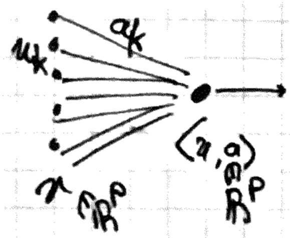
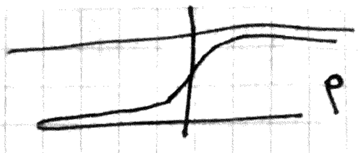
Multilayer Perceptron with 1 Hidden Layer

0 hidden layer (Perceptron)

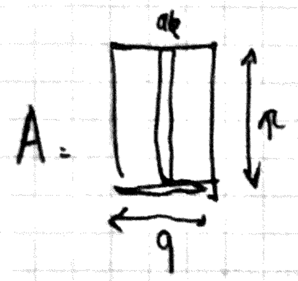


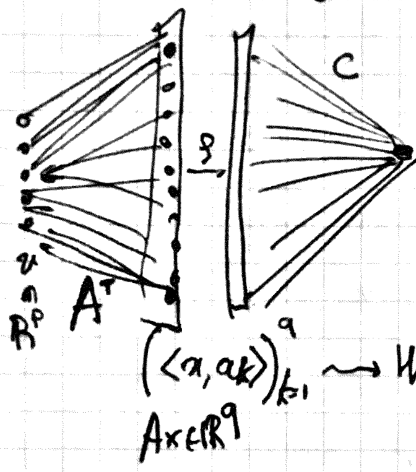
$\longrightarrow \varphi(\langle x, a\rangle) \in \mathbb{R}$ ex. $\varphi(x) = \dfrac{e^x}{1+e^x}$



$\Rightarrow$ logistic regression, convex.

1 hidden layer



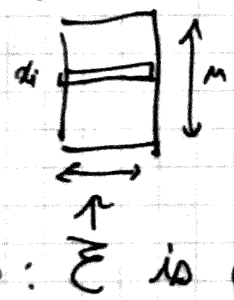


$$f(x) = \sum_{k=1}^{q} c_k \, \varphi(\langle x, a_k\rangle) = \langle \varphi(A^T x), c\rangle$$

Rmq: include a bias $\quad \underset{\in\mathbb{R}^p}{\langle x, a_k\rangle} + \underset{\in\mathbb{R}}{b_k} = \langle \underbrace{[x;1]}_{\in\mathbb{R}^{p+1}}, [a_k, b_k]\rangle$

Training (Regression): $\displaystyle \min_{A,c} \mathcal{E}(A,c) \triangleq \frac{1}{2n}\sum_{i=1}^{n} |f(x_i) - y_i|^2 = \frac{1}{n}\sum_{i=1}^{n} |\langle \varphi(A^T x_i), c\rangle - y_i|$

$$\mathcal{E} = \frac{1}{2n} \| \varphi(\underset{\in\mathbb{R}^{n\times p}}{X} \times \underset{\in\mathbb{R}^{p\times q}}{A}) \, \underset{\in\mathbb{R}^{q}}{c} - \underset{\in\mathbb{R}^{n}}{y} \|^2$$



Pbm: $\mathcal{E}$ is convex wrt c but non convex wrt A ...

Gradient descent: $\nabla_c \mathcal{E}(A,c) = \varphi(XA)^T \big(\underbrace{\varphi(XA)c - y}_{\triangleq R \in \mathbb{R}^n} \big)$

$$\nabla_A \mathcal{E}(A,c) = \underbrace{X^T}_{\in\mathbb{R}^{p\times n}} \big[\underset{\in\mathbb{R}^{n\times q}}{\varphi'(XA)} \odot (\underset{\in\mathbb{R}^{n}}{R}\, \underset{\in\mathbb{R}^{q}}{c^T}) \big] \in \mathbb{R}^{p\times q}$$

Approximation theory : $f(x) = \langle c, \varphi([A;b]^T[x;1])\rangle$

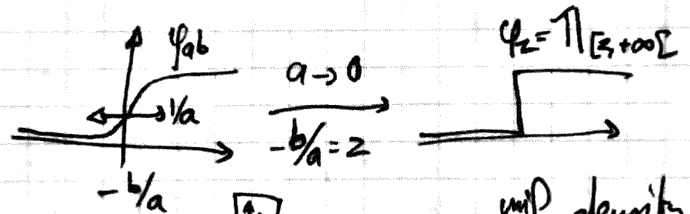
$$= \sum_{k=1}^{q} c_k \underbrace{\varphi(\langle a_k, x\rangle + b_k)}_{\triangleq\ \varphi_{a_k,b_k}(x)}.$$

$\rightsquigarrow$ linear approx° in an adaptive dictionary $(\varphi_{a_k b_k})_{k=1}^{q}$ of q atoms

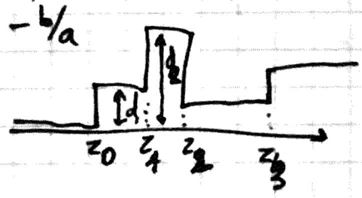
Thm : [Cybenko, 1989] "Universality" Let $\Omega \subset \mathbb{R}^n$ be compact, $\varphi : \mathbb{R} \to \mathbb{R}$ continuous and bounded, $\varphi(x) \begin{cases} \xrightarrow{x\to+\infty} 1 \\ \xrightarrow{x\to-\infty} 0 \end{cases}$

Then finite sums $\sum_{k=1}^{q} a_k \varphi_{a_k b_k}$ are dense in $L^\infty(\Omega)$.

Intuition in 1-D, $n=1$



$\sum_{k=1}^{q} d_k(\varphi_{z_k} - \varphi_{z_{k-1}}) =$



$\longrightarrow$ unif. density of piecewise constant f° in $L^\infty(\Omega)$ (unif. continuity on compact)

Rmq : Thm means that $\forall f \in \mathcal{C}(\Omega), \forall \varepsilon > 0, \exists q, \exists A, b, c$ tq.

$$\max_{x\in\Omega} |f(x) - \sum_{k=1}^{q} c_k \varphi(\langle x, a_k\rangle - b_k)| \leq \varepsilon$$

$\rightsquigarrow$ Pbm : how does q depends on f and n (convergence speed)

$\rightsquigarrow$ Require smoothness hypothesis on f

Thm : [Barron 93] : if $\int \|\omega\| |\hat{f}(\omega)| d\omega \leq C$ then

$$\int_{\|x\|\leq r} |f(x) - \sum_{k=1}^{q} c_k \varphi(\langle x, a_k\rangle - b_k) + c_0|^2 dx \leq \mathrm{Vol}(B(0,r)) \times \frac{(2rC)^2}{q}$$

Proof of the universality theorem of Cybenko.

Prop 1: If φ is such that

$$\left[\forall (a,b) \int_\Omega \varphi(\langle a,x\rangle + b)\, d\mu = 0\right] \Rightarrow \mu = 0 \qquad \text{(Discr)}$$

$\mu \in \mathcal{M}_b(\Omega) = \mathcal{C}(\Omega)^*$ Radon measures, bounded $\mu(\Omega)<+\infty$

Then the universality thm is true.

Proof: Let $\mathcal{S} = \left\{ \sum_k c_k \varphi_{a_k b_k} : \begin{matrix} q\in\mathbb{N} \\ a_k\in\mathbb{R}^p \\ b_k\in\mathbb{R} \end{matrix} \right\} \subset \mathcal{C}(\Omega)$ is a linear space

Let $\bar{\mathcal{S}}$ be its closure in $\mathcal{C}(\Omega)$ for $\|.\|_\infty$ (wich is a Banach sp)

If $\bar{\mathcal{S}} \neq L^\infty(\Omega)$, pick $g \neq 0$, $g \in L^\infty(\Omega) \setminus \bar{\mathcal{S}}$

We define a linear ~~operator~~ form L on $\bar{\mathcal{S}} \oplus \mathrm{Span}(g)$ by

$\forall s \in \bar{\mathcal{S}}, \quad L(s + \lambda g) = \lambda \qquad$ (so $L \equiv 0$ on $\bar{\mathcal{S}}$)

L is a bounded linear form, so by Hahn-Banach theorem, it can be extended in a bounded linear form $\bar{L} : \mathcal{C}(\Omega) \to \mathbb{R}$

Since $\bar{L} \in \mathcal{C}(\Omega)^* \sim \mathcal{M}(\Omega)$, it can be represented as

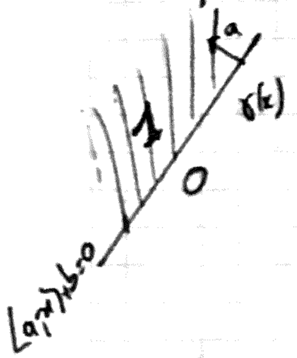
$$\bar{L}(f) = \int_\Omega f(x)\, d\mu(x) \qquad \text{with } \mu \neq 0$$

But $\bar{L} \equiv 0$ on $\bar{\mathcal{S}}$, so $\int_\Omega \varphi_{ab}\, d\mu = 0 \ \forall (a,b)$, so by (Discr), $\mu = 0$, contradiction ∎

Prop 2: if $\varphi \underset{+\infty}{\overset{-\infty}{\rightrightarrows}} \begin{matrix} 0 \\ 1 \end{matrix}$ is continuous, then it satisfies (DiscR).

Proof:

$$\varphi_{\lambda a, \lambda b + t}(x) = \varphi\big(\lambda(\langle a,x\rangle + b) + t\big) \xrightarrow{\lambda \to +\infty} \begin{cases} \overbrace{1 \text{ if } \langle a,x\rangle + b > 0}^{H_{ab}} \\ 0 \text{ if } \langle a,x\rangle + b < 0 \\ \underbrace{\varphi(t) \text{ if } \langle a,x\rangle + b = 0}_{\Pi_{ab}} \end{cases} \triangleq \gamma(x)$$



By Lebesgue dominated convergence, (bounded by 1 on compact set)

$$\int \varphi_{\lambda a, \lambda b + t}\, d\mu \xrightarrow{\lambda \to +\infty} \int \gamma\, d\mu = \varphi(t) \cdot \mu(\Pi_{ab}) + \mu(H_{ab})$$

where $\Pi_{ab} = \{x : \langle a,x\rangle + b = 0\}$ $H_{ab} = \{x : \langle a,x\rangle + b = 0\}$

If μ is such that $\int \varphi_{a',b'}\, d\mu = 0 \ \forall (a',b')$ then $\forall (a,b,t)$, $\varphi(t)\mu(\Pi_{ab}) + \mu(H_{ab}) = 0$

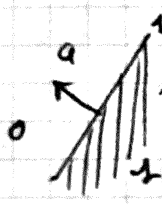
By selecting (t,t') such that $\varphi(t) \neq \varphi(t')$, one has that $\forall (a,b)$ $\begin{cases} \mu(H_{a,b}) = 0 \\ \mu(\Pi_{a,b}) = 0 \end{cases}$

One now needs to show that $\mu = 0$.

For $\begin{cases} h \in L^\infty(\mathbb{R}) \\ a \in \mathbb{R}^? \end{cases}$, let $F(h) \triangleq \int_\Omega h(\langle a, x\rangle)\, d\mu$.

$F : L^\infty(\mathbb{R}) \to \mathbb{R}$ is a bounded linear form since $\|F(h)\| \leq \|F\|_\infty \cdot \underbrace{\mu(\Omega)}_{<+\infty}$

Let $h = \mathbb{1}_{[b,+\infty[}$ then $F(\mathbb{1}_{[b,+\infty[}) = \int \mathbb{1}_{[b,+\infty[}\, d\mu = \mu(\{x : \langle a, x\rangle - b \geq 0\})$

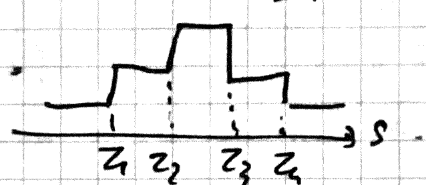


$\mathbb{1}_{[b,+\infty[}(\langle x, a\rangle)$

$$= \underset{0}{\underset{\shortparallel}{\mu(\Pi_{a,-b})}} + \underset{0}{\underset{\shortparallel}{\mu(H_{a,-b})}} = 0$$

Similarly as in 1D, for $h(s) = \sum c_i \left(\mathbb{1}^{(s)}_{[z_i,+\infty[} - \mathbb{1}^{(s)}_{[z_{i+1},+\infty[}\right)$

h :



One has $F(h) = 0$ for all piecewise constant f^o h.

By density in $L^\infty(\mathbb{R})$, $F(h) = 0$ $\forall h \in L^\infty(\mathbb{R})$.

Taking $h(x) = e^{ix}$, one has

$$\int \exp(i\langle x, a\rangle)\, d\mu(x) = \hat{\mu}(a) = 0 \quad \forall a.$$

By injectivity of the Fourier transform, $\mu = 0$ ■