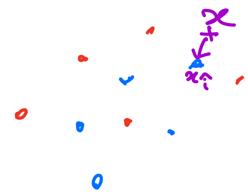
Kernelized 1-Nearest Neighbor

$$S = \{(x_i, y_i)\}_{i=1}^{N}$$

standard 1-NN: $f_{1\text{-NN}}(x) = y_{\hat{i}}$ such that $\hat{i} = \underset{i \in \{1..n\}}{\arg\min} \|x_i - x\|$



$f(x) = \text{blue}$

Assume we want to compute the 1-NN in the space $\{(\phi(x_i), y_i)\}_{i=1}^{N}$

We need a distance in the ϕ space.

$$\|\phi(x) - \phi(x')\|^2 = (\phi(x) - \phi(x'))^T(\phi(x) - \phi(x'))$$
$$= \phi(x)^T\phi(x) + \phi(x')^T\phi(x') - 2\phi(x)^T\phi(x')$$
$$= k(x,x) + k(x',x') - 2k(x,x')$$

simply compute $\hat{i} = \arg\min k(x_i, x_i) + k(x, x) - 2k(x_i, x)$

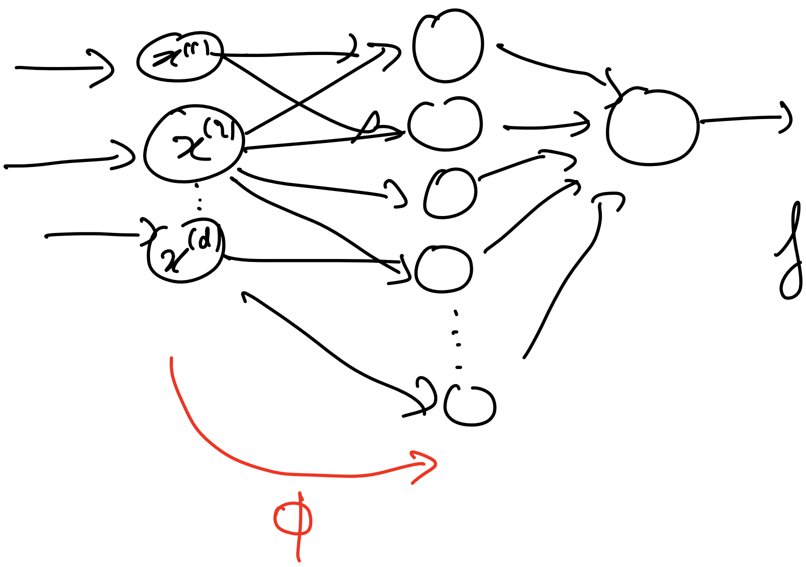
$f_{1\text{-NN kernelized}}(x) = y_{\hat{i}}$

Relation between Neural Nets and Kernels

1) Random Feature Model (RFM)
2) Neural tangent Kernels (NTK)

1) RFM

Consider a simple neural net:



relu or else

$$f(x) = \frac{1}{\sqrt{K}} \sum_{j=1}^{K} a_j \, \sigma(B_j^T x)$$

$a \in \mathbb{R}^K$

$B \in \mathbb{R}^{d \cdot K}$

$x \in \mathbb{R}^d$

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

Assume B is unchanged, we train only on a.

We can see $f(x) = a^T \Phi(x)$

where $\Phi(x) = \begin{pmatrix} \frac{1}{\sqrt{K}} \sigma(B_1^T x) \\ \vdots \\ \frac{1}{\sqrt{K}} \sigma(B_K^T x) \end{pmatrix}$

Assume $K \to \infty$.

What is the underlying kernel?

$$k(x, x') = \lim_{K \to \infty} \langle \phi(x), \phi(x') \rangle$$

$$= \lim_{K \to \infty} \frac{1}{K} \sigma(B^T x)^T \sigma(B^T x')$$

$$\sigma(B^T x) = \begin{pmatrix} \sigma(B_1^T x) \\ \vdots \\ \sigma(B_K^T x) \end{pmatrix}$$

$$= \mathbb{E}_{\omega \sim \mathcal{N}(0, I_d)} \left[\sigma(\omega^T x)^T \sigma(\omega^T x') \right]$$

$$= \mathbb{E}_{(u,v) \sim \mathcal{N}(0, \Lambda)} \left[\sigma(u)^T \sigma(v) \right]$$

$$\Lambda = \begin{bmatrix} \|x\|^2 & x^T x' \\ x^T x' & \|x'\|^2 \end{bmatrix}$$

For Relu:

$$k(x, x') = \frac{1}{\pi} \left((x^T x') (\pi - \arccos (x^T x')) + \sqrt{1 - (x^T x')^2} \right)$$

2) NTK

Idea: • lift the assumption over freezed weights.
All weights are learnt
• applys to any sufficiently large network.

An arbitrary NN is a function

$$f(x; w) \to \mathbb{R}$$

↳ whole weights.

Define $f_x(\omega) = f(x, \omega)$

let ω_0 be the initial weight set.

$$f_x(\omega) = \underbrace{f_x(\omega_0) + \langle \omega - \omega_0 \rangle^T \nabla_{\omega_0} f_x(\omega_0) + \text{higher order terms}}_{\text{Taylor decomposition.}}$$

For simplicity, assume $f_x(\omega_0) = 0$

At neighborhood of ω_0, we have:

$$f_x(\omega) \approx \langle \omega - \omega_0 \rangle^T \nabla_{\omega_0} f_x(\omega_0)$$

$$= \omega^T \phi(x) + \text{cste}$$

with $\phi(x) = \nabla_{\omega_0} f_x(\omega_0)$

$$k(x, x') = \langle \nabla_{\omega_0} f_x(\omega_0), \nabla_{\omega_0} f_{x'}(\omega_0) \rangle$$