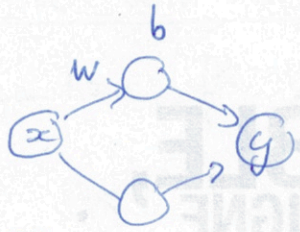
16.4.16

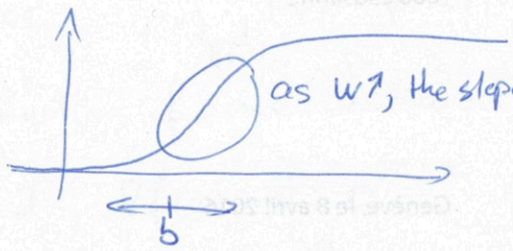
Prelim: A simple proof that a neural net can approximate any function

(↔ universality thm) (Michael Nielsen)

Take $\phi(z) = \frac{1}{1+e^{-z}}$ e.g.



- Top hidden neuron outputs $\phi(wx+b)$:

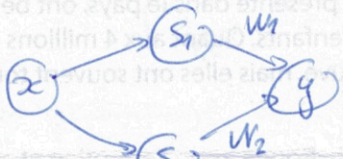


⟹ step fn, where you want $1_{x>s}$, $s = -\frac{b}{w}$

⟹ neuron idealization outputs $1_{x>s}$; single parameter s

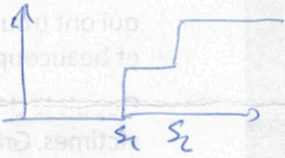
CNB: works for "every" activation fn $\phi(x) \xrightarrow[-\infty]{} 0$, $\phi(x) \xrightarrow[+\infty]{} 1$)

- Now:

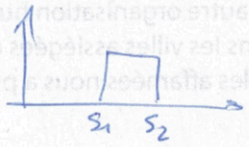


output $y = w_1 1_{x>s_1} + w_2 1_{x>s_2}$

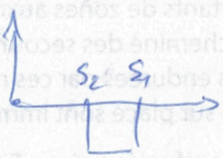
(forget the last ϕ + bias b)



$w_2 = -w_1$ ⟹

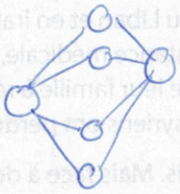


or

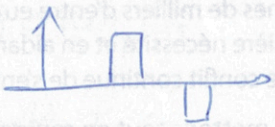


∈ adjustable height

- Two pairs of neurons:



⟹ two bumps:



- ⟹ staircase with 2n neurons ↔ n intervals

→ approx of any f

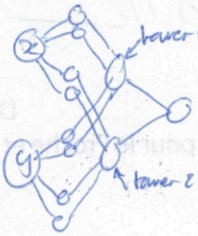
(cf. Riemann sums!)

Generalization to multivariate x:

towers

as

⚠ two hidden layers:



⚠ need h large & bias b ... (read again)

(?)

↳ one layer: circular towers!

enko: sigmoid functions can be used for approx: fine (as sines, polynomials)

ran. Preliminary on Fourier transform

I. What kind of function $f: \mathbb{R}^d \to \mathbb{R}$ will we be able to approximate?

- Consider first Γ, the set of functions f s.t. its FT defined as

$$f(x) = \int_{\mathbb{R}^d} e^{i\omega \cdot x} \tilde{f}(\omega)\, d\omega \qquad (*)$$

satisfies $\int_{\mathbb{R}^d} |\omega| \cdot |\tilde{f}(\omega)|\, d\omega < \infty$

NB: integrability of $\tilde{f}$ $\leftrightarrow$ regularity of f

Such functions f are continuously differentiable $\rightarrow$

(indeed, $\nabla f(x) = \int_{\mathbb{R}^d} e^{i\omega \cdot x}\, i\omega\, \tilde{f}(\omega)\, d\omega$ is well defined) $\forall x\ |\nabla f(x)| \le C$

- Then consider Γ_C for $C > 0$, the set of functions f

such that $C_f = \int_{\mathbb{R}^d} |\omega| \cdot |\tilde{f}(\omega)|\, d\omega \le C$. [examples later?] sect. IX

NB: can be generalized to $\tilde{f}(\omega)\, d\omega = \tilde{F}(d\omega)$ a measure, but we will skip this (bef)

- Note: even if C_f is finite, $(*)$ might be an ill defined integral

skip remedy to this: $f(x) - f(0) = \int_{\mathbb{R}^d} (e^{i\omega \cdot x} - 1)\, \tilde{f}(\omega)\, d\omega$

is well defined, as $|e^{i\omega \cdot x} - 1| \le |\omega \cdot x| \le |\omega| \cdot |x|$

II. With what functions f_n will we approximate f?

Let ϕ be a (fixed) bounded measurable fn on $\mathbb{R}$ with $\phi(z) \xrightarrow[z \to -\infty]{} 0$, $\phi(z) \xrightarrow[z \to +\infty]{} 1$.

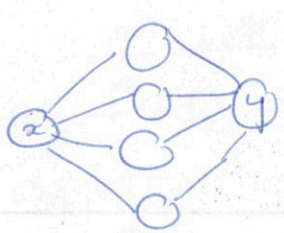
typically continuous!

Typically a cdf, even more typically a sigmoid: $\boxed{\phi(z) = \frac{1}{1+e^{-z}}}$

One layer network:

$$f_n(x) = \sum_{j=1}^{n} c_j\, \phi(a_j \cdot x + b_j) + c_0 \qquad (**)$$

($c_j \in \mathbb{R}$, $a_j \in \mathbb{R}^d$, $b_j \in \mathbb{R}$)



n nodes

In total: $(d+2)n+1$ parameters $(O(n))$

5. Theorem 1 (NB: B_r → general $B \ni 0$ possible bounded)

Let us consider: ϕ "a sigmoid"
$B_r = \{x \in \mathbb{R}^d : |x| \le r\}$, $r > 0$ } fixed
(think uniform or empirical meas.) μ a probability distribution on B_r

Then $\forall f \in \Gamma_C$ and $\forall n \ge 1$, $\exists f_n$ of the form (**)

such that $\int_{B_r} (f(x) - f_n(x))^2 \mu(dx) \le \frac{(2rC_f)^2}{n}$ (1)

& $\sum_{j=1}^{n} |c_j| \le 2C_f$

Why is this theorem important/interesting?

here, only $O(nd)$

- classical approx methods (polynomial, Fourier, ...)
  → "exponential" number of parameters in n^d for same error $O(\frac{1}{n})$
- fixed basis of $O(n)$ functions → error $O(\frac{1}{n^{2/d}})$ at least
  → curse of dimensionality! [proof later? sect. X]
  (⚠ C also depends on d, ~~boundedness~~)

(practical, low-complexity algorithm to find the parameters → next time!)

⊙ 2 proofs: not iterative / iterative

- generalization: $B_r \to B$

⊕ p. 932, col. 2:

= only approximation error here

~~also training error and else~~

$\mu = \frac{1}{N}\sum_{j=1}^{N} \delta_{x_j}$: $\frac{1}{N}\sum_{j=1}^{N} (f(x_j) - \hat{f}_n(x_j))^2 \le \frac{C_f'}{n}$

"training" error (?) → chosen to minimize the expression / or iterative

$\frac{C_f^2}{n} + \frac{nd \lg N}{N}$ cf. [11]!

⊕ computational complexity

(1) Note: $f \in \Gamma_C$ is defined on $\mathbb{R}^d$ (the whole space) differentiable
one can as well consider $f : B_r \to \mathbb{R}$ s.t. $\exists$ an extension $\tilde{f}$ of f which coincides on B_r and is in Γ_C.

Proof

Pisier: "remarks on an unpublished result by Maurey" 1980 (!)

Lemma 1 (NB: ref of Maurey & Pisier!)

Let H be a Hilbert space (scalar prod $\langle \cdot,\cdot \rangle$, corr. norm $\|\cdot\|$)

$G \subset H$ be a set with $\|g\| \le b \quad \forall g \in G$

$\overline{co(G)}$ denotes the closure of the convex hull of G (clear?)

Assume $\bar f \in \overline{co\,G}$. Then $\forall n \ge 1$ and $\forall c' > b^2 - \|\bar f\|^2$,

$\exists\, g_1, \dots, g_n \in G$ and $\alpha_1, \dots, \alpha_n \ge 0$ s.t. $\sum_{j=1}^n \alpha_j = 1$ with

$$\left\| \bar f - \sum_{j=1}^n \alpha_j g_j \right\|^2 \le \frac{c'}{n}$$

NB: The surprise here is that an error $O(\frac{1}{n})$ can be obtained with only n points [exactly what we want]

Proof (probabilistic 😊)

- $\bar f \in \overline{co\,G}$, so $\forall \delta > 0$ & $\forall n \ge 1$, $\exists f^* \in co\,G$ s.t. $\|\bar f - f^*\| \le \frac{\delta}{n}$
- $f^* \in co\,G$, i.e. $\exists m$ suff. large s.t. $f^* = \sum_{k=1}^m \gamma_k g_k^*$
  for some $\gamma_1, \dots, \gamma_m \ge 0$, $\sum_{k=1}^m \gamma_k = 1$, $g_1^* \dots g_m^* \in G$
- take now g drawn from $g_1^* \dots g_m^*$ with probabilities $\mathbb{P}(g = g_k^*) = \gamma_k$
  and pick $g_1 \dots g_n$ iid (randomly) $\sim g$, $f_n = \frac{1}{n}\sum_{j=1}^n g_j$
- $\mathbb{E}(f_n) = \mathbb{E}(g) = f^*$
- $\mathbb{E}\left(\|f_n - f^*\|^2\right) = \mathbb{E}\left(\left\|\frac{1}{n}\sum_{j=1}^n (g_j - f^*)\right\|^2\right) = \frac{1}{n^2}\,\mathbb{E}\left(\sum_{j,\ell=1}^n \langle g_j - f^*, g_\ell - f^* \rangle\right)$

(Var $O(\frac{1}{n})$)
$$= \frac{1}{n^2}\,\mathbb{E}\left(\sum_j \|g_j - f^*\|^2\right) + \frac{1}{n^2}\sum_{j \ne \ell} \langle \mathbb{E}(g_j - f^*), \mathbb{E}(g_\ell - f^*) \rangle$$

(the cross terms $\to 0$)

key!
$$= \frac{1}{n}\,\mathbb{E}\left(\|g - f^*\|^2\right) = \frac{1}{n}\,\mathbb{E}\left(\|g\|^2 - 2\langle g, f^* \rangle + \|f^*\|^2\right)$$
$$= \frac{1}{n}\left(\mathbb{E}\left(\|g\|^2\right) - 2\langle f^*, f^* \rangle + \|f^*\|^2\right) = \frac{1}{n}\left(\mathbb{E}\left(\|g\|^2\right) - \|f^*\|^2\right)$$
$$\le \frac{1}{n}\left(b^2 - \|f^*\|^2\right)$$
hyp.

- If the preceding inequality holds in expectation, this means $\exists$ actual $g_1 \dots g_n \in G$, $f_n = \frac{1}{n}\sum_{j=1}^{n} g_j$ s.t.

  ↳ NB: uniform weights α

$$\|f_n - f^*\| \le \frac{1}{n}\left(b^2 - \|f^*\|^2\right)$$

(skip details)

- Now, $\|f_n - \bar{f}\| \le \|f_n - f^*\| + \|f^* - \bar{f}\| \le \dfrac{b^2 - \|f^*\|^2 + \delta}{n}$

& $\|\bar{f} - f^*\| \ge \|\bar{f}\| - \|f^*\| \Rightarrow \|f^*\| \ge \|\bar{f}\| - \frac{\delta}{n} \Rightarrow \|f^*\|^2 \ge \|\bar{f}\|^2 - \frac{2\delta\|\bar{f}\|}{n}$

so $\|f_n - \bar{f}\| \le \dfrac{b^2 - \|\bar{f}\|^2 + \delta(1 + 2\|\bar{f}\|)}{n}$ ← can be taken arbitrarily close to $b^2 - \|\bar{f}\|^2$ by taking δ small #

Application

- $G_\phi = \{ \gamma\, \phi(a \cdot x + b),\ |\gamma| \le 2C_r,\ a \in \mathbb{R}^d,\ b \in \mathbb{R} \}$
- $\Gamma_{C, B_r} = \{ f : B_r \to \mathbb{R}$ s.t. $\exists$ regular extension $\tilde{f} : \mathbb{R}^d \to \mathbb{R}$ with $\int_{\mathbb{R}^d} |\omega|\, |\hat{\tilde{f}}(\omega)|\, d\omega \le C \}$

Theorem 2

$\forall f \in \Gamma_{C, B_r}$, the function $f(x) - f(0)$ is in the closure of the convex hull of G_ϕ, where the closure is taken with respect to the $L^2(B_r, \mu)$ norm: $\|f - g\|^2 = \int_{B_r} (f(x) - g(x))^2 d\mu(x)$

(Therefore the claim) [up to the constant: to come]

Proof steps: 1) $\Gamma_{C, B_r} \subset \overline{\text{co}}\, G_{\cos}$

$$G_{\cos} = \left\{ \frac{\gamma}{|\omega|} \left(\cos(\omega \cdot x + b) - \cos(b)\right) : \omega \ne 0,\ |\gamma| \le C,\ b \in \mathbb{R} \right\}$$

rem: A convex
⇒ $\bar{A}$ convex
⇒ ok!

2) $G_{\cos} \subset \overline{\text{co}}\, G_{\text{step}}$ (i.e. $\phi(z) = 1_{\{z \ge 0\}}$)

$$G_{\text{step}} = \{ \gamma\, 1_{\{\alpha x \ge t\}} : |\gamma| \le 2C_r,\ |\alpha| \le 1,\ |t| \le r \}$$

(detail omitted)

3) $G_{\text{step}} \subset \overline{\text{co}}\, G_\phi$

15.4.16
5

Proof of 1)

Let $\tilde{f}$ be the extension to $\mathbb{R}^d$ of f : → FT $\tilde{f}(\omega)$

$x \in B$: $f(x) - f(0) = \int_{\mathbb{R}^d} (e^{i\omega \cdot x} - 1)\, \tilde{f}(\omega)\, d\omega$, $C_f = \int_{\mathbb{R}^d} |\omega|\, |\tilde{f}(\omega)|\, d\omega \le C$ by assumption

Write : $\tilde{f}(\omega) = e^{i\Theta(\omega)} \cdot R(\omega)$

Because f is real-valued,

$$\bar{f}(x) = f(x) - f(0) = Re\left(\int_{\mathbb{R}^d} (e^{i\omega x} - 1)\, e^{i\Theta(\omega)}\, R(\omega)\, d\omega \right)$$

$$= \int_{\mathbb{R}^d} \left(\cos(\omega \cdot x + \Theta(\omega)) - \cos(\Theta(\omega)) \right) R(\omega)\, d\omega$$

$$= \int_{\mathbb{R}^d} \underbrace{\frac{C_f}{|\omega|} \left(\cos(\omega \cdot x + \Theta(\omega)) - \cos(\Theta(\omega)) \right)}_{= g(x,\omega)} \underbrace{\frac{|\omega| \cdot R(\omega)}{C_f}}_{= p(\omega)\ \text{prob. density}} d\omega$$

$\left(|g(x,\omega)| \le \frac{C_f}{|\omega|} \cdot |\omega \cdot x| \le C_f r \le C r \right.$ well def.) (*) (**)

So $\bar{f}$ = ∞-convex combination of functions in G_{cos}

$\Rightarrow \bar{f} \in \overline{co}\, G_{cos}$:

draw iid $\omega_1 .. \omega_n \sim p(\omega)$ and define $f_n(x) = \frac{1}{n} \sum_{j=1}^{n} g(x, \omega_j)$

$$\mathbb{E}\left(\|\bar{f} - f_n\|^2 \right) = \mathbb{E}\left(\int_{B_r} (\bar{f}(x) - f_n(x))^2\, \mu(dx) \right)$$

$$= \int_{B_r} \mathbb{E}\left(\left(\bar{f}(x) - \frac{1}{n} \sum_{j=1}^{n} g(x, \omega_j) \right)^2 \right) \mu(dx)$$

$$= \mathbb{E}\left(\left(\frac{1}{n} \sum_{j=1}^{n} (\bar{f}(x) - g(x,\omega_j)) \right)^2 \right) = \frac{1}{n}\, \mathbb{E}\left((\bar{f}(x) - g(x,\omega))^2 \right) \le \frac{1}{n}\, Var(g(x,\omega))$$

(ω_j indep & $\mathbb{E}(g(x,\omega)) = \bar{f}(x)$)

$$\le \frac{(Cr)^2}{n}$$

$\Rightarrow \exists$ an actual sequence $f_n = \frac{1}{n} \sum_{j=1}^{n} g(x, \omega_j)$ s.t. $\|\bar{f} - f_n\|^2 \le \frac{(Cr)^2}{n}$. #

(*) BTW; Any $g \in G_{cos}$ satisfies $\|g\|^2 = \int_{B_r} (g(x))^2\, \mu(dx) \le \sup_{x \in B_r} |g(x)|^2 \le \underbrace{(Cr)^2}_{= b^2}$

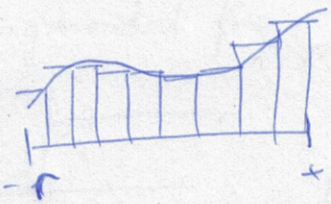
(**) $\nabla_x g(x,\omega) = \frac{C_f}{|\omega|} \sin(\omega \cdot x + \Theta(\omega)) \cdot \omega$: $|\nabla g| \le C_f \le C$

Proof of 2) $\left[G_{cos} \subset \overline{co\, G_{step}} \right]$

$f \in G_{cos} : f = g \circ h(x)$ $\begin{cases} g(z) = \frac{\gamma}{|\omega|}\left(\cos(|\omega| z + b) - \cos(b)\right) \\ z = h(x) = \frac{\omega \cdot x}{|\omega|} \in [-r, r] \end{cases}$

g unif. continuous on $[-r,r]$, $|g'(z)| \le |\gamma| \le C$ ($\Rightarrow$ total variation of g on $[-r,r] \le 2Cr$

$\Rightarrow$ can be uniformly well approximated by step fns



more precisely, $g(0)=0$ so for $z>0$, $g(z) \simeq \sum_{i=1}^{k-1} (g(t_i) - g(t_{i-1}))\, 1_{z \ge t_i}$

add the two! $z<0$ $g(z) \simeq \sum_{i=1}^{k-1} (g(-t_i) - g(-t_{i-1}))\, 1_{z \le -t_i}$

& $\sum_{i=1}^{k-1} |g(t_i) - g(t_{i-1})| \le \int_0^r |g'(z)|\, dz \le |\gamma| r \le Cr$, $\sum_{i=1}^{k-1} |g(-t_i) - g(-t_{i-1})| \le Cr$

g = linear combination of functions $1_{z \ge t_i}$, $1_{-z \ge t_i}$ with coeffs $|\cdot| \le 2Cr$
(with $k \to \infty$)

$\in \overline{co}\{ f_0(z) = \gamma 1_{z \ge t},\ |\gamma| \le 2Cr,\ |t| \le r \}$

$\Rightarrow f \in \overline{co}\{ f_0(z) = \gamma 1_{\alpha \cdot x \ge t},\ |\alpha| \le 1,\ |\gamma| \le 2Cr,\ |t| \le r \}$ #

NB: all this with respect to $\|\cdot\|_\infty$ norm
$\longrightarrow$ also with respect to $L^2(B_r, \mu)$ norm !

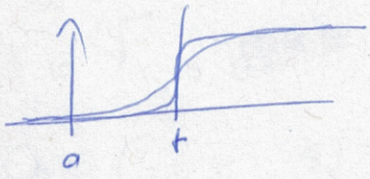
Proof of 3):

detail: $t \in [-r,r] \to t \in$ continuity points of ~~[illegible]~~ the distribution of $Z = \alpha \cdot x$ induced by μ on B_r

= dense set! $\to G^{\mu}_{step}$! $G_{\cos} \subset \overline{co\, G^{\mu}_{step}}$

$G^{(\mu)}_{step} \subset \overline{co\, G_\phi}$:

take sigmoidal fns: $\phi(|a|(\alpha \cdot x - t))$ with $|a| \to \infty$



$\xrightarrow[|a|\to\infty]{} 1_{\alpha \cdot x \geq t}$

(except for x's st $\alpha \cdot x = t$ but $\cdot \mu(\{x : \alpha x = t\}) = 0$ by restriction on t) #

Constants (end of proof)

Claim: $f \in \Gamma_{C,B_r} \Rightarrow \|\bar{f} - \bar{f}_n\|^2_{L^2(B_r,\mu)} \leq \frac{(2rC)^2}{n}$

$\bar{f}(x) = f(x) - f(0)$

- if $\|\bar{f}\| = 0$, then [illegible], $\bar{f} \equiv 0$: clear ($C = 0$ ok)
- assume $\|\bar{f}\| > 0$.

$G_\phi = \{ \gamma \cdot \phi(a \cdot x - b), |\gamma| \leq 2Cr, a \in \mathbb{R}^d, b \in \mathbb{R} \}$

$f \in G_\phi \Rightarrow \|f\|_{L^2(B_r,\mu)} \leq \|f\|_\infty \leq 2Cr$ ($|\phi| \leq 1$) → add hyp.

$\Gamma_{C,B_r} \subset \overline{co\, G_\phi} \Rightarrow$ lemma 1: $\exists f_n \in G_\phi$ st $\|\bar{f} - f_n\|^2_{L^2(B_r,\mu)} \leq$ [illegible]

$\forall C' > (2Cr)^2 - \|\bar{f}\|^2$

$\Rightarrow$ one can choose $C' = (2Cr)^2$ #

- $|\phi| \not\leq 1$: ok - - -