19/03/2025

What we usually do in ML is to minimize the Empirical Risk

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

on what we called $\mathcal{P}$. From now on we will always assume that

$\mathcal{P}$ = "set of NN with a fixed architecture"

$f \in \mathcal{P}$ can always be represented a

$$f(\cdot; w)$$

w ≈ weights of the NN

Therefore if we use this representation in the empirical risk we obtain the following

$$L_N(f(\cdot; w)) = \frac{1}{N} \sum_{i=1}^{N} \ell(f(x_i; w), y_i)$$

$\ell(a,b) = \frac{1}{2}|a-b|^2$

Rem. Once ℓ is fixed, for istance is a quadratic loss or Cross Entropy loss or ....

since (x_i, y_i) are also given because they are the points of a given training set then

■ = is a function solely of the weights.

$$L_N(f(\cdot, w)) = \hat{L}_N(w) \quad \forall w \in \mathbb{R}^n$$

n = number of weights of the NN

$n = 6$

Therefore the minimization of L_N is equivalent to the minimization of $\hat{L}_N$

Side note

$$x_i = \sigma\left(\sum_j w_{ij} x_j\right)$$

$$= \sigma\left(\sum_j w_{ij} x_j + b_i\right)$$

↑ bias.

The problem that we need to solve is

$$\min_{w \in \mathbb{R}^n} \hat{L}_N(w)$$

Rem We are going to solve tackle this problem using "gradient-based" methods, which of course involve the computation of $\nabla \hat{L}_N$

$$\hat{L}_N(w) = \frac{1}{N} \sum_{i=1}^{N} \ell(f(x_i; w), y_i),$$

$$\nabla \hat{L}_N(w) = \frac{1}{N} \sum_{i=1}^{N} \partial_1 \ell(f(x_i, w), y_i) \, \partial_w f(x_i, w)$$

↓

yesterday we called this $\varphi(w)$

2. So $\nabla_1 \ell$ is usually given. For example: if $\ell(a,b) = \frac{1}{2}(a-b)^2$

$$\nabla_1 \ell(a,b) = (a-b)$$

2. $\nabla_w f(x_i, w)$ is the gradient of a NN with respect to its weights.
And we saw that this can be computed efficiently $O(n)$ using backpropagation

Continuous Optimization

New Notation

$\hat{L}_N \to f$

$w \to x$

Brief introduction on Hilbert spaces (and informal)

We say that X is an Hilbert space if it is a vector space (you know how to + and multiply by scalars) with an inner product

$$\_ \cdot \_ : X \times X \to \mathbb{R}$$

$$(x,y) \longmapsto x \cdot y$$

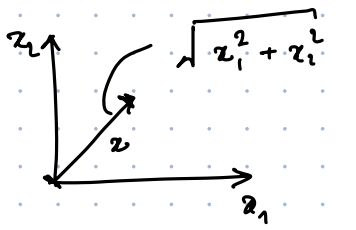
if $X = \mathbb{R}^n$ a natural choice for $\cdot$ is the standard scalar product in $\mathbb{R}^n$

$$x \cdot y = \sum_{i=1}^{n} x_i y_i$$

Once you have an inner product then you can define a norm $\| \cdot \|_X$

$$\| x \|_X = \sqrt{x \cdot x}$$

In $\mathbb{R}^n$ $|x| = \sqrt{\sum_{i=1}^{n} (x_i)^2}$



X is Hilbert if it is complete with respect to this norm

complete = Cauchy sequences converge.

A sequence (x_n) is a Cauchy sequence if $\forall \varepsilon > 0$ $\exists M > 0$ s.t. $\forall m, n > M$

$$\| x_n - x_m \|_X < \varepsilon .$$

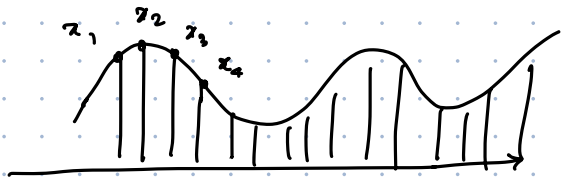
Example 1. $\ell^2 = \{ (x_n) : \sum_{n=1}^{+\infty} x_n^2 < +\infty \}$.

$x_1, x_2, x_3, x_4, \ldots$

and $x \cdot y = \sum_{i=1}^{+\infty} x_i y_i$

Ex. Prove that l^2 is complete and therefore it is Hilbert.

Example 2.



$$L^2(a,b) = \{ f : (a,b) \to \mathbb{R} : \int_a^b f^2 \, dx < +\infty \}.$$

$$f \cdot g = \int_a^b f \cdot g \, dx$$

Now the problem that we want to solve is:

$$\min_{x \in X} f(x) \quad \text{where } X \text{ is Hilbert and } f : X \to \mathbb{R}.$$

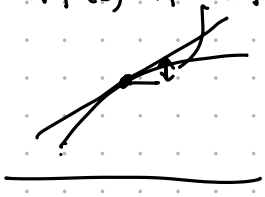
Gradients

X^* is called the dual space of X and it is the space of all continuous linear functionals on X

$$X^* = \{ L : X \to \mathbb{R} : L \text{ is linear and continuous} \}.$$

$$L(\alpha x + \beta y) = \alpha L(x) + \beta L(y)$$

Def. The differential of f at a point x that we denote $df(x)$ is if it exists an element of X^* such that



$$f(y) = f(x) + \langle df(x), x - y \rangle + o(|x-y|)$$

$$\langle L, x \rangle = L(x) \quad \text{if } L \in X^* \text{ and } x \in X$$

If the differential $df(x)$ exists for every $x \in X$ then we call f differentiable and we write

$$f \in C^1(X; \mathbb{R})$$

Many times we say that f is Fréchet - differentiable.

Theorem (Riesz) If X is hilbert then every element $L \in X^*$ can be represented as follows

$$\forall x \in X \qquad \underbrace{\langle L, x \rangle}_{L(x)} = v \cdot x \qquad (*)$$

in the sense that it exists a vector $v \in X$ such that $(*)$ holds.

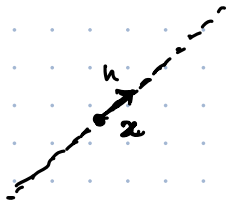
Since X is Hilbert $df(x) \in X^*$ by Riesz $\exists$ a vector in X that we denote $\nabla f(x)$ (and we call it gradient of f in the point x) such that

$$\langle df(x), z \rangle = \nabla f(x) \cdot z \qquad \forall z \in X$$

In infinite dimensions (in Hilbert) the generalization of directional derivative is called Gateau derivative.

Fix $x \in X$ and define $\phi(t) = f(x + th)$ where $h \in X$ and $h \neq 0$.

↑ I'm looking at how f behaves around x in the direction suggested by h



$$df(x;h) = \phi'(0)$$ is defined to be the "Gateau derivative / differential" along the direction h.

In the Calculus of Variations this is also what is usually called first variation.

Observation: if f is differentiable in x, then

$$f(x+th) = f(x) + \langle df(x), th \rangle + o(|th|)$$

$$\frac{f(x+th) - f(x)}{t} = \langle df(x), h \rangle + o(1)$$

$(o(1) = o(|h|)$ for functions of the variable $t)$

If you take the limit $t \to 0$. , $o(1) \to 0$

$$\boxed{f(x+th) = \phi(t)}$$
$$f(x+0) = \phi(0)$$

So $\frac{f(x+th) - f(x)}{t} = \frac{\phi(t) - \phi(0)}{t}$

as $t \to 0$ $\frac{\phi(t) - \phi(0)}{t} \to \phi'(0) \equiv d$

So we obtain the usual relation between directional derivative and partial derivatives

$$\boxed{df(x;h) = \langle df(x), h \rangle}$$

Proposition 1. Let X be Hilbert and $f \in C^1(X;\mathbb{R})$, $x \in X$ such that $df(x) \neq 0$.

Let $\nabla f(x)$ be the gradient of f at x (Riesz). Then

$$\frac{\nabla f(x)}{\|\nabla f(x)\|} \in \operatorname{argmax}\{|df(x;h)| : h \in X \ \|h\| = 1\}.$$

the proof of this is basically an application of Cauchy Schwartz inequality.

Cauchy-Schwartz: $\forall\, x, y \in X$

$$|x \cdot y| \leq \|x\| \cdot \|y\|$$

and $=$ holds only if $x = \alpha y$ where $\alpha \in \mathbb{R}$.

Let us now try to prove Proposition 1.

Choose $h = \dfrac{\nabla f(x)}{\|\nabla f(x)\|}$

$$df(x;h) = \nabla f(x) \cdot \frac{\nabla f(x)}{\|\nabla f(x)\|} = \frac{\|\nabla f(x)\|^2}{\|\nabla f(x)\|} = \|\nabla f(x)\|.$$

Now for any other direction

$$|df(x;h)| = |\nabla f(x) \cdot h| \leq \|\nabla f(x)\| \cdot \|h\| = \|\nabla f(x)\|.$$

This is basically the observation that motivates gradient based optimization algorithms whose most famous representative is GD. (gradient descent.

$$\boxed{x^{k+1} = x^k - \tau \nabla f(x^k)}$$ Gradient Descent.

Curves of maximal slope. (Ennio De Giorgi)

Are a way to extend gradient flows in metric spaces.

Metric spaces $= (\mathcal{J}, d)$

$\mathcal{J}$: space; d: distance

Gradient Flows.

Are solutions of

$$\begin{cases} u' = -\nabla f(u) \\ u(0) = u_0 \end{cases}$$ It is an ODE on $(0, +\infty)$

Gradient flows are "the continuous version" of Gradient descent.

As often happens it is easier to analyze dynamics in continuous time. because often you have simple closed form solutions..

Example of GF

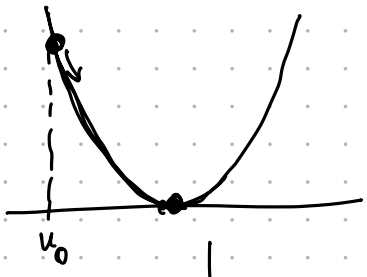
$f(u) = \frac{u^2}{2}$

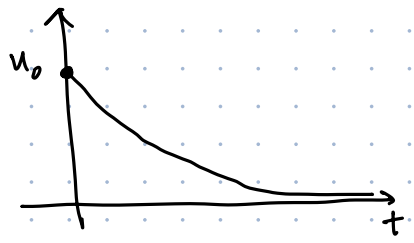
(*) $\begin{cases} u'(t) = -u(t) \\ u(0) = u_0 \end{cases}$ $\forall\, t > 0$

The general solution is $u(t) = A e^{-t}$

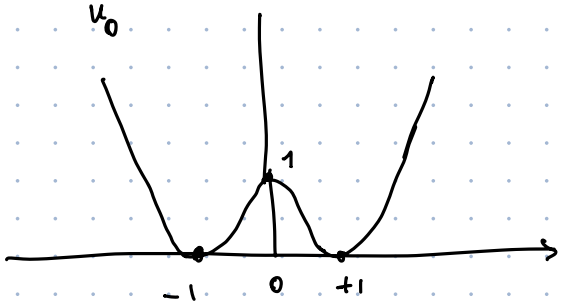
$u(0) = A = u_0$

The solution of (*) is $u(t) = u_0 e^{-t}$





Exercise



$(x - (x-1)^2)^2$

Double Well or "Mexican hat"

$f(u(t))$ this decreases and indeed

$$\frac{d}{dt} f(u(t)) = \nabla f(u(t)) \cdot u'(t) = -(\nabla f(u(t)))^2 \leq 0$$

Curves of maximal slope ("Gradient Flows in Metric Spaces" Ambrosio, Gigli, Savaré)

↑ Very hard/difficult book

$u'(t) = -\nabla\phi(u(t))$ Gradient Flow

In order to write this equation you need two things:

1. u' derivative of $t \mapsto u(t)$
2. $\nabla\phi(u(t))$ you need the gradient. (for instance if $\phi \in C^1$ in hilbert or banach ok!)

If you don't have this two objects you can't even write the equation.
In metric spaces you don't have this.

IDEA Even you can't define u' and $\nabla\phi$ you can define something that generalises $|u'|$ and $|\nabla\phi|$.

For instance if $t \to v(t)$ is a curve in $\mathcal{S}$ and it is absolutely continuous. then you can define

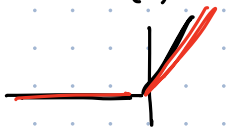
$$|v'|(t) = \lim_{h\to 0} \frac{d(v(t+h), v(t))}{h} \leftarrow \text{metric derivative.}$$

Obs. You do not define v' and then you take the norm. (You can't do this) you directy define $|v'|$

In the same way you can define what is called an upper gradient which is a surrogate of the $|\nabla\phi|$.

Usually the interesting object is the "Local slope"

$(\ )^+ = \text{ReLu}(\ )$



$$|\partial\phi|(v) = \limsup_{w\to v} \frac{(\phi(w) - \phi(v))^+}{d(w,v)}$$

Ok I have metric analogues of $|u'|$ and $|\nabla\phi(u)|$ but the gradient flow equation is not just a statement on norms of u' an $\nabla\phi$, direction are important.

In other words

$$u' = -\nabla\phi(u) \quad \xrightarrow{\text{take } |\cdot|} \quad |u'| = |\nabla\phi|$$

$(\bullet)$ $\qquad$ $(\bullet\ \bullet)$

(••) is surely not equivalent to (•)

The informative content of (••) is less that that of (•)

De Giorgi idea:

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

(hilbert; you are recovering the sign of (•))

$$= -\frac{1}{2}|D\phi|^2 - \frac{1}{2}|u'|^2$$

(by using (••))

Now (•) is equivalent to

- $\frac{d}{dt}\phi(u(t)) = -\frac{1}{2}|u'|^2 - \frac{1}{2}|\nabla\phi(u)|^2$.

$$\int_s^t \frac{1}{2}|u'(r)|^2 + \frac{1}{2}|\nabla\phi(u(r))|^2\,dr = \phi(u(s)) - \phi(u(t))$$

You say that u is a curve of maximal slope if it exists g such that

$$\int_s^t \frac{1}{2}|u'(r)|^2 + \frac{1}{2}|g(u(r))|^2\,dr \leq \phi(u(s)) - \phi(u(t))$$

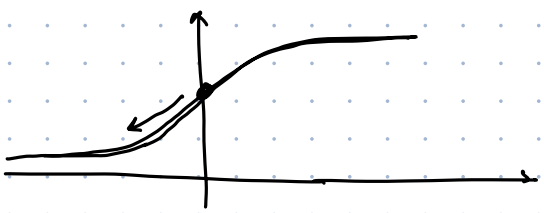
($=$)

$\forall$ $\mathcal{L}^1$-a.e. t, s with $s<t$.

Proof of convergence of GD under assumptions that are ML-friendly.

Basically we want to give a proof of convergence of GD without using the usual convexity assumption.

Next time we'll try to consider also the convex case to see what we can say more that the non convex



✓ Hilbert.

What is GD? It is a procedure that computes a sequence of points $(x^n) \subset X$ defined as follows

$$(GD) \quad \begin{cases} 1. \quad x^0 = x_0 \in X & x_0 \text{ is a fixed point.} \\ 2. \quad x^{k+1} = x^k - \tau \nabla f(x^k) & k = 0, 1, 2, 3, \dots \quad \tau > 0 \end{cases}$$

↘ in ML is called "learning rate"

Obs 1. $x^{k+1} = x^k - \tau \nabla f(x^k)$ is an explicit Euler scheme with step size τ for the gradient flow $x' = -\nabla f(x)$

Explicit Euler = Forward Euler
Implicit Euler = Backward Euler.

$x' = F(x) \rightarrow x^{k+1} = x^k + \tau F(x^k)$ explicit

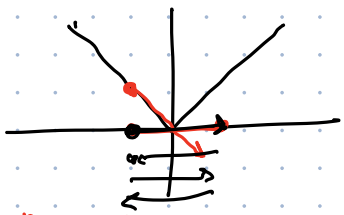
↓

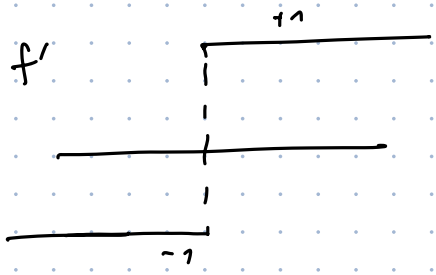
$x^{k+1} = x^k + \tau F(x^{k+1})$ implicit.

Theorem (Convergence of GD): Let $f \in C^1(X; \mathbb{R})$ and suppose that ∇f is L-Lipshitz, then if $0 < \tau < \frac{2}{L}$ then (GD) converges in the sense that

$$|\nabla f(x^k)| \to 0 \quad \text{as} \quad k \to +\infty$$

Ex $f(x) = |x|$





$x^{k+1} = x^k - \tau(-1)$

$= x^k + \tau$

$x^{k+2} = x^{k+1} - \tau(1) = x^{k+1} - \tau = x^k$