Advanced Deep Learning — Lecture 2 26/09/2022

Loss function

$$\ell : \mathbb{R}^p \times \mathbb{R}^p \to \overline{\mathbb{R}}_+$$

$$\ell(a,b) = \frac{1}{2}\|a-b\|^2 \quad \forall\, a,b \in \mathbb{R}^p$$

In ML $f : X \to Y$ $\qquad \ell(f(x), y)$

$f(x)$: prediction on x; y: lable

<u>Def</u>

A functional $R : \mathcal{X} \to \mathbb{R}$ is a map from a space $\mathcal{X}$ (usually ∞-dimensional) to real numbers

$$\pi \to \int_\Omega F(z)\, \underline{d\pi(z)} \qquad z \sim (x,y)$$

$d\pi(z) \to dz$ $\qquad d\pi(z) \to \underline{g(z)}\, \underline{dz}$

If $\pi = \mathcal{L}$ (Lebesgue measure) $\quad d\pi(x,y) = dx\,dy$

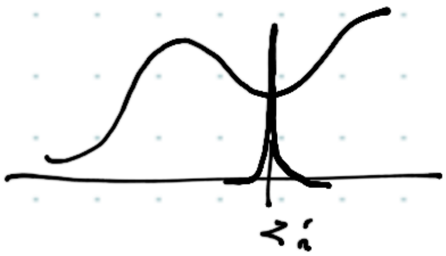
Going from the Risk functional to the empirical risk functional

$$R(f) = \int_\Omega \ell(f(x), y)\, d\pi(x,y)$$

$$\pi \to \pi_n = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$$

$$\frac{1}{n}\int_\Omega \ell(f(x),y) \sum_{i=1}^n d\delta_{z_i}(x,y) = \frac{1}{n}\sum_{i=1}^n \int_\Omega \ell(f(x),y)\, d\delta_{z_i}(x,y)$$

$d\delta_{z_i}(x,y) \to \delta(x-x_i)\,\delta(y-y_i)\; dx\,dy$



$$= \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) = R_n(f)$$

Empirical risk functional

$$f_n^* \in \boxed{\underset{f \in \mathcal{F}}{\operatorname{argmin}} \; R_n(f)}$$

This is a set

$$\cancel{f_n^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; R_n(f)}$$

function

set of functions

The set $X^Y = \{ f : X \to Y \}$

Uniform Convergence

$$R(f_n^*) - \underline{\underline{\inf_{f \in \mathcal{F}} R(f)}} = \big(R(f_n^*) - R_n(f_n^*)\big) \quad ①$$

$$+ \big(R_n(f_n^*) - R_n(f^*)\big) \quad ②$$

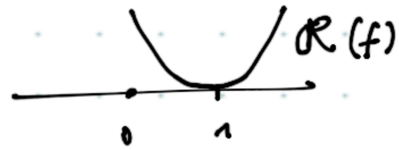
$$+ \big(R_n(f^*) - \underline{\underline{R(f^*)}}\big) \quad ③$$

$$f^* \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; R(f)$$

$$② = R_n(f_n^*) - R_n(f^*)$$

Finite-dimensional example $\quad R_n(f) = f^2 =$

$f_n^* = 0$ $\qquad$ $R(f) = (f-1)^2$



$f^* = 1$

$$R_n(f_n^*) = 0$$
$$R(f^*) = 0$$
$$R_n(f^*) = 1$$

$$R_n(f_n^*) - R_n(f^*) = 0 - 1 = -1 \leq 0$$

We conclude that ② ≤ 0

$$R(f_n) - \inf_{f \in \mathcal{F}} R(f) \leq \left(R(f_n^*) - R_n(f_n^*)\right)$$

$$+ \left(R_n(f^*) - R(f^*)\right)$$

The ③ term can be controlled using LLN

n $\quad (X_i) \quad \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu$

are drawn from π

$$R_n(f^*) = \frac{1}{n}\sum_{i=1}^{n} \ell(f^*(x_i), y_i)$$

$$R(f^*) = \int_\Omega \ell(f^*(x), y)\, d\pi(x, y)$$

$$\left| R_n(f^*) - R(f^*) \right| \xrightarrow{n \to +\infty} 0 \quad \text{by LLN}$$

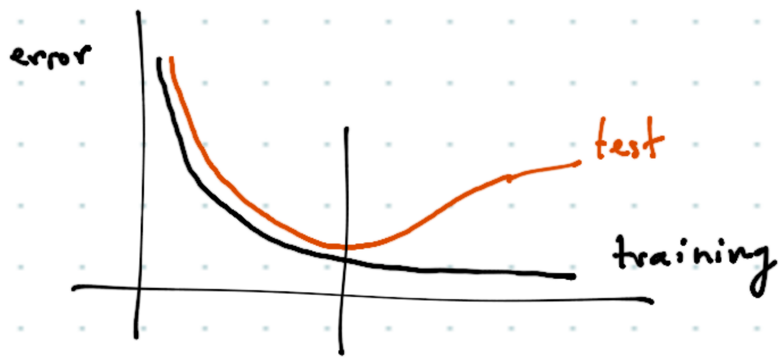
The difficult term is ①

$$\left(R(f_n^*) - R_n(f_n^*) \right) \leq \sup_{f \in \mathcal{F}} \left| R(f) - R_n(f) \right|$$

If the loss is "well behaved" we can prove

$$\sup_{f \in \mathcal{F}} \left| R(f) - R_n(f) \right| \to 0$$

Uniform law of large numbers



Early stopping

Regularization

Gradient flow is a solution to the ODE

$$(*) \begin{cases} w'(t) = -\nabla\phi(w(t)) \\ w(0) = w^0 \end{cases} \qquad \leftarrow \quad \phi \text{ any } C^{1,1}(\mathbb{R}^N;\mathbb{R})$$

We are interested in the case $\phi = R_n$

$t \mapsto w(t) \in \mathbb{R}^N$

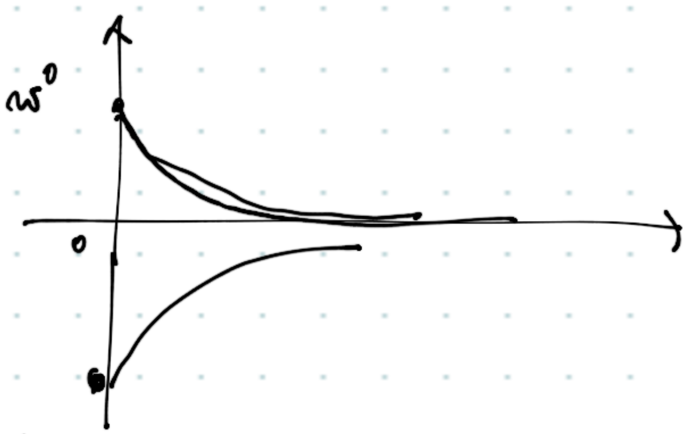
Example $N=1$

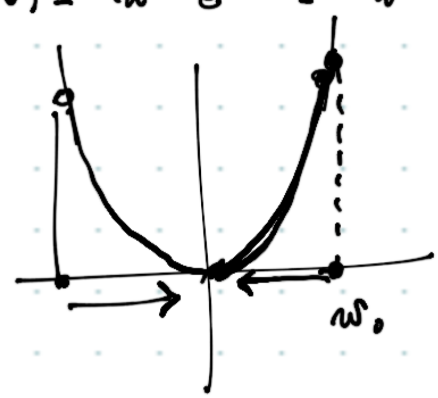
$$\phi(x) = \frac{x^2}{2} \qquad \nabla\phi(x) = x$$

$$\nabla\phi(w(t)) = w(t)$$

Problem $(*)$ becomes

$$\begin{cases} w'(t) = -w(t) \\ w(0) = w^0 \end{cases} \qquad \begin{aligned} w(t) &= w^0 e^{-t} \\ w'(t) &= -w^0 e^{-t} = -w(t) \\ w(0) &= w^0 e^{-0} = w^0 \end{aligned}$$



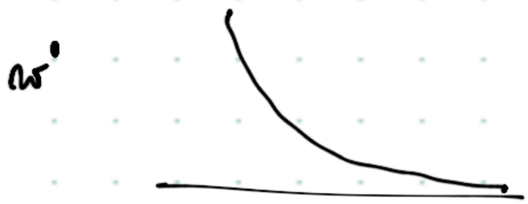


" ϕ decreases along the gradient flow " $\quad \phi : \mathbb{R}^N \to \mathbb{R}$

$$\frac{d}{dt}\phi(w(t)) = \nabla\phi(w(t)) \cdot w'(t) = -w'(t) \cdot w'(t)$$

Solves $(*)$ — dot prod in $\mathbb{R}^N$ — Because $w' = -\nabla\phi$

$= -\|w'\|^2$, $\frac{d}{dt}\phi(w(t)) = -\|w'\|^2 \leq 0$



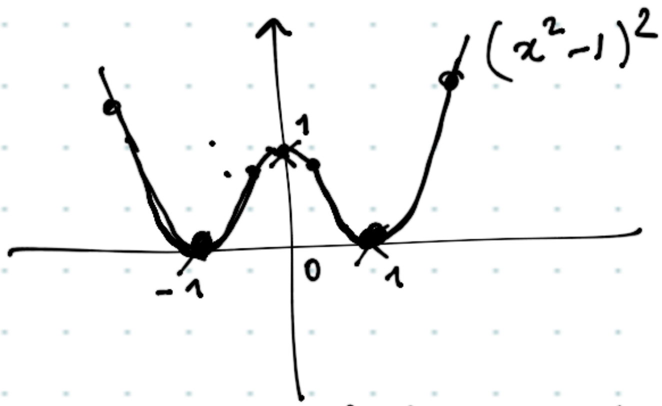
$\phi \downarrow$

<u>Exercise</u> Find the gradient flow when $N=1$

$$\phi(x) = (x^2-1)^2 \quad \text{i.e. solve}$$

$$\nabla\phi(x) = 2(x^2-1)\,2x = 4x(x^2-1)$$

$$\begin{cases} w'(t) = -4\,w(t)\left((w(t))^2-1\right) \\ w(0) = w^0 \end{cases}$$

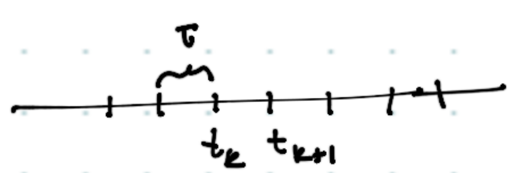


these solution reaches the minima of ϕ <u>asymptotically</u>

Gradient Descent

Gradient descent can be interpreted as an Euler method to solve the gradient flow (*)

$$w'(t) = f(w(t), t)$$



$$w(t_k) = w^k$$

$$\frac{w^{k+1} - w^k}{\tau} = \begin{cases} f(w^k, t_k) & \text{explicit Euler method} \\ f(w^{k+1}, t_{k+1}) & \text{implicit Euler method} \end{cases}$$

If you apply this to (*):

$$\frac{w^{k+1} - w^k}{\tau} = -\nabla \phi (w^k) \quad \text{explicit}$$

$$\frac{w^{k+1} - w^k}{\tau} = -\nabla \phi (w^{k+1}) \quad \text{implicit}$$

Which are usually written as

$$w^{k+1} = w^k - \overset{\text{Learning rate}}{\tau} \nabla \phi (\underline{\underline{w^k}})$$

$$w^{k+1} = w^k - \tau \nabla \phi (\underline{\underline{w^{k+1}}})$$

A "better" way to define GD

$$(1) \quad w^{k+1} \in \underset{s \in \mathbb{R}^N}{\operatorname{argmin}} \quad \phi(s) + \frac{1}{2\tau} \| s - w^k \|^2$$

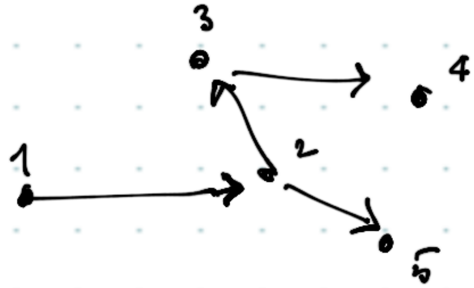
$$(2) \quad w^{k+1} \in \underset{s \in \mathbb{R}^N}{\operatorname{argmin}} \quad \phi(w^k) + \nabla \phi (w^k) \cdot (s - w^k) + \frac{1}{2\tau} \| s - w^k \|^2$$

Exercise Prove (1) is equivalent to the implicit GD
" (2) " " , " explicit GD

Back prop. on a dag

$G = (V, A)$

$\uparrow$ set of vertices



$A = \{ 1 \to 2,\ 2 \to 3,\ 3 \to 4,\ \vdots \}$

$1 \to 2$ is an arc of the dag.

These are not Dags

given $i \in V$ $\qquad pa(i) = \{ j \in V : j \to i \in A \}$

$j \in pa(i)$

$ch(i) = \{ j \in V : i \to j \in A \}.$

$j \in ch(i)$

What is a FNN? In general is a dag + $i \mapsto x^i$, $j \to i \to w_{ij}$

$x^3 = \sigma(w_{31}\ x_1\ )$

With the following computational rule

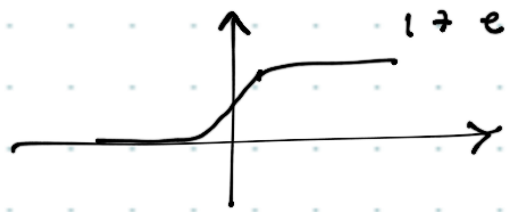
$$x^i = \begin{cases} \sigma\Big(\underbrace{\sum_{j \in pa(i)} w_{ij}\, x^j}_{a^i}\Big) & \text{if } i \in H \cup O \\ y^i & \text{if } i \in I \end{cases}$$

$I = \{ i \in V : pa(i) = \phi \}$

$O = \{ i \in V : ch(i) = \phi \}$

$H = V \setminus \{ O \cup I \}$.

$\sigma \to$ activation function (for instance $\sigma(x) = \dfrac{1}{1+e^{-x}}$



)

z is the input to the network

With this computational rules a network describes a function

$$z \to f(z, w) \quad \text{(which are the values of } x^i \text{ on } O\text{)}$$

What is Backprop?

It is a clever (optimal) way to compute

$$\nabla R_n(w)$$

Which is the key ingredient for GD on R_n

when $R_n(w) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(f(z_i, w), y_i \right)$

↑ this is a NN

$$\frac{\partial}{\partial w_{ij}} \ell = \frac{\partial \ell}{\partial a^i} \frac{\partial a^i}{\partial w_{ij}} \qquad a^i = \sum_{j \in pa(i)} w_{ij} z^j$$

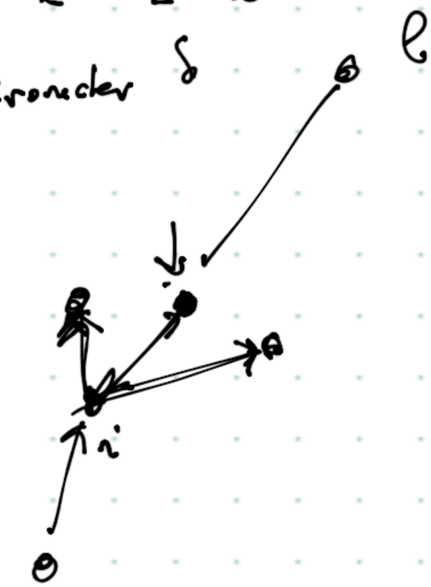
$$\frac{\partial a^i}{\partial w_{ij}} = \frac{\partial}{\partial w_{ij}} \sum_{k \in pa(i)} w_{ik} z^k$$

$$= \sum_{k \in pa(i)} \delta_{kj} z^k = x^j$$

↑ Kronecker δ

$$\frac{\partial \ell}{\partial a^i} = \delta_i \quad \forall i \in H \cup O$$

$$\frac{\partial \ell}{\partial a^i} = \sum_{k \in ch(i)} \frac{\partial \ell}{\partial a^k} \frac{\partial a^k}{\partial a^i}$$



$$= \sum_{k \in ch(i)} \delta_k \frac{\partial a^k}{\partial a^i}$$

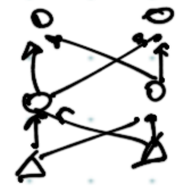
$$a^k = \sum_{j \in pa(k)} w_{kj} x^j = \sum_{j \in pa(k)} w_{kj} \sigma(a^j)$$

$$\frac{\partial a^k}{\partial a^i} = w_{ki} \sigma'(a^i)$$

$$\boxed{\delta_i := \sigma'(a_i) \cdot \sum_{k \in ch(i)} \delta_k w_{ki}}$$

$$\frac{\partial \ell}{\partial w}$$

$\sigma' \in [0, 1/4]$



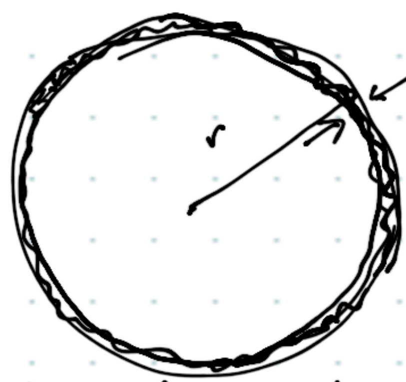
$\sigma' \; \sigma'$

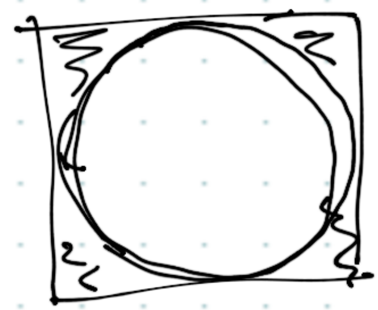
$$\left(\frac{1}{4}\right)^{10}$$

ReLU 

Vanishing gradient problem

Exercise





$$\frac{V_s}{V_c} \xrightarrow{n \to +\infty} 0$$

All the volume lies on the surface as $n \to +\infty$

$$\frac{\overbrace{\mathrm{vol}(B_r) - \mathrm{vol}(B_{r-\varepsilon})}}{\mathrm{vol}(B_r)}$$

$$\mathrm{vol}(B_r) = \omega_n \, r^n$$

ω_n: volume of the unit Ball in $\mathbb{R}^n$

$$0 < \frac{\varepsilon}{r} < 1$$

$$\frac{\omega_n r^n - \omega_n (r-\varepsilon)^n}{\omega_n r^n} = 1 - \Big(\overbrace{1 - \frac{\varepsilon}{r}}^{1}\Big)^n \to 1 - 0 = 1$$