Current Deep Learning Algorithms are based on

1. Back Propagation
2. Stochastic Gradient Descent

ML basics : Empirical Risk minimization

Classical Statistical ML setting :

We have a probability space $(\Omega, \mathcal{F}, \pi)$ — Ω: set; $\mathcal{F}$: measurable subsets of Ω; π: probability measure.

Usually, e.g. in Supervised Learning

$\Omega = X \times Y$ ← labels

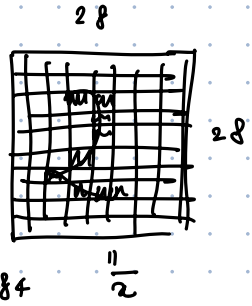
↑ data, examples, perceptual inf.

$(x, y) \in \Omega$

x = image

MNIST

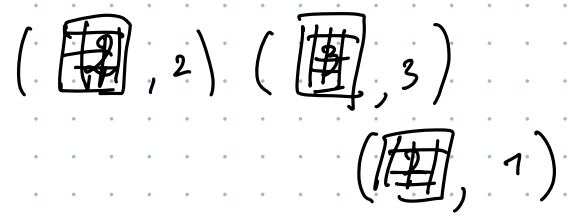
y = label



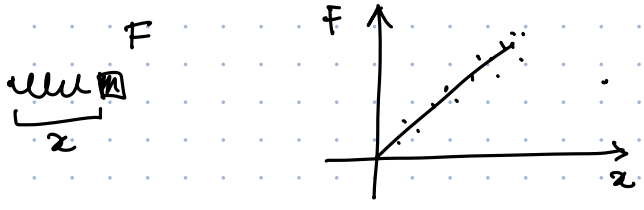
$X = \mathbb{R}^{784}$

Let's talk about π

(x, y)



$y = 2$



General problem in ML is to minimize the Functional Risk

Usually what you want to do is to make predictions on your examples. Meaning that you want to find $f : X \to Y$ $\quad x \mapsto y$

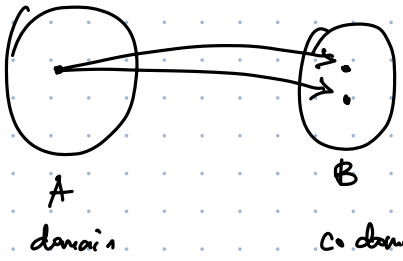
$f(\boxed{2}) = 2$

ℓ is called "loss function" and measures the "goodness" of your predictions.

$\ell : Y \times Y \to Y$

Functions

f is a "rule" that associate elements from a set A to elements of a set B with the following property

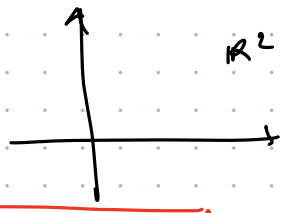


f cannot associate one element of A to two distinct elements of B.

$f : A \to B$

$A \times B$: cartesian product of A and B $\quad A \times B = \{ (a, b) : a \in A, \ b \in B \}$.

$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$$
$$= \{ (x, y) : x \in \mathbb{R}, \ y \in \mathbb{R} \}$$



$$L(f) = \mathbb{E}\, \underset{\uparrow}{\ell} \circ \underset{\uparrow}{(f \times id)} = \int_{\Omega} \ell(f(x, y)) \, d\pi(x, y)$$ Functional Risk

Example $\quad \Omega = \mathbb{R}^2 \quad$ A is any subset of $\mathbb{R}^2$

- $\pi(A) = C \int_A e^{-(x-y)^2} e^{-x^2} \, dx \, dy$ $\quad C$ is a normalization constant.

Exercise: Compute C

$\pi(\mathbb{R}^2) = 1$ Normalization of the probability

$$\pi(\mathbb{R}^2) = C \int_{\mathbb{R}^2} e^{-(x-y)^2} e^{-x^2} \, dx \, dy = 1$$

$$\int_{\mathbb{R}} e^{-x^2} \underbrace{\left(\int_{\mathbb{R}} e^{-(x-y)^2} \, dy \right)}_{\sqrt{\pi}} dx$$

$$\int_{\mathbb{R}} e^{-(x-y)^2}\,dy \qquad y-x=t \quad , dy = dt$$

$$= \int_{\mathbb{R}} e^{-t^2}\,dt = \sqrt{\pi}$$

$$\int_{\mathbb{R}} e^{-x^2}\sqrt{\pi}\,dx = \sqrt{\pi}\int_{\mathbb{R}} e^{-x^2}\,dx = \pi$$

$$\pi(\mathbb{R}^2) = C\,\pi = 1 \qquad C = \frac{1}{\pi}$$

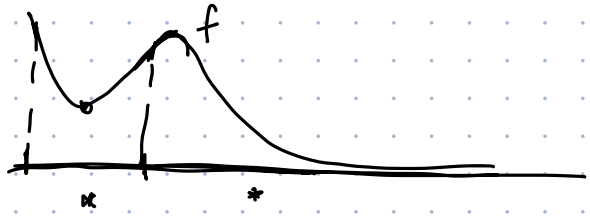
$$p(x,y) = \frac{1}{\pi}\, e^{-(x-y)^2}\, e^{-x^2}$$

$f : X \to Y$

Take $\ell(a,b) = (a-b)^2$

$$\ell(f(x),y) = (f(x)-y)^2$$

$$L(f) = \frac{1}{\pi}\int_{\mathbb{R}^2} (f(x)-y)^2\, e^{-(x-y)^2}\, e^{-x^2}\,dx\,dy \leftarrow$$



Problem is Minimize L on (some space)

$$\text{dom}(L) = \{ f : X \to Y : L(f) < +\infty \}$$

$f(x) = \sin(x) \qquad L(\sin) < +\infty$
$= +\infty$

Calculus of variations

$$\min_{x \in X} F(x)$$

"X = space of functions"

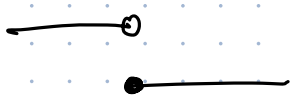
$C^0(\mathbb{R};\mathbb{R})$ continuous functions from $\mathbb{R}$ to $\mathbb{R}$

$f(x) = x$

$f \in C^0(\mathbb{R};\mathbb{R})$

$\sin(x) \quad e^x$

δ is not a function



$$H(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

And in F which are integrals. In particular we are interested in functionals of the form

$$(*) \qquad F(u) = \int_a^b \mathcal{L}(\underset{\Downarrow}{u}(x), \underset{\Downarrow}{u'(x)}, \underset{\Downarrow}{x})\, dx \qquad u \in X$$
$$\qquad\qquad\qquad\quad u \quad p \quad x$$

$$F(u) = \int_0^1 u(x)\, dx \qquad F(id) = \int_0^1 x\, dx = \frac{1}{2}$$

therefore we want to find minima of Functionals of the form $(*)$.

Algorithm: 1. Find a candidate solution $u^*(x)$ that satisfies Euler-Lagrange. $\left(\frac{d}{dx}\mathcal{L}_p - \mathcal{L}_u = 0\right)$ "$\nabla f = 0$"

2. Check if this is indeed a minima, for instance by proving that

$$F(u) \geq F(u^*).$$

$\mathcal{L}$ is usually called Lagrangian. $\mathcal{L}(u, p, x)$

$\mathcal{L}_u$ = partial derivative of $\mathcal{L}$ with respect to the I argument

$\mathcal{L}_p$ = partial derivative of $\mathcal{L}$ with respect to the II argument.

Remark: if $\mathcal{L}$ does not depend on p then E-L equations are just

$$\mathcal{L}_u = 0$$

$$L(f) = \frac{1}{\pi} \int_{\mathbb{R}^2} (f(x) - y)^2\, e^{-(x-y)^2} e^{-x^2}\, dx\, dy.$$

$$= \int \mathcal{L}(f(x), x)\, dx$$

$$L(f) = \frac{1}{\pi} \int_{\mathbb{R}^2} \left(f(x)^2 + y^2 - 2y f(x) \right) e^{-(x-y)^2} e^{-x^2} \, dx\, dy$$

$$= \frac{1}{\pi} \int_{\mathbb{R}} \left[f^2(x)\, e^{-x^2} \left(\int_{\mathbb{R}} e^{-(x-y)^2} \, dy \right) + e^{-x^2} \left(\int_{\mathbb{R}} y^2 e^{-(x-y)^2} \, dy \right) \right.$$

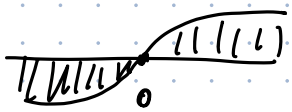
$$\left. - 2 f(x) e^{-x^2} \left(\int_{\mathbb{R}} y e^{-(x-y)^2} \, dy \right) \right] dx$$

$$\int_{\mathbb{R}} e^{-(x-y)^2} \, dy = \sqrt{\pi}$$

$$\int_{\mathbb{R}} y^2 e^{-(x-y)^2} \, dy \qquad y - x = z \qquad y = z + x$$

$$\int_{\mathbb{R}} (z+x)^2 e^{-z^2} \, dz = \int_{\mathbb{R}} \left(z^2 + 2xz + x^2 \right) e^{-z^2}$$

$$= \int_{\mathbb{R}} z^2 e^{-z^2} \, dz + \underbrace{2x \int_{\mathbb{R}} z e^{-z^2} \, dz}_{=\,0} + x^2 \int_{\mathbb{R}} e^{-z^2} \, dz.$$



$$\int_{\mathbb{R}} z^2 e^{-z^2} \, dz = -2 \cdot \frac{1}{2} \int_{\mathbb{R}} z \cdot \frac{d}{dz} e^{-z^2} \, dz = 2 \cdot \frac{1}{2} \int_{\mathbb{R}} e^{-z^2} \, dz = 2 \cdot \frac{1}{2} \sqrt{\pi}$$

$$\boxed{\int_{\mathbb{R}} y^2 e^{-(x-y)^2} \, dy = 2\sqrt{\pi} + x^2 \sqrt{\pi}}$$

$$y - x = z$$

$$\int_{\mathbb{R}} y e^{-(x-y)^2} \, dy = \int_{\mathbb{R}} (x+z) e^{-z^2} \, dz = x\sqrt{\pi}$$

$$L(f) = \frac{1}{\pi} \int_{\mathbb{R}} \left[\underbrace{f^2(x) \sqrt{\pi}\, e^{-x^2} + e^{-x^2} \left(2\sqrt{\pi} + x^2 \sqrt{\pi} \right) - 2 f(x) \sqrt{\pi}\, x\, e^{-x^2}}_{\mathcal{L}(f)} \right] dx$$

So the stationarity conditions for L, which are the E-L equations

$$\mathcal{L}_f = 0$$

$$2f(x)\sqrt{\pi}\, e^{-x^2} - 2\sqrt{\pi}\, x\, e^{-x^2} = 0$$

$$e^{-x^2} > 0$$

$$\boxed{f(x) = x}$$

We still need to check that this is a minimum.

In this example I gave you the True distribution of data and therefore I could find explicitly a minimum of the functional risk.

However, in general, we NEVER know the true distribution π but we only can aspire to know independent sample from π. Meaning that we usually know a data set

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N) : (x_i, y_i) \in \Omega\}.$$

The main idea then is to approximate π with

$$\pi_N = \left(\frac{1}{N}\right) \sum_{i=1}^{N} \delta_{(x_i, y_i)}$$

π_N: empirical distribution

What happens when you use π_N instead of π in the definition of the risk

$$\int_\Omega \ell(f(x), y)\, d\pi_N(x, y) = \frac{1}{N} \sum_{i=1}^{N} \int_\Omega \ell(f(x), y)\, d\delta_{(x_i, y_i)}$$

$$\boxed{L_N(f) := \frac{1}{N} \sum_{i=1}^{N} \ell(f(x_i), y_i)}$$

Empirical Risk

Functional Risk $\longrightarrow$ Empirical Risk

Functional Space $\longrightarrow$ $\mathbb{R}^n$ (finite dimensional space)

Basically we restrict the domain of optimization only to PARAMETRIC FUNCTIONS

$$\mathcal{P} = \{ g \in \mathrm{dom}(L) : g(x) = f(x; w) \quad w \in \mathbb{R}^n \}.$$

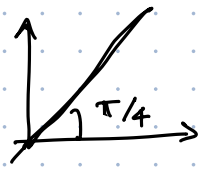
We will choose $\mathcal{P}$ to be the set of (let us say all neural networks with some architecture.

$$\min_{w \in \mathbb{R}^n} \; L_N\left(f(\cdot; w) \right)$$

We have shown that $f^*(x) = x$ is a stationary point (a candidate to be a minimum) of

$$L(f) = \frac{1}{\pi} \int_{\mathbb{R}^2} (f(x) - y)^2 \, e^{-(x-y)^2} e^{-x^2} \, dx \, dy$$

$f^* =$



$L(f^*)$?

$$L(f^*) = \frac{1}{\pi} \int_{\mathbb{R}^2} (x-y)^2 \, e^{-(x-y)^2} e^{-x^2} \, dx \, dy$$

$$= \frac{1}{\pi} \int_{\mathbb{R}} e^{-x^2} \left(\int_{\mathbb{R}} (x-y)^2 e^{-(x-y)^2} dy \right) dx$$

$\int_{\mathbb{R}} (x-y)^2 e^{-(x-y)^2} dy = 2\sqrt{\pi}$ $\qquad x - y = t$

We have already computed

$$\int_{\mathbb{R}} z^2 e^{-z^2} dz = \frac{\sqrt{\pi}}{2}$$

$$L(f^*) = 2 \frac{\sqrt{\pi}}{\pi} \int_{\mathbb{R}} e^{-x^2} dx = \frac{1}{2}$$

$$\boxed{L(f^*) = \frac{1}{2}}$$

$$L(f) \geq \frac{1}{2} \quad \forall f \in \mathrm{dom}(L)$$