \exp\left\{-L \underbrace{\sqrt{(x - z)^\top M (x - z)}}_{\text{Mahalanobis distance.}}\right\}$$

Algorithm: $M = I$

$$K_M(x, z) = \exp\left\{-L\sqrt{(x-z)^T M (x-z)}\right\}$$

(1) Solve kernel regression with K_M

$$K_{train} = K_M(X, X)$$

$$\alpha = y K_{train}^{-1}$$

$$\hat{f}(x) = \alpha K_M(X, x)$$

$\hat{f}$

(2) $M = \frac{1}{n} \sum_{x^{(p)} \in X} \nabla \hat{f}(x^{(p)}) \nabla \hat{f}(x^{(p)})^T$ $\Big\}$ $G(\hat{f})$

(3) Iterate steps 1 & 2.

$x \in \mathbb{R}^{100}$ $\quad f^*(x) = x_1 x_2$ $\Big\}$

$$K(x, z) = \exp\left\{-L \|x - z\|_2\right\}$$

$$f^*(x) = g(u^T x)$$

$$\nabla f^*(x) = g'(u^T x)\, u$$

$$\mathbb{E}_x\left[\nabla f^*(x)\, \nabla f^*(x)^T\right] = \mathbb{E}_x\left[g'(u^T x)^2\, u u^T\right]$$

$$= \underbrace{\mathbb{E}_x\left[g'(u^T x)^2\right]}_{C'} u u^T$$

$$= C' \boxed{u u^T} \leftarrow$$

$u \in \mathbb{R}^{50}$

$g(u^T x) \rightarrow C' u u^T$

$$\Sigma = \begin{bmatrix} 1 & \xi \\ \xi & 1 \end{bmatrix}$$

$$\exp\left\{-\frac{1}{2(1-\xi^2)} \begin{bmatrix} u & v \end{bmatrix} \begin{bmatrix} 1 & -\xi \\ -\xi & 1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}\right\}$$

$$\begin{bmatrix} u - v\xi, & -u\xi + v \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

$$u^2 - uv\xi - uv\xi + v^2 = u^2 + v^2 - 2uv\xi$$