Deep Learning Theory - Lecture 3

Q: how about higher dimensions? $d > 1$

<u>Idea</u>: We would like to approximate with sums.

$$\prod_e \mathbb{1}(x \in A_e) \qquad A_e = \bigtimes_{j=1}^{d} [b_{e,j}, b'_{e,j}).$$

However, this cannot be done (at least directly) with shallow nets.

Or can it?

Consider $\sigma(z) = \cos(z)$. Then

$$\begin{aligned}
\sigma(z) \cdot \sigma(y) &= \cos(z)\cos(y) \\
&= \cos(z+y) + \cos(z-y). \\
&= \sigma(z+y) + \sigma(z-y).
\end{aligned}$$

So, at least in principle, using the above property of cosines and the Fourier approximation of step functions:

$$g(x) \approx \sum_j \alpha_j \prod_e^d \mathbb{1}(x_e \in [b_{j,e}, b'_{j,e})) \approx \sum_j \alpha_j \prod_e \sum_k \beta_k \cos(\gamma_k x_e)$$

algebra

$$\approx \sum_j \alpha_j \sum_k \beta_k \prod \cos(\gamma_k x_e) \approx \sum_j \alpha_j \sum_k \beta \sum_m \delta_m \cos\left(\sum \gamma'_{ekm} x_e\right)$$

$\overset{\text{expressivity}}{\longrightarrow} \approx \sum_j \alpha_j \sum_e \beta_e \sum \delta_m \sum \eta_n \mathbb{1}[\gamma_n' \cdot x \geq b_n]$

In practice, this is painful, but the heavy lifting was done for us:

Thm 2.5 (Stone-Weierstrass): Let $\mathcal{F} \subseteq C(\mathcal{X})$ for compact $\mathcal{X} \subseteq \mathbb{R}^d$ satisfy.

a) for every $x \in \mathcal{X}$, there exists $f \in \mathcal{F}$ such that $f(x) \neq 0$

b) for every pair $x, x' \in \mathcal{X}$ with $x \neq x'$ there exists $f \in \mathcal{F}$ with

$$f(x) \neq f(x') \qquad (\mathcal{F} \text{ separates points})$$

c) $\mathcal{F}$ is closed under pointwise multiplication ($\mathcal{F}$ is an algebra)

then $\mathcal{F}$ is an universal approximator.

Lemma 2.6 $\mathcal{F}_{\cos}$ is universal

Pf: a) each $f \in \mathcal{F}_{\cos}$ is continuous (finite sum of cont. functions)

b) $\cos(0 \cdot x) = 1 \quad \forall x \in \mathcal{X}$

c) $x \neq x' \implies f(z) = \cos\left(\frac{(z-x')\cdot(x-x')}{\|x-x'\|^2}\right)$ satisfies $\begin{cases} f(x') = 1 \\ \quad \neq \\ f(x) = 0 \end{cases}$

d) already checked. $\square$

Thm 2.7 Suppose $\sigma \in C(\mathbb{R})$ is sigmoidal: $\begin{cases} \lim_{z \to -\infty} \sigma(z) = 0 \\ \lim_{z \to \infty} \sigma(z) = 1 \end{cases}$

then $\mathcal{F}_\sigma$ is universal.

Also, $\mathcal{F}_{\mathrm{ReLU}}$ is universal

Pf (sketch): By Lemma 2.7 we have there exists $n \in \mathbb{N}$,

$$h_n(x) = \sum_{j=1}^{n} \tilde{a}_j \cos(\tilde{\omega}_j \cdot x + \hat{b}_j) \in \mathcal{F}_{\cos}$$

with $\|h - g\|_\infty \leq \frac{\varepsilon}{2}$

Then, since $h_{n,j}(x) = \hat{a}_j \cos(\hat{\omega}_j \cdot x + \hat{b}_j) \in C(\mathcal{X})$, by exercise we have $\exists f_{n,j} \in \mathcal{F}_{\text{sigmoid}} : \|f_j - h_{n,j}\| \leq \frac{\varepsilon}{2n}$

$\Longrightarrow$ for $f(x) = \sum_j f_{n,j}(x) \in \mathcal{F}_{\text{sigmoid}}$

$$\|f_n - g\|_\infty \leq \|f_n - h_n\| + \|h_n - g\| \leq \sum_{j=1}^{n} \frac{\varepsilon}{2n} + \frac{\varepsilon}{2} \leq \varepsilon. \qquad \square$$

Note: the algebra condition does not hold for polynomials of bounded degree. In fact.

Thm (Leshno, 1993): $\mathcal{F}_\sigma$ is universal iff $\sigma \in C(\mathcal{X})$ is not a polynomial

Multilayer neural networks

Def: Let $L\in\mathbb{N}$. A fully connected feedforward neural network of widths $(n_1,\dots n_L)\in\mathbb{N}^L$ is a function of the form

$$f_\theta(x) = \sigma_{L+1}\left(z^{L+1}(x)\right)$$

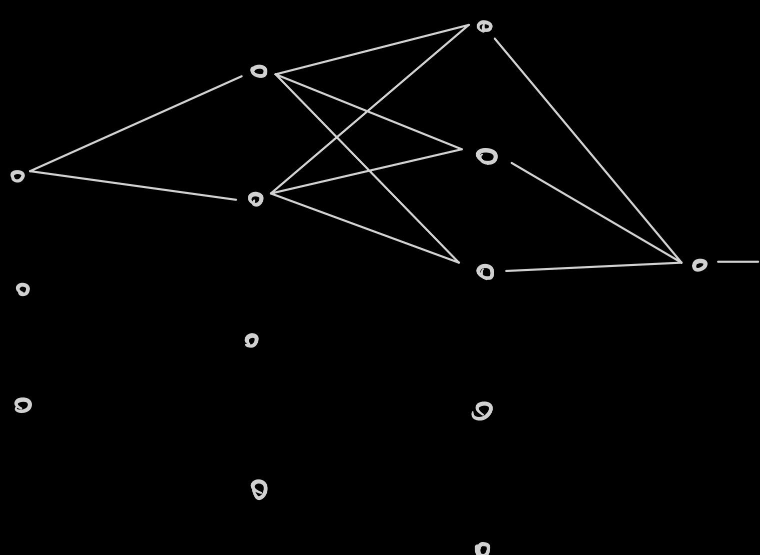
where the preactivations are given by:

$$z_j^\ell(x) = \alpha_{\ell-1}\sum_{k=1}^{n_{\ell-1}} w_{jk}^\ell\, \sigma_\ell\left(z_k^{\ell-1}(x)\right) + b_j^\ell \qquad j\in\{1,\dots,n_\ell\}$$
$$\ell\in\{2,\dots L+1\}$$

$$z_j^1(x) = \sum_{k=1}^{d} w_{jk}^1 x_k + b_j^1$$

for a choice of $\theta = \left(\left(w_{jk}^\ell\right)_{j=1,k=1}^{n_\ell,n_{\ell-1}}, \left(b_j^\ell\right)_{j=1}^{n_\ell}\right)_{\ell=1}^{L+1} \in \mathbb{R}^{\sum_{\ell=1}^{L}(n_{\ell-1}+1)n_\ell}$

<u>Note</u>: • This network can be represented as a graph:



$f_\theta(x)$

- Letting σ_ℓ act componentwise we can write

$$f_\theta(x) = \sigma_{L+1}\left(W_{L+1}\,\sigma_L\left(\cdots W_2\,\sigma_1\left(W_1 x + b_1\right) + b_2 \cdots + b_{L+1}\right)\right)$$

for $W^\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$, $b^\ell \in \mathbb{R}^{n_\ell}$

- σ_L are the same as in the single layer setting.

<u>Lemma</u>: Deep neural networks with ReLU activation are universal approximators.

<u>Proof</u>: since for $z \in \mathbb{R}$ $\quad \mathbb{1}(z) = -(-z)_+ + z_+ = -1\,\sigma((-1)z) + 1\cdot\sigma(1\cdot z)$
we can construct a network of depth L combining $L-1$ layers of $\mathbb{1}$ with the network from <u>Thm 2.7</u>.

Q: Why using deep neural networks?

<u>Thm</u> (Telgarsky 2015): for any $L \geq 2$ there exists a depth $2L^2+4$ ReLU NN with $3L^2+6$ nodes f_L such that for any depth L NN with $\leq 2^L$ nodes we have $\|g - f_L\|_1 \geq \frac{1}{32}$.

Neural network training

For a given dataset $\mathcal{D}_n$, we aim to minimize the empirical risk

$$\hat{R}(\vartheta) = \hat{R}(f_\vartheta) = \frac{1}{n}\sum_{j=1}^{n} \ell(f_\vartheta(x_j), y_j) \qquad \text{(training error)}$$

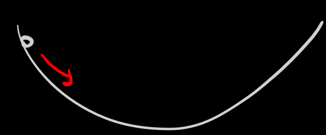
For an algorithm $\mathcal{A}$, we aim to characterize the optimization error

$$\hat{R}(\mathcal{A}(\mathcal{D}_n)) - \inf_{f \in \mathcal{F}} \hat{R}(f)$$

While in some cases this can be done explicitly (e.g. linear regression) in general the problem of finding the minimum of a function $\hat{R}$ is hard.

One method to (hopefully) solve this problem: move sequentially in the direction (in Θ) of steepest descent of ϑ by updating

$$\vartheta \longleftarrow \vartheta - \gamma D_\vartheta R(\vartheta)$$



for a small timestep parameter γ. This method is called gradient descent:

the update reads: $\vartheta_{k+1} = \vartheta_k - \gamma_k D_\vartheta R(\vartheta_k)$.

Note: Why using this and not trying to solve $D_\vartheta R(\vartheta) = 0$?

Computation of $D_\vartheta R$ is cheap: consider

$$D_{W^1_{11}} f(\vartheta) = \sigma'_{L+1}(z_{L+1}(x))\, W^L \sigma'_L(z_L(x)) \cdots \sigma'_1(z_1(x)) \cdot x$$

In the above update, provided that we know σ_e' and z_e we are computing a complicated derivative by taking a product of known numbers (z_e were evaluated to find $f_\theta(x)$) so that

$$\vartheta_{k+1} = \vartheta_k - \gamma_k D_\theta \hat{R}(\vartheta_k)$$

$$= \vartheta_k - \gamma_k \frac{1}{n} \sum_{j=1}^{n} \underbrace{D_\theta f_{\vartheta_k}(x_j)}_{\text{"easy"}} \underbrace{D_f \ell(f_{\vartheta_k}(x_j), y_j)}_{\text{"easy": just evaluate } f_{\vartheta_k}}$$

To study the dynamics of ϑ_k we write formally

$$\vartheta_{k+1} = \vartheta_k - \gamma D_\theta R(\vartheta_k) \longrightarrow \frac{\vartheta_{k+1} - \vartheta_k}{\gamma} = - D_\theta R(\vartheta_k)$$

Based on the above, one expects that $\vartheta_k \approx \bar{\vartheta}_{k\gamma}$ where $\bar{\vartheta} : \mathbb{R}_+ \to \Theta$ solves

$$\begin{cases} \frac{d}{dt} \bar{\vartheta}_t = - D_\theta R(\bar{\vartheta}_t) \\ \bar{\vartheta}_0 = \vartheta_0 \end{cases} \qquad \text{(gradient flow)}$$

<u>Lemma</u>: Let $D_\theta R$ be Lipschitz. For every $T > 0$ there exists $C > 0$ s.t.

$$\| \bar{\vartheta}_{k\gamma} - \vartheta_k \| \le C\gamma \qquad \forall\, k \in \{0, \dots, \lfloor \tfrac{T}{\gamma} \rfloor\}$$

<u>Proof</u>: $\bar{\vartheta}_{t+\gamma} = \bar{\vartheta}_t - \gamma D_\theta R(\vartheta_t) + C\gamma^2$

$$e_{k+1} = \| \bar{\vartheta}_{(k+1)\gamma} - \vartheta_{k+1} \| = \| \bar{\vartheta}_{k\gamma} - \gamma D_\theta R(\bar{\vartheta}_{k\gamma}) - \vartheta_k + \gamma D_\theta R(\vartheta_k) \| + C\gamma^2$$

$$\le (1 + \gamma\lambda) e_k + C\gamma$$

$$\Longrightarrow e_{k+1} = C\gamma^2 \sum_{j=1}^{k} (1+\lambda\gamma)^j = \frac{C\gamma^2}{\lambda\gamma}\left((1+\lambda\gamma)^{k+1} - 1\right) \le \frac{C\gamma}{\lambda}\left(e^{\lambda\gamma(k+1)} - 1\right) \le C'\gamma \qquad \square$$

This problem simplifies when $\hat{R}$ is convex and has Lipschitz derivative.

Def: $\hat{R}(\vartheta)$ is λ strongly convex if

$$\hat{R}(\vartheta') \ge \hat{R}(\vartheta) + \langle D_\vartheta \hat{R}(\vartheta), \vartheta' - \vartheta \rangle + \frac{\lambda}{2}\|\vartheta' - \vartheta\|^2 \qquad \text{for any } \vartheta, \vartheta' \in \Theta$$

$$\Longrightarrow \hat{R}(\vartheta') \ge \hat{R}(\vartheta) + \langle D_\vartheta \hat{R}(\vartheta'), \vartheta - \vartheta' \rangle + \frac{\lambda}{2}\|\vartheta - \vartheta'\|^2$$

adding the two inequalities we have

$$\langle D_\vartheta \hat{R}(\vartheta') - D_\vartheta \hat{R}(\vartheta), \vartheta' - \vartheta \rangle \ge \lambda \|\vartheta - \vartheta'\|^2$$

Thm: Let $R(\vartheta)$ be λ-convex, then there exists a unique ϑ^* and

$$\|\bar{\vartheta}_t - \vartheta_*\|^2 \le \|\bar{\vartheta}_0 - \vartheta_*\|^2 e^{-2\lambda t}$$

Proof: $$\frac{d}{dt}\frac{1}{2}\|\bar{\vartheta}_t - \vartheta_*\|^2 = \langle \bar{\vartheta}_t - \vartheta_*, \frac{d}{dt}\vartheta_t \rangle$$

$$= -\langle \bar{\vartheta}_t - \vartheta_*, D_\vartheta R(\bar{\vartheta}_t) - D_\vartheta R(\vartheta_*) \rangle$$

$$\le -\lambda \|\bar{\vartheta}_t - \vartheta_*\|^2$$

$\longrightarrow$ Grönwall