Departamento de Matemáticas, Facultad de Ciencias
Universidad Autónoma de Madrid

Teoría de flujos gradientes y aplicaciones a redes neuronales.

Trabajo de fin de grado

Grado en Matemáticas

Autor: Jorge Moral
Tutor: Matteo Bonforte

Curso 2023-2024

Resumen

La teoría de flujos gradientes es una herramienta bastante útil a la hora de comprender y analizar sistemas dinámicos. En esencia, se centra en visualizar la evolución de un sistema como el movimiento de partículas en un campo de fuerza, donde la dirección del flujo está determinada por el gradiente de una función de energía o potencial. Este enfoque proporciona una perspectiva intuitiva para entender cómo los sistemas se mueven hacia estados de menor energía o mayor estabilidad.

En el contexto de las redes neuronales, la teoría de flujos gradientes juega un papel fundamental. Las redes neuronales se entrenan mediante un proceso iterativo de ajuste de los pesos de las conexiones entre las neuronas para minimizar una función coste, que representa la discrepancia entre las predicciones del modelo, basadas en un conjunto de muestras de entrenamiento, y los resultados obtenidos con los outputs de la red neuronal. Este proceso de optimización se puede entender como un flujo en el espacio de parámetros de la red neuronal, donde la dirección del flujo está determinada por el gradiente de la función de pérdida con respecto a los pesos de la red neuronal. En el caso particular de este trabajo, se intenta optimizar el coste de la red neuronal empleando el método del descenso del gradiente estocástico que lleva consigo asociado una ecuación de Fokker-Plank. En un modelo general, la ecuación de Fokker-Planck asociada tiene un carácter muy degenerado, dado que no es necesariamente cierto que la función de coste sea globalmente λ-convexa. Trabajaremos con un modelo simplificado para poder explotar las herramientas de la teoría de flujos gradientes en el caso de energías globalmente λ-convexas. El caso general todavía es un problema abierto.

Abstract

The theory of gradient flows is a rather useful tool when it comes to understanding and analyzing dynamic systems. Essentially, it focuses on visualizing the evolution of a system as the movement of particles in a force field, where the flow direction is determined by the gradient of an energy or potential function. This approach provides an intuitive perspective for understanding how systems move towards states of lower energy or higher stability.

In the context of neural networks, the theory of gradient flows plays a fundamental role. Neural networks are trained through an iterative process of adjusting the weights of connections between neurons to minimize a loss or cost function, which represents the discrepancy between the model's predictions, based on a set of training samples, and the results obtained with the outputs of the neural network. This optimization process can be understood as a flow in the parameter space of the neural network, which reflects the architecture and composition of the neural network, where the flow direction is determined by the gradient of the loss function with respect to the weights of the neural network. In the particular case of this work, the aim is to optimize the cost of the neural network using the stochastic gradient descent method, which has a Fokker-Planck equation associated. In a general model, the associated Fokker-Planck equation is highly degenerate, as it is not necessarily true that the cost function is globally λ-convex. We will work with a simplified model to exploit the tools of gradient flow theory in the case of globally λ-convex energies. The general case remains an open problem.

Índice general

CAPÍTULO 1
Introducción y preliminares

1.1. Inroducción

La teoría de flujos gradientes nos permite estudiar la evolución temporal de sistemas dinámicos en términos de gradientes de funciones de energía o potenciales. En esencia, considera que el comportamiento de un sistema está determinado por la tendencia a minimizar o maximizar una cierta función de energía, similar al movimiento de una partícula que está sometida a las interacciones y fuerzas del campo sobre el que está.

Este enfoque se basa en la idea fundamental de que los sistemas naturales tienden a evolucionar hacia estados de menor energía, reflejando un principio básico en física y termodinámica. La función de energía o potencial asociada describe cómo varía la energía del sistema en función de su estado, y el gradiente de esta función indica la dirección y la tasa de cambio más rápida en el espacio de estados del sistema.

Por otro lado, uno de los aspectos más importantes sobre la teoría de flujos gradientes es que nos proporciona una interpretación geométrica e intuitiva sobre la dinámica y evolución del sistema que estamos estudiando y de como está afectado por las condiciones y restricciones iniciales del propio sistema.

Antes de empezar a desarrollar la teoría de flujos gradientes es necesario asentar las bases sobre la noción de convexidad de una función $f : \mathbb{R}^n \to \mathbb{R}$ y varios resultados claves asociados a la convexidad de la misma. Esto se debe a que la teoría de flujos gradientes puede ser vista como una generalización sobre la teoría de funciones convexas, pero ahora sobre espacios de Hilbert más generales. En concreto, para las aplicaciones que presentaremos al final del trabajo tomaremos como espacio de Hilber $L^2(\mathbb{R}^k, \mu)$ y asumiremos que estamos trabajando con potenciales convexos.

Una vez que hayamos profundizado en la teoría de flujos gradientes, podremos beneficiarnos de una serie de resultados fundamentales que nos permitirán abordar aspectos cruciales en la optimización de funciones. Entre estos resultados destaca el teorema de Brezis-Komura, que presentaremos y estudiaremos a fondo en el segundo capítulo de este trabajo, del cual obtendremos la existencia y unicidad de un flujo gradiente, así como métodos para determinar la dirección de máximo decrecimiento

de la función. Estas nociones de existencia y unicidad serán fundamentales para establecer la garantía de que, bajo ciertas condiciones, nuestro proceso de optimización convergerá a un mínimo local o global.

Una vez se haya desarrollado toda la teoría fundamental sobre los flujos gradientes, presentaremos como puede ser empleada en casos prácticos de la actualidad, en concreto, nos centraremos en su aplicación a las redes neuronales. Para ello, haremos una introducción sobre la composición y funcionamiento de las redes neuronales y sobre el método de aprendizaje de la misma. Este método de aprendizaje trata de minimizar cierta función de coste dada por la diferencia entre los outputs recibidos y esperados de una muestra de entrenamiento de partida. Para minimizar esta función de coste emplearemos el método del descenso del gradiente estocástico, que se trata de un proceso estocástico discreto. Este proceso discreto puede ser aproximado por un proceso continuo que viene dado por una ecuación diferencial estocástica cuya función de densidad de probabilidad es la solución de una ecuación de tipo Fokker-Planck.

Es fundamental señalar que la minimización de la función de coste asociada a una red neuronal general es un problema extremadamente complejo y que continúa siendo un tema de investigación activo en la actualidad. Para este trabajo, con el propósito de emplear las herramientas de la teoría de flujos gradientes, utilizaremos un modelo simplificado en el cual asumiremos que nuestra función de coste es convexa.

La última parte de este trabajo se centrará en estudiar la ecuación de Fokker-Plank, viendo su relación con la ecuación clásica del calor, y luego, entendiéndola como la trayectoria del descenso del gradiente de un cierto potencial de energía.

1.2. Resultados preliminares

En esta primera sección del trabajo haremos una breve introducción sobre la teoría de las funciones convexas unidimensionales y multidimensionales, y enunciaremos algún resultado importante de forma que nos permita generalizar estos conceptos a la teoría de flujos gradientes.

Definición 1.1. (Conjunto convexo). Sea $S \subset \mathbb{R}^n$ un conjunto distinto del vacío. Decimos que S es convexo si para todo par de puntos $x, y \in S$ y para todo $\alpha \in [0,1]$ se cumple

$$\alpha x + (1-\alpha)y \ \in S$$

A la combinación $\alpha x + (1-\alpha)y$ la llamaremos una combinación convexa entre los vectores x e y.

Definición 1.2. (Función convexa). Sea $S \subset \mathbb{R}^n$, $S \neq \emptyset$ un conjunto convexo. Diremos que una función $f : S \to \mathbb{R}$ es convexa si para todo $\alpha \in [0,1]$ y para todo par de vectores $x, y \in S$ satisface

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$$

Diremos que una función g es cóncava si bajo las mismas condiciones se satisface que

$$g(\alpha x + (1-\alpha)y) \geq \alpha g(x) + (1-\alpha)g(y)$$

Teorema 1.3. *Sea $S \subset \mathbb{R}^n$ un conjunto abierto y convexo y sea $f : S \to \mathbb{R}$ una función convexa. Entonces se tiene que f es una función continua.*

Definición 1.4. (Epigrafo e Hipografo). Sea $S \subset \mathbb{R}^n$ un conjunto convexo y sea $f : S \to \mathbb{R}$ una función convexa. Definimos los conjuntos

1. $epi(f) = \{(x, y) \in \mathbb{R}^{n+1} : \ x \in S, f(x) \leq y\}$
2. $hipo(f) = \{(x, y) \in \mathbb{R}^{n+1} : \ x \in S, f(x) \geq y\}$

Teorema 1.5. *Sea $S \subset \mathbb{R}^n$ un conjunto convexo y sea $f : S \to \mathbb{R}$ una función. Entonces f es convexa si y solo si $epi(f)$ es un conjunto convexo.*

Definición 1.6. (Subgradiente de una función convexa). Sea $S \subset \mathbb{R}^n$ un conjunto convexo y sea $f : S \to \mathbb{R}$ una función convexa. Se dice que el vector $\xi \in \mathbb{R}^n$ es un subgradiente de f en el punto x_0 si cumple

$$f(x) \geq f(x_0) + \xi \cdot (x - x_0) \quad \forall x \in S.$$

Ahora nos enfocaremos en un caso particularmente importante dentro de la teoría de funciones convexas, donde además de ser convexa, nuestra función también es derivable. Esta extensión de la teoría de funciones convexas nos permite estudiar propiedades aprovechando los conocimientos que disponemos sobre el cálculo diferencial.

Proposición 1.7. Sea $S \subset \mathbb{R}^n$ un conjunto abierto y convexo y sea $f : S \to \mathbb{R}$ una función convexa y derivable. Entonces, para todo $x_0 \in S$ existe un único subgradiente $\xi \in \mathbb{R}^n$ de f en x_0, además $\xi = \nabla f(x_0)$.

Teorema 1.8. *Sea $S \subset \mathbb{R}^n$ un conjunto abierto y convexo y sea $f : S \to \mathbb{R}$ una función derivable. Entonces f es convexa si y solo si para todo $x_0 \in S$ satisface*

$$f(x) \geq f(x_0) + \nabla f(x_0) \cdot (x - x_0) \quad \forall x \in S.$$

Si ahora aumentamos las restricciones pidiendo que nuestra función $f : S \to \mathbb{R}$ sea de clase $C^2(S)$. Entonces podremos dar la siguiente caracterización de función convexa.

Teorema 1.9. *Sea $f : S \to \mathbb{R}$ una función de clase $C^2(S)$. Entonces f es una función convexa si y solo si la matriz Hessiana $H_f(x)$ es semidefinida positiva para todo $x \in S$.*

A continuación, vamos a exponer una serie de resultados vinculados con la minimización de funciones convexas. Estos resultados son fundamentales para la aplicación de la teoría de funciones convexas y la teoría de flujos gradientes en campos como la programación lineal, problemas de optimización y, en nuestro caso, para la aplicación a redes neuronales.

El problema que vamos a estudiar es minimizar una función convexa f sobre un conjunto factible S también convexo, es decir, hallar

$$\min_{x \in S} f(x). \tag{1.1}$$

Teorema 1.10. *Sea $S \subset \mathbb{R}^n$ un conjunto convexo y sea $f : S \to \mathbb{R}$ una función convexa. Consideramos el problema de minimizar $f(x)$ sujeto a $x \in S$. Supongamos que tenemos $\bar{x}$ es una solución local del problema, es decir, que existe un entorno U de $\bar{x}$ tal que se cumple que $f(y) \geq f(\bar{x})$ para todo $y \in U$. Entonces $\bar{x}$ es una solución global del problema. Si además $\bar{x}$ es un mínimo local estricto o la función f es estrictamente convexa, entonces $\bar{x}$ es la única solución global del problema.*

Teorema 1.11. *Sea $S \subset \mathbb{R}^n$ un conjunto convexo y sea $f : S \to \mathbb{R}$ una función convexa. Consideramos el problema de minimizar $f(x)$ sujeto a $x \in S$. Entonces un punto $\bar{x} \in S$ es un mínimo global si y solo si f tiene un sub gradiente ξ en $\bar{x}$ tal que*

$$\xi^t(x - \bar{x}) \geq 0 \quad \forall x \in S$$

Proposición 1.12. Si consideramos el problema descrito en 1.1, y suponemos además que f es una función diferenciable, entonces un punto $\bar{x} \in S$ es un mínimo global si y solo si

$$\nabla f(\bar{x})^t(x - \bar{x}) \geq 0 \quad \forall x \in S$$

Corolario 1.1. *En el caso que nuestro conjunto $S = \mathbb{R}^n$, entonces la condición se reduce a ver si para $\bar{x}$ se cumple*

$$\nabla f(\bar{x}) = 0.$$

CAPÍTULO 2
Teoría de flujos gradientes

Esta sección se centrará en una introducción a la teoría y aplicaciones de los flujos gradientes. Para ello, estaremos trabajando en un espacio de Hilbert que denotaremos como $\mathcal{H}$ con un producto escalar $\langle \cdot, \cdot \rangle$ y su norma asociada $\|\cdot\|$, mientras que $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ será una función cuyo dominio finito $\{f < \infty\}$ se denotará por $\mathcal{D}(f)$.

El caso clásico es cuando estamos considerando nuestro espacio de Hilbert $\mathcal{H} = \mathbb{R}^n$, siendo $f : \mathbb{R}^n \to \mathbb{R}$ una función diferenciable en todo el espacio. Entonces un flujo gradiente de f empezando en $\bar{x} \in \mathcal{H}$ es una solución a la ecuación diferencial ordinaria

$$\begin{cases} \text{x'(t)=-}\nabla f(x(t)) & t > 0 \\ x(0) = \bar{x}. \end{cases}$$

Si nuestra función f es además lipschitz, entonces la existencia y unicidad viene garantizada del teorema de Cauchy-Lipschitz. Para los resultados de este trabajo, la condición de que f sea derivable y lipschitz puede ser demasiado restrictiva, ya que en las aplicaciones sobre ecuaciones en derivadas parciales donde nos gustaría encontrar soluciones en espacios de funciones, a priori, la diferenciabilidad de una función falla, y por eso se imponen unas condiciones de regularidad un poco más débiles que son la convexidad y la semicontinuidad inferior de la función f.

2.1. Funciones λ-convexas

En esta sección permitiremos una generalización aún mayor de nuestra función f que es la λ-convexidad de la misma.

Definición 2.1. (λ-convexidad). Dado $\lambda \in \mathbb{R}$, decimos que una función f es λ-convexa si $f - \frac{\lambda}{2} \|\cdot\|^2$ es convexa. Se puede notar que para los valores $\lambda < 0$ se tiene una noción más débil de convexidad.

Definición 2.2. (λ-Subdiferencial). El λ-subdiferencial de una función f en $x \in \mathcal{D}(f)$ es el conjunto

$$\partial_\lambda f(x) := \left\{ p \in \mathcal{H} : f(y) \geq f(x) \ + \langle p, y - x \rangle \ + \frac{\lambda}{2} \|y - x\|^2 \ \forall y \in \mathcal{H} \right\} \tag{2.1}$$

2.2. Flujos gradientes de funciones λ-convexas

Definición 2.3. (Funciones absolutamente continuas). Una función $F : [a, b] \to \mathcal{H}$ se dice que es absolutamente continua si para todo $\epsilon > 0$ existe un $\delta > 0$ tal que para cualquier familia finita de sub intervalos disjuntos $\{(a_j, b_j)\}_{j=1}^n \subseteq [a, b]$ que cumpla $\sum_{j=1}^n (b_j - a_j) < \delta$, entonces se tiene que

$$\sum_{j=1}^{n} \|F(b_j) - F(a_j)\| < \epsilon \tag{2.2}$$

Al conjunto de todas las funciones absolutamente continuas en $[a, b]$ se las denota por AC([a,b]).

Definición 2.4. (Flujo gradiente de una función f λ-convexa). Decimos que $x : (0, +\infty) \to \mathcal{D}(f)$ es un flujo gradiente de f si $x \in AC_{loc}((0, +\infty); \mathcal{H})$ y

$$x'(t) \in -\partial_\lambda f(x(t)) \text{ para c.t.p } t \in (0, +\infty). \tag{2.3}$$

Decimos que x(t) empieza en $\bar{x} = \lim_{t \mapsto 0^+} x(t)$

Definición 2.5. ("Evolution Variational Inequality" EVI). Una curva localmente absolutamente continua $x := x(t)$ $(x \in AC_{loc}((0, \infty), \mathcal{H}))$ se denota como una solución EVI_λ si para todo $y \in \mathcal{D}(f)$ si cumple

$$\frac{d}{dt}\left(\frac{1}{2}\|x(t) - y\|^2\right) + \frac{\lambda}{2}\|x(t) - y\|^2 + f(x(t)) \leq f(y) \quad \text{para c.t.p } t \in (0, +\infty). \tag{2.4}$$

Decimos que $x(t)$ empieza en $\bar{x}$ si $\bar{x} = \lim_{t \to 0^+} x(t)$

Lema 2.6. *Para toda función $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ existe un conjunto $\mathcal{D} \subset \mathcal{H}$ que cumple que para todo $y \in \mathcal{D}(f)$ existe una sucesión $(y_n)_{n \in \mathbb{N}} \subset \mathcal{H}$ con $y_n \to y$ y $f(y_n) \to f(y)$.*

Definición 2.7. Una función $f : \mathcal{H} \to \mathbb{R}$ se dice que es semicontinua inferior en un punto $\bar{v} \in \mathcal{H}$ si cumple que para cualquier sucesión $v_n \to \bar{v}$ se tiene que $\liminf_{v_n \to \bar{v}} f(v_n) \geq f(\bar{v})$.

Lema 2.8. *Sea $f : \mathcal{H} \to \mathbf{R} \cup \{+\infty\}$ una función semicontinua inferior. Entonces, para todo $\lambda \in \mathbf{R}$ y $\bar{x} \in \mathcal{D}(f)$ existe como mucho una solución EVI_λ x(t) que empieza en $\bar{x}$.*

Teorema 2.9. *(Caracterización de flujos gradientes). Para una función λ-convexa y semicontinua inferior, una curva localmente absolutamente continua x=x(t) es un flujo gradiente si y solo si es una solución EVI_λ.*

Demostración. Sea x(t) un flujo gradiente, entonces x(t) satisface la propiedad 2.4 notando que

$$\frac{d}{dt}\left(\frac{1}{2}\|x(t) - y\|^2\right) = \langle x'(t), x(t) - y\rangle = \langle -x'(t), y - x(t)\rangle$$

y además, por la definición de flujo gradiente $x'(t) \in -\partial_\lambda f(x(t))$, lo que implica que

$$\langle -x'(t), y - x(t)\rangle \leq -\frac{\lambda}{2}\|x(t) - y\|^2 - f(x(t)) - f(y) \quad \forall y \in \mathcal{H}$$

Ahora veamos el recíproco. Sea x una solución EVI, haciendo uso del lema 2.6 sabemos que para todo $y \in \mathcal{D}$ (nuestro conjunto denso en energía) se cumple la desigualdad

$$\langle -x'(t), y_n - x(t)\rangle + \frac{\lambda}{2}\|x(t) - y_n\|^2 + f(x(t)) \leq f(y_n) \quad c.t.p \;\; t \in (0, +\infty).$$

Pasando al límite cuando $n \to \infty$ obtenemos

$$\langle -x'(t), y - x(t)\rangle + \frac{\lambda}{2}\|x(t) - y\|^2 + f(x(t)) \leq f(y)$$

□

Definición 2.10. Denotamos como ∇f al elemento de norma mínima del conjunto $\partial_\lambda f(x)$ cuando $\partial_\lambda f(x) \neq \emptyset$.

Proposición 2.11. (Propiedades de un flujo gradiente). Dado x(t):=$S_t\bar{x}$ un flujo gradiente de una función λ-convexa y semicontinua inferior, se cumplen

1. La derivada métrica por la derecha

   $$\left|x'_+(t)\right| = \lim_{h\to 0^+} \frac{\|x(t+h) - x(t))\|}{h}$$

   existe para todo t>0 y se tiene que la función $t \to e^{\lambda t}\left|x'_+(t)\right|$ no es creciente en el intervalo $(0, \infty)$

2. La función $t \to f(x(t))$ es localmente lipschitz, no creciente y se cumple

   $$\frac{d}{dt}f(x(t)) = \langle \nabla f(x(t)), x'(t)\rangle = -|\nabla f|^2(x(t)) \quad para\ c.t.p \;\; t \in (0, \infty)$$

   Donde $\nabla f(x(t))$ denota el único elemento de norma mínima en $\partial f(x(t))$

3. $x'(t) = -\nabla f(x(t))$ para c.t.p $t \in (0, +\infty)$

4. Para el caso particular $\lambda = 0$, para todo t>0 se tiene

   $$f(x(t)) \leq \inf_{v\in\mathcal{D}(f)}(f(v) + \frac{1}{2t}\|\bar{x} - v\|^2)$$

Teorema 2.12. *(Brézis-Komura). Asumamos que f es λ-convexa para algún $\lambda \in \mathbb{R}$ y además, que es semicontinua inferior. Entonces para todo $\bar{x} \in \overline{\mathcal{D}(f)}$ existe un único flujo gradiente $x(t) = S_t\bar{x}$ que empieza en $\bar{x}$. Además, la familia de operadores $S_t : \overline{\mathcal{D}(f)} \to \mathcal{D}(f)$ satisface:*

- *Propiedad de semigrupos: $S_{t+l} = S_t \circ S_l$*

- *Propiedad de contractividad:* $\|S_t\bar{x} - S_t\bar{y}\| \leq e^{-2\lambda t}\|\bar{x} - \bar{y}\| \quad \forall x, y \in \mathcal{H}$

Demostración. Empecemos viendo la unicidad y la propiedad de contractividad de nuestro flujo gradiente x(t). Asumamos $\lambda = 0$ por simplicidad, ya que el caso general $\lambda \in \mathbb{R}$ es análogo. Sean $x(t)$ y $\tilde{x}(t)$ dos flujos gradientes que empiezan en $\bar{x}$ y $\tilde{\bar{x}}$ respectivamente, sabemos que dados $p \in \partial f(y)$ y $q \in \partial f(z)$, tenemos que

$$\langle p - q, y - z\rangle \geq 0$$

para $y = x(t)$, $z = \tilde{x}(t)$, $p = -x'(t)$, $q = -\tilde{x}'(t)$ y t>0. Obtenemos de esta forma

$$\frac{d}{dt}\left(\|x(t) - \tilde{x}(t)\|^2\right) = 2\langle x'(t) - \tilde{x}'(t), x(t) - x'(t)\rangle \ \leq 0$$

Por tanto, la función $t \to \|x(t) - \tilde{x}(t)\|^2$ es decreciente y se tiene $\|x(t) - \tilde{x}(t)\| \leq \|x(s) - \tilde{x}(s)\|\ \forall s \in (0, t)$. Pasando al límite cuando $s \to 0^+$ se obtiene la propiedad de contractividad. Para la unicidad hay que notar que para una misma condición inicial $\bar{x} = \tilde{x} \in \mathcal{D}(f)$ se tendría que $\|x(t) - \tilde{x}(t)\| \leq 0$, de lo que se deduce qué $x(t) = \tilde{x}(t)$.

El resto de la demostración consistirá en construir nuestro flujo gradiente y para ello, dividiremos la demostración en varios pasos diferentes.

Paso 1: Reducimos nuestro problema al caso de energía finita, es decir, al caso donde $f(\bar{x}) < \infty$. Asumamos por simplicidad que $\lambda = 0$ aunque la demostración es análoga para cualquier $\lambda \in \mathbb{R}$ y asumamos también que la solución $(S_t\bar{x})_{t>0}$ está definida para toda condición inicial $\bar{x}$ en el dominio de f, y veamos que podemos extenderlo para toda condición inicial $\bar{x}$ en la clausura de $\mathcal{D}(f)$.

Tomemos una sucesión $(\bar{x}_n) \subset \mathcal{D}(f)$ convergente a nuestra condición inicial $\bar{x} \in \overline{\mathcal{D}(f)}$. Por la propiedad de contractividad del enunciado del teorema, para todo t>0 se tiene que la secuencia $(S_t(\bar{x}_n))$ es de Cauchy, por lo que converge a un elemento $x(t) = S_t\bar{x}$. Por tanto, tenemos que probar que

- $\lim_{t\to 0^+} x(t) = \bar{x}$
- $x(t)$ es un flujo gradiente

Empezamos probando que $\bar{x} = lim_{t\to 0^+} x(t)$. Por la propiedad de contractividad tenemos que para todo t>0 y para n,p $\in \mathbb{R}$ se cumple

$$\|S_t(\bar{x}_{n+p}) - S_t(\bar{x}_n)\| \leqslant \|\bar{x}_{n+p} - \bar{x}_n\|$$

Por tanto, cuando $p \to \infty$ tenemos

$$\|x(t) - S_t(\bar{x}_n)\| \leqslant \|\bar{x} - \bar{x}_n\|$$

Ahora sumando y restando términos y aplicando la desigualdad triangular obtenemos

$$\|\bar{x}(t) - \bar{x}_n\| \leq \|\bar{x}(t) - S_t(\bar{x}_n)\| + \|S_t(\bar{x}_n) - \bar{x}_n\| + \|\bar{x}_n - \bar{x}\| \leq 2\|\bar{x}_n - \bar{x}\| + \|S_t(\bar{x}_n) - \bar{x}_n\|$$

para cualquier n arbitrario. Entonces para un $\epsilon > 0$ elegimos $n \in \mathbb{N}$ tal que $2\|\bar{x}_n - \bar{x}\| \leq \epsilon$, y como $\lim_{t\to 0+} S_t(\bar{x}_n) = \bar{x}_n$, podemos tomar el límite superior en la igualdad anterior y obtener

$$\limsup_{t\to 0} \|x(t) - \bar{x}\| < \epsilon$$

Ahora pasemos a probar que $x(t) = \lim_{n\to\infty} S_t(x_n)$ es un flujo gradiente con la suposición adicional que las curvas $\bar{x}_n(t)$ son localmente equi-Lipschitz en el intervalo $(0, +\infty)$ y, por tanto, x(t) es también localmente equi-Lipschitz. Escribiendo $\bar{x}_n(t) := S_t(\bar{x}_n)$ para simplificar la notación, tenemos que la función $t \to \bar{x}_n(t)$ es una solución EVI_0 para todo $y \in \mathcal{D}(f)$ tenemos

$$\frac{d}{dt}(\frac{1}{2}\|x_n(t) - y\|^2) + f(x_n(t)) \leq f(y)$$

tanto en $\mathcal{L}^\infty$ como en el sentido de las distribuciones. Además, como f es semicontinua inferior y se tiene que

$$\frac{d}{dt}(\frac{1}{2}\|x_n(t) - y\|^2) \to \frac{d}{dt}(\frac{1}{2}\|x(t) - y\|^2)$$

volviendo a la ecuación anterior se obtiene la propiedad EVI de x(t).

Ahora nos falta ver que las curvas $x_n(t)$ son localmente equi-Lipschitz en $(0, +\infty)$. Dado $\epsilon > 0$, como las curvas $x_n(t)$ son flujos gradientes tenemos que

$$\left\|x_n'(t)\right\|^2 = |\nabla f|^2 (x_n(t)) \text{ } para \text{ } casi \text{ } todo \text{ } t > 0.$$

Por tanto, podemos estimar para t>ϵ y v con $\partial f(v) \neq \emptyset$

$$\left\|x_n'(t)\right\|^2 = |\nabla f|^2 (x_n(t)) \leq |\nabla f|^2 (v) + \frac{1}{t^2}\|v - \bar{x}_n\|^2 \leq |\nabla f|^2 (v) + \frac{1}{\epsilon^2}\|v - \bar{x}_n\|^2$$

Lo que prueba que (x_n') está uniformemente acotada en $L^\infty(\epsilon, \infty)$. Ahora, usando la propiedad EVI, para cada $y \in \mathcal{H}$

$$f(x_n(t)) - f(y) \leq -\frac{d}{dt}\left(\frac{1}{2}\|x_n(t) - y\|^2\right) \leq \left\|x_n'(t)\right\| \|x_n(t) - y\| \leq L_\epsilon \|x_n(t) - y\|.$$

Tomando entonces $y = x(s)$ para $s \in (\epsilon, \infty)$ obtenemos

$$|f(x(t)) - f(x(s))| \leq L_\epsilon |x(t) - x(s)| \leq L_\epsilon \int_t^s \left\|x'(r)\right\| dr \leq L_\epsilon^2 |s - t|.$$

Paso 2: Una vez asumido que $f(\bar{x}) < \infty$, para un $\tau > 0$ definimos recurrentemente la secuencia $(y_i)_{n\in\mathbb{N}}$ como $y_0 = \bar{x}$ y para $i \geq 1$ se tiene y_{i+1} como el mínimo de la función

$$g_i(y) = f(y) + \frac{1}{2\tau}\|y - y_i\|^2.$$

Ahora, construimos una solución discreta interpolando los puntos (y_i, y_{i+1})

$$\hat{x}_\tau(t) = \begin{cases} \bar{x} & \text{si t=0} \\ y_i & \text{si } i \geq 1 \text{ y } t \in ((i-1)\tau, i\tau] \end{cases}$$

a la que denotaremos $(S_\tau(\bar{x},t))_{t>0}$.
Observación: Sea $x(t)$ una solución EVI y (t_i) la secuencia de tiempos definidos como $t_i := \tau i$ para $i \in \mathbb{N}$, la derivada

$$\frac{d}{dt}(\frac{1}{2}\|x(t)-v\|^2)$$

puede aproximarse cuando $\tau \to 0$ por el ratio de cambio

$$\frac{1}{\tau}(\frac{1}{2}\|x(t_{i+1})-v\|^2 - \frac{1}{2}\|x(t_i)-v\|^2)$$

Gracias a esta observación podemos definir una versión discreta de las EVI, denotadas como EVI_τ y que además, la sucesión $(y_i)_{i\in\mathbb{N}}$ satisface esa propiedad.

Proposición 2.13. La sucesión $(y_i)_{i\in\mathbb{N}}$ satisface la siguiente propiedad, conocida como EVI_τ :

$$\frac{\|y_{i+1}-v\|^2 - \|y_i-v\|^2}{2\tau} + f(y_{i+1}) \le f(v) \quad \forall v \in \mathcal{H}, \quad \forall i \in \mathbb{N}$$

Demostración. Tomemos un $v \in \mathcal{D}(f)$ y $i \ge 1$. Definimos para cada $t \in (0,1)$

$$\gamma(t) := (1-t)y_{i+1} + tv$$

Por la definición de y_{i+1} tenemos que para todo $t \in (0,1)$ se cumple

$$f(y_{i+1}) + \frac{1}{2\tau}\|y_{i+1}-y_i\|^2 \le f(\gamma(t)) + \frac{1}{2\tau}\|\gamma(t)-y_i\|^2 .$$

Por convexidad de la función f se tiene $f(\gamma(t)) \le (1-t)f(y_{i+1}) + tf(v)$, y usando

$$\frac{1}{2\tau}\|\gamma(t)-y_i\|^2 = \frac{1-t}{2\tau}\|y_{i+1}-y_i\|^2 + \frac{t}{2\tau}\|v-y_i\|^2 - \frac{t(1-t)}{2\tau}\|y_{i+1}-v\|^2$$

podemos obtener

$$\begin{aligned} f(y_{i+1}) + \frac{1}{2\tau}\|y_{i+1}-y_i\|^2 \le & (1-t)f(y_{i+1}) + tf(v) + \frac{1-t}{2\tau}\|y_{i+1}-y_i\|^2 + \\ & + \frac{t}{2\tau}\|v-y_i\|^2 - \frac{t(1-t)}{2\tau}\|y_{i+1}-v\|^2 . \end{aligned}$$

Ahora, restando en ambos lados $(1-t)f(y_{i+1}) + \frac{1-t}{2\tau}\|y_{i+1}-y_i\|^2$ y posteriormente dividiendo entre t>0 obtenemos

$$f(y_{i+1}) \le f(v) + \frac{1}{2\tau}\|v-y_i\|^2 - \frac{1-t}{2\tau}\|y_{i+1}-v\|^2 .$$

Y, por tanto, cuando $t \to 0^+$ se obtiene la propiedad EVI_τ deseada.

□

Paso 3: Por último, veamos lo que ocurre cuando hacemos $\tau \to 0^+$. Para este paso vamos a asumir que estamos trabajando con una función convexa y no negativa.

Proposición 2.14. Sea $f : \mathcal{H} \to [0, \infty]$ una función convexa, y sea $\bar{x} \in \mathcal{D}(f)$. Entonces se cumple:

1. Para todo $t \geq 0$ la familia de curvas $(S_\tau(\bar{x}, t))_{\tau>0}$ es de Cauchy cuando $\tau \to 0^+$, y, por tanto, existe $S(\bar{x}, t) := \lim_{\tau\to 0^+} S_\tau(\bar{x}, t)$
2. La curva $S(\bar{x}, t)$ definida en el apartado anterior es un flujo gradiente que empieza en $\bar{x}$.
3. Podemos obtener una estimación de la diferencia entre las curvas $S_\tau(\bar{x}, t)$ y su límite que viene dada por $\|S_\tau(\bar{x}, t) - S(\bar{x}, t)\| \leq C\sqrt{\tau f(\bar{x})}$ donde $C = 2(\sqrt{2}+1)$, para cada $\tau > 0$ y t>0.

Esta última proposición nos permite ver que la solución discreta que habíamos construido en el paso 2 converge a un flujo gradiente que empieza desde la condición inicial $\bar{x}$, y, por tanto, habríamos probado la existencia de un flujo gradiente tal y como indicaba el teorema de Brézis-Komura.

□

Definición 2.15. (Pendiente descendente). Para $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ y $x \in \mathcal{D}(f)$, la pendiente descendente de f en x, denotada como $|\nabla^- f|(x)$, se define como

$$\left|\nabla^- f\right|(x) = \limsup_{y\to x} \frac{(f(y) - f(x))_-}{|y - x|} \tag{2.5}$$

Notemos que para el caso donde nuestra función f es convexa, para x e y con $x \neq y$ las combinaciones convexas $x_t := (1 - t)x + ty,\ t \in (0, 1)$ cumplen $f(x) - f(x_t) \geq t(f(x) - f(y))$ y, por tanto,

$$\frac{f(x) - f(y)}{|x - y|} \leq \frac{f(x) - f(x_t)}{|x - x_t|}$$

lo que nos permite ver $|\nabla^- f|(x)$ como

$$|\nabla^- f|(x) = \sup_{y\neq x} \frac{(f(y) - f(x))_-}{|y - x|}.$$

Para el caso general de una función λ-convexa se tiene, empleando un argumento similar, que

$$|\nabla^- f|(x) = \sup_{y\neq x} \left[\frac{f(y) - f(x)}{|y - x|} - \frac{\lambda}{2}|x - y|\right]_-$$

Teorema 2.16. *Para cualquier función λ-convexa y cada $x \in \mathcal{D}(f)$ se cumple que $\partial_\lambda f(x) \neq \emptyset$ si y solo si $|\nabla^- f|(x) < \infty$, y en este caso se tiene*

$$|\nabla f(x)| = \left|\nabla^- f\right|(x) \tag{2.6}$$

donde $\nabla f(x)$ denotaba el elemento de norma mínima de $\partial_\lambda f(x)$.

Demostración. Por simplicidad asumamos que $\lambda = 0$ (para el caso general, los coeficientes cuadráticos proporcionales a λ no afectan al cociente incremental en x). Si tenemos $p \in \partial_\lambda f(x)$ entonces se cumple

$$f(x) - f(y) \leq < p, x - y > \leq |p||x - y|$$

y, por tanto, $|\nabla^- f|(x) \leq |p|$. Minimizando respecto a p obtenemos la desigualdad

$$|\nabla^- f|(x) \leq |\nabla f(x)|.$$

Por otro lado, nos falta ver que $\partial_\lambda f(x) \neq \emptyset$ y que $|\nabla^- f|(x) \geq |\nabla f(x)|$, siempre que $L = |\nabla^- f|(x)$ sea finita. Sin pérdida de generalidad podemos asumir que estamos trabajando en $x = 0$ y $f(0) = 0$, por lo que

$$L = \sup_{y \neq 0} \frac{(f(y))_-}{|y|}.$$

Geométricamente, la desigualdad $f(y) \geq -L|y|$ implica que el epigrafo de f, $epi(f)$ está por encima del cono

$$C := \{(y, -L\,|y|) : y \in \mathcal{H}\} \subset \mathcal{H} \times \mathbb{R}.$$

Por otro lado, sabemos que existe un hiperplano que separa $epi(f)$ y C, es decir, existe un $p \in \mathcal{H}$ tal que

$$f(y) \geq \langle p, y \rangle \geq -L|y| \quad \forall y \in \mathcal{H}.$$

La primera desigualdad nos indica que $p \in \partial_0 f(0)$ y la segunda implica que $|p| \leq L$ y, por tanto,

$$|\nabla f(0)| \leq |p| \leq |\nabla^- f|(0).$$

□

CAPÍTULO 3
Redes neuronales

3.1. Introducción sobre las redes neuronales

Las redes neuronales son un modelo computacional inspirado en el funcionamiento del cerebro humano, diseñado para procesar información y aprender patrones complejos. Estas redes están compuestas por nodos interconectados, denominados neuronas, que trabajan en conjunto para realizar tareas específicas, como reconocimiento de patrones o clasificación de datos. El concepto fundamental detrás de las redes neuronales es la capacidad de aprender a partir de datos. Esto se logra ajustando los pesos y las conexiones entre las neuronas durante un proceso de entrenamiento, en el cual el modelo se expone a un conjunto de datos de entrada junto con sus correspondientes clasificaciones. A medida que se procesa esta información, la red neuronal adapta sus parámetros internos para minimizar la diferencia entre las respuestas predichas y las reales.

La construcción de una red neuronal está basado en una máquina presentada en los años sesenta por el investigador F. Rosemblatt llamada Perceptron. Esta máquina realizaba una clasificación binaria de un input (un vector $x \in \mathbb{R}^N$) al que se le realizaba una suma ponderada de cada una de sus coordenadas por unos ciertos pesos, es decir, se realizaba una trasformación afín $T : \mathbb{R}^N \to \mathbb{R}$, dada por $T(x) = \sum_{i=1}^{N} x_i w_i + w_0$, donde w_i representan los pesos de la ponderación, y w_0 refleja el sesgo o término de error. Finalmente, al resultado obtenido por la trasformación afín se le aplicaba una función (generalmente no lineal) conocida como **función de activación** $\sigma : \mathbb{R} \to \mathbb{R}$, que es de la que provendría la clasificación de los datos.

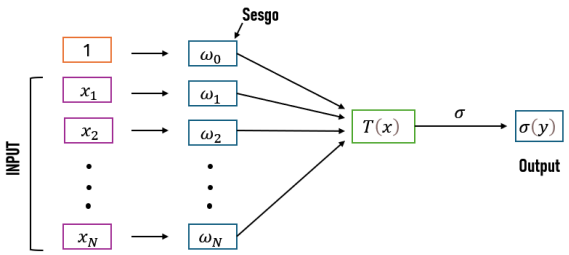

Figura 3.1: Esquema perceptron

Hay que destacar la existencia de diferentes tipos de redes neuronales, pero en nuestro caso nos centraremos en las redes neuronales "feedforward" totalmente conectadas. Estas redes neuronales actuales están conformadas por una concatenación de perceptrones en paralelo, donde cada neurona del perceptrónn en la posición l está conectada a todas las neuronas del perceptrón $l+1$, estos perceptrones se denominan como las capas de la red neuronal.

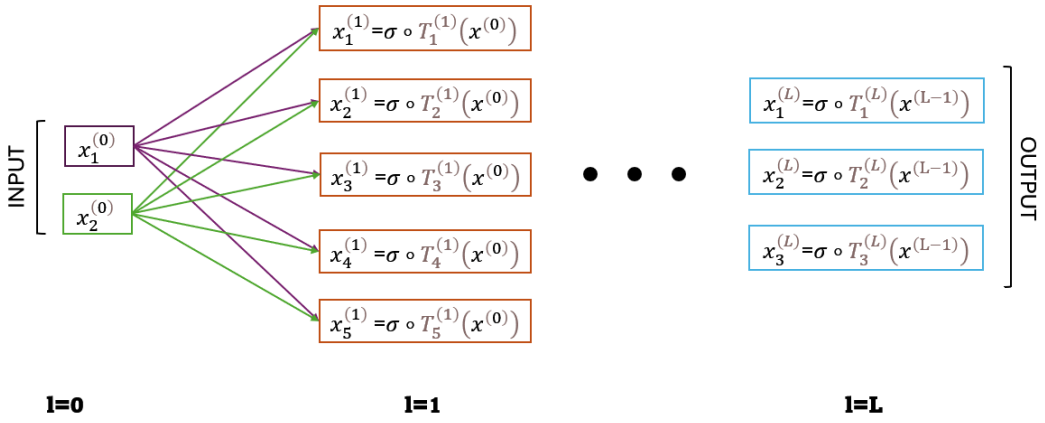

Figura 3.2: Red Neuronal totalmente conectada

Escribiendo esta idea de forma matricial, para las neuronas en la capa $l \in \{1, \dots, L\}$ tenemos

(3.1)
$$\begin{bmatrix} a_1^{(l)} \\ a_2^{(l)} \\ . \\ . \\ a_{n_l}^{(l)} \end{bmatrix} = \sigma \left(\begin{bmatrix} \omega_{11}^{(l)} & . & . & \omega_{n_{l-1}1}^{(l)} \\ \omega_{12}^{(l)} & . & . & \omega_{n_{l-1}2}^{(l)} \\ . & & & . \\ . & & & . \\ . & & & . \\ \omega_{1n_l}^{(l)} & & & \omega_{n_{l-1}n_l}^{(l)} \end{bmatrix} \begin{bmatrix} a_1^{(l-1)} \\ . \\ . \\ a_{n_{l-1}}^{(l-1)} \end{bmatrix} + \begin{bmatrix} w_0^{(l)} \\ . \\ . \\ w_0^{(l)} \end{bmatrix} \right) = \sigma \left(W^{(l)} a^{(l-1)} + b^{(l)} \right)$$

donde $W^{(l)}$ denota la matriz de los pesos de cada una de las capas y $a^{(l-1)}$ denota la activación de las neuronas de la capa anterior, teniendo en cuenta que hay que añadir el sesgo. Por tanto, dando un input $\bar{x} \in \mathbb{R}^N$, el output esperado de la red neuronal se puede expresar como

(3.2)
$$a^{(L)} = \sigma \circ z^{(L-1)} \circ \cdots \circ \sigma \circ z^{(1)}$$

siendo $z^{(l)}$ la pre activación de las neuronas de la capa l, $z^{(l)} = W^{(l)} a^{(l-1)}$. Por otra parte, el interés general sobre las redes neuronales radica en su capacidad de aprender de manera autónoma y adaptativa.

El proceso de aprendizaje de una red neuronal se centra en ajustar la matriz de pesos $W^{(l)}$ de cada capa empleando una muestra de entrenamiento M que consta de unos inputs $\bar{x}_i \in \mathbb{R}^N$ de los que se conoce su clasificación $\bar{y}_i$. El objetivo es minimizar una función de error entre los outputs esperados y recibidos. Esta función de coste puede variar dependiendo de cada red neuronal y cada problema específico, pero en este caso trabajaremos con el promedio de los errores cometidos de la muestra de entrenamiento $M = \{(x_i, y_i)\}_{i=1}^K$ medidos en la norma euclídea. Por tanto, nuestra función de coste dada por la muestra de entrenamiento $M = \{(x_i, y_i)\}_{i=1}^K$ y la arquitectura específica de la red $p = \left\{(W^{(l)}, b^{(l)})\right\}_{l=1}^L \in \mathbb{R}^s$ se define como

$$f_M(p) = \frac{1}{K} \sum_{i=1}^{K} \frac{1}{2} \left\| y^{(i)} - a^{(L)}(x^{(i)}, p) \right\|_2^2 \tag{3.3}$$

siendo $a^{(L)}(x^{(i)}, p)$ el output recibido al haber introducido como input el vector $x^{(i)}$ en la red neuronal cuya arquitectura viene especificada por $p = \left\{(W^{(l)}, b^{(l)})\right\}_{l=1}^L$. También se puede definir la función de error de cada una de las muestras $(x^{(i)}, y^{(i)})$ como

$$f_i(p) = \frac{1}{2} \left\| y^{(i)} - a^{(L)}(x^{(i)}, p) \right\|_2^2$$

y, por tanto, se tiene qué $f_M(p) = \frac{1}{K} \sum_{i=1}^K f_i(p)$.

3.2. Método de descenso del gradiente estocástico

El objetivo está claro, hay que encontrar la arquitectura óptima para nuestra red neuronal basándonos en la muestra de entrenamiento M, es decir, tenemos que encontrar la configuración de pesos $W^{(l)}$ y de sesgos $b^{(l)}$ que hagan que nuestra función de costo se minimice.

La forma clásica de abordar este problema es emplear una iteración para encontrar el mínimo de una función f conocida como el método del gradiente. Este método consiste en que dada una función $f : \mathbb{R}^N \to \mathbb{R}$ diferenciable, tenemos que la dirección de máximo descenso de la función f en un punto $p \in \mathbb{R}^s$ con $\nabla f(p) \neq 0$, es una solución del problema

$$\min_{d \in \mathbb{R}^s, \|d\|=1} \nabla f(x) \cdot d$$

Demostración. Por la desigualdad de Cauchy-Swartz tenemos

$$|\nabla f(p) \cdot d| \leq \|\nabla f(p)\| \, \|d\| = \|\nabla f(p)\|$$

y, por tanto, se cumple

$$-\|\nabla f(p)\| \leq \nabla f(p) \cdot d \leq \|\nabla f(p)\|$$

Ahora tomando $d = -\frac{\nabla f(p)}{\|\nabla f(p)\|}$ tenemos que

$$\nabla f(p) \cdot d = \nabla f(p) \cdot \frac{-\nabla f(p)}{\|\nabla f(p)\|} = -\|\nabla f(p)\|$$

y, por tanto, $d = \frac{-\nabla f(p)}{\|\nabla f(p)\|}$ es la solución a nuestro problema de máximo descenso. □

Por tanto, el método del gradiente consistirá en evaluar primero ∇f en el punto $p \in \mathbb{R}^s$ y luego actualizar nuestro punto por $p \to p - \eta \nabla f(p)$ siendo η un número real conocido como la tasa de aprendizaje. Dependiendo del método empleado, esta tasa de aprendizaje puede depender de cada paso de la iteración, pero en nuestro caso, trabajaremos una tasa fija, lo suficientemente pequeña para garantizar una buena aproximación al mínimo, pero teniendo en cuenta también el coste computacional del algoritmo.

Realmente este método no se utiliza debido a su falta de eficiencia a la hora de trabajar con muestras de gran tamaño o con un gran número de parámetros. Por tanto, una de las estrategias generalmente usadas es utilizar el **método del gradiente estocástico**. Este método consiste en aproximar la media de todos los puntos por una muestra que haya sido seleccionada de forma aleatoria. Esto consistiría en

1. Elegimos de forma aleatoria los enteros $k_1, k_2, ..., k_S \in \{1, ... , n\}$
2. Actualizamos nuestro punto $p \to p - \frac{\eta}{S} \sum_{i=1}^{S} \nabla f_i(p)$

Por tanto, para aplicar el método del descenso del gradiente estocástico necesitamos ser capaces de calcular los distintos gradientes ∇f_i que van a determinar la dirección de descenso máxima de cada f_i. El problema es que pese a ser posible obtener una expresión analítica de los gradientes de forma relativamente sencilla, la evaluación numérica de dicha expresión puede volverse costosa desde un punto de vista computacional. Para enfrentar este problema, recurrimos al **método de retropropagación**.

3.3. Retropropagación

Para aplicar el método del descenso del gradiente estocástico es necesario calcular los gradientes cada una de las funciones de coste. Para ello, intentaremos ver la relación existente entre la función de coste y los diferentes pesos de cada una de las capas de la red neuronal y de los sesgos correspondientes.

Para ello, denotaremos como $z^{(l)}$ al input de la capa l antes de su activación

$$z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)} \in \mathbb{R}^{n_l}$$

y, por tanto, se cumple $a^{(l)} = \sigma(z^{(l-1)})$.

Definición 3.1. Definimos el gradiente local $\delta^{(l)} \in \mathbb{R}^{n_l}$ como el vector cuyas componentes satisfacen

$$\delta_j^{(l)} = \frac{\partial f_i}{\partial z_j^{(l)}} \quad 1 \leq j \leq n_l \ y \ 1 \leq l \leq L$$

Esta expresión se suele llamar el *error* en la j-ésima neurona de la capa l.

Definición 3.2. Se define el *producto de Hadamard* para un par de vectores $u, v \in \mathbb{R}^n$ como $u \odot v := (u_1 v_1, \ \dots u_n v_n)^t \in \mathbb{R}^n$

Proposición 3.3. Con la notación anterior se tiene

$$\delta^{(L)} = \sigma'(z^{(L)}) \odot (a^{(L)} - y) \tag{3.4}$$

$$\delta^{(l)} = \sigma'(z^{(l)}) \odot (W^{(l+1)})^t \ \delta^{(l+1)} \quad para\ 1 \leq l \leq L-1 \tag{3.5}$$

$$\frac{\partial f_i}{\partial b_j^{(l)}} = \delta_j^{(l)} \quad para\ 1 \leq l \leq L \tag{3.6}$$

$$\frac{\partial f_i}{\partial w_{jk}^{(l)}} = \delta_j^{(l)} a_k^{(l-1)} \quad para\ 1 \leq l \leq L \tag{3.7}$$

Demostración. Empecemos probando la ecuación 3.4. Usando la regla de la cadena se puede obtener

$$\delta_j^{(L)} = \frac{\partial f_i}{\partial z_j^{(L)}} = \frac{\partial f_i}{\partial a_j^{(L)}} \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}}, \quad para 1 \leq j \leq n_l$$

Como hemos visto antes que $a^{(L)} = \sigma(z^{(L)}$, tenemos que

$$\frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} = \sigma'(z_j^{(L)})$$

Por otro lado, tenemos que

$$\frac{\partial f_i}{\partial a_j^{(L)}} = \frac{\partial}{\partial a_j^{(L)}} (\frac{1}{2} \left\| y - a^{(L)} \right\|_2^2) = -(y_j - a_j^{(L)})$$

Y, por tanto,

$$\delta_j^{(L)} = (a_j^{(L)} - y_j)\sigma'(z_j^{(L)})$$

que se trata de cada una de las componentes de vector deseado.

Para probar 3.5 nos damos cuenta que podemos expresar $z_k^{(l+1)}$ en función de $z_k^{(l)}$ como

$$z_k^{(l+1)} = \sum_{s=1}^{n_l} w_{ks}^{(l+1)} \sigma(z_s^{(l)}) + b_k^{(l+1)}$$

y, por tanto,

$$\frac{\partial z_k^{(l+1)}}{\partial z_j^{(l)}} = w_{kj}^{(l+1)} \sigma'(z_j^{(l)}).$$

Aplicando los resultados anteriores podemos expresar el gradiente local de cada una de las capas *ocultas* como

$$\delta_j^{(l)} = \frac{\partial f_i}{\partial z_j^{(l)}} = \sum_{k=1}^{n_{l+1}} \frac{\partial f_i}{\partial z_k^{(l+1)}} \frac{\partial z_k^{(l+1)}}{\partial z_j^{(l)}} = \sum_{k=1}^{n_{l+1}} \delta_k^{(l)} w_{kj}^{(l+1)} \sigma'(z_j^{(l)}) = \sigma'(z_j^{(l)})((W^{(l+1)})^t \delta^{(l+1)})_j$$

y al igual que antes, hemos obtenido la expresión de cada una de las coordenadas del vector deseado.

Para probar 3.6 vemos que los $z_j^{(l)}$ están relacionados con los $b_j^{(l)}$ como

$$z_j^{(l)} = (W^{(l)}\sigma(z^{(l-1)}))_j + b_j^{(l)}$$

y como $z^{(l-1)}$ no depende de los $b_j^{(l)}$, obtenemos

$$\frac{\partial f_i}{\partial b_j^{(l)}} = \frac{\partial f_i}{\partial z_j^{(l)}} \frac{\partial z_j^{(l)}}{\partial b_j^{(l)}} = \delta_j^{(l)}.$$

Por último, para probar 3.7 trabajaremos con las componentes del vector $z^{(l)}$. Por definición, tenemos

$$z_j^{(l)} = \sum_{k=1}^{n_{l-1}} w_{jk}^{(l)} a_k^{(l-1)} + b_j^{(l)}$$

y, por tanto, se cumple

$$\frac{\partial z_j^{(l)}}{\partial w_{jk}^{(l)}} = a_k^{(l-1)} \quad y \quad \frac{\partial z_s^{(l)}}{\partial w_{jk}^{(l)}} = 0, \quad para\ s \neq j.$$

Juntando todos estos resultados, obtenemos

$$\frac{\partial f_i}{\partial w_{jk}^{(l)}} = \sum_{s=1}^{n_l} \frac{\partial f}{\partial z_s^{(l)}} \frac{\partial z_s^{(l)}}{\partial w_{jk}^{(l)}} = \frac{\partial f_i}{\partial z_j^{(l)}} \frac{\partial z_j^{(l)}}{\partial w_{jk}^{(l)}} = \delta_j^{(l)} a_k^{(l-1)}$$

□

Observaciones: Recordemos que para calcular $a^{(L)}$ basta emplear las fórmulas presentadas en 3.1 y 3.2 para ir calculando los diferentes $a^{(l)}$ y $z^{(l)}$ de uno en uno. Una vez hecho esto, por 3.4 se puede calcular $\delta^{(L)}$ de forma inmediata. Por tanto, por la ecuación 3.5 podemos ir calculando de forma recursiva los valores $\delta^{(L-1)}$, $\delta^{(L-2)}$, etc. Y por última, empleando 3.6 y 3.7 podemos calcular las derivadas parciales de la función de coste y asi computar los gradientes. Este proceso es el que se conoce como **retropropagación**.

En conclusión, el proceso de aprendizaje de una red neuronal consiste en ir actualización la arquitectura de nuestra red, es decir, ir actualizando las entradas o "pesos"de las matrices $W^{(l)}$ y de los vectores $b^{(l)}$, siguiendo el siguiente algoritmo que se presenta:

1. Se escoge una configuración inicial de la arquitectura de la red
2. Se escoge un punto x aleatorio de nuestra muestra
3. Usamos el método de retropropagación para calcular ∇f_i

4. Actualizamos la arquitectura de nuestra red usando el método del gradiente estocástico

5. Repetimos los pasos 2, 3 y 4.

CAPÍTULO 4
Ecuación de Fokker-Planck

En este capítulo nos centraremos en estudiar sin profundizar como la iteración del método del gradiente estocástico, que se trata de una iteración discreta, puede aproximarse por un proceso estocástico continuo que estará caracterizado por una cierta ecuación diferencial.

La idea de esta parte se centra en observar que la iteración del descenso del gradiente es una discretización de la ecuación diferencial

$$\frac{dx}{dt} = -\nabla f(x)$$

y, por tanto, surge la pregunta de si el método del descenso del gradiente estocástico procede de alguna ecuación diferencial continua.

Una motivación es reescribir la iteración del descenso del gradiente estocástico como

$$x_{k+1} = x_k - \eta \nabla f(x_k) + \sqrt{\eta} V_k(x_k, \gamma_k)$$

donde $V_k(x_k, \gamma_k) = \sqrt{\eta}(\nabla f(x_k) - \nabla f_{\gamma_k}(x_k))$ es una variable aleatoria n-dimensional con

- $\mathbb{E}\left[V_k | x_k\right] = 0$
- $cov\left[V_k, V_k | x_k\right] = \eta \Sigma(x_k)$
- $\Sigma(x_k) := \mathbb{E}\left[(\nabla f_{\gamma_k}(x_k) - \nabla f(x_k)(\nabla f_{\gamma_k}(x_k) - \nabla f(x_k))^t\right]$

Sin entrar en detalles, podemos aproximar esta iteración por el proceso estocástico de Itô continuo X_t descrito por la siguiente ecuación diferencial estocástica

$$dX_t = -\nabla f(X_t)dt + \sqrt{\eta D(X_t)} dW_t, \quad X_0 = x_0$$

donde $D(x) = \frac{1}{n}\sum_{i=1}^{n}(\nabla f(x) - \nabla_i f(x))(\nabla f(x) - \nabla_i f(x))^t$, W representa el movimiento browniano y X_t refleja la evolución de los pesos $w_{jk}^{(l)}$ en el tiempo, es decir, podemos reescribir la ecuación como

$$d\bar{w}_t = -\nabla f(\bar{w}_t)dt + \sqrt{\eta D(\bar{w}_t)} dW_t$$

Esta ecuación diferencial estocástica tiene asociada una ecuación de Fokker-Plank, que refleja la evolución temporal de la función de densidad v($\bar{x}$, t) de los pesos $\bar{w}$ de la red neuronal. La función anterior cumple

$$\partial_t v(t,\bar{w}) = div\left(\nabla f(\bar{w})v(t,\bar{w}) + \frac{1}{2}\eta\ div\left(D(\bar{w})v(t,\bar{w})\right)\right)$$

En nuestro caso, nos centraremos en el modelo más "sencillo"de esta ecuación considerando $D(\bar{w}) = Id$ y $\eta = 2$, de forma que estudiaremos la ecuación

$$\partial_t v = \Delta_{\bar{w}} + div_{\bar{w}}(f(\bar{w})v) = div_{\bar{w}}(\nabla_{\bar{w}} v + f(\bar{w})v) \tag{4.1}$$

que es la ecuación de Fokker-Planck genérica.

4.1. Relación entre la ecuación de Fokker-Planck y la ecuación del calor

El problema clásico de la ecuación del calor n-dimensional viene dado por la ecuación en derivadas parciales

$$u_t - \Delta u = 0, \quad x \in \mathbb{R}^N, t > 0 \tag{4.2}$$

Suponiendo que la cantidad de calor se conserva en el medio, es decir, que se cumple $\int_{\mathbb{R}^n} u(x,t)dx = cte, \quad \forall t > 0$ tenemos que la solución fundamental de la ecuación del calor es la Gaussiana

$$w(t,x) = \frac{C}{(\sqrt{t})^n} \cdot e^{\frac{-|x|^2}{4t}}$$

y ajustando la constante C para que la cantidad de calor sea 1 tenemos qué $C = \left(\frac{1}{4\pi}\right)^{\frac{N}{2}}$.

La relación entre la ecuación del calor descrita en 4.2 y la ecuación de Fokker-Plank genérica descrita en 4.1 viene dada por un cambio de variables de la ecuación del calor.

Proposición 4.1. Sea $u(\tau, y)$ una solución de la ecuación del calor, haciendo el cambio de variables

$$u(\tau,y) = (1-2\tau)^{-\frac{N}{2}} v(t,x), \quad t = \frac{1}{2}\log(1+2\tau), \quad x = \left(\sqrt{1+2\tau}\right)^{-1} y \tag{4.3}$$

obtenemos la ecuación de Fokker-Plank genérica 4.1.

Demostración. Por ser $u(\tau, y)$ una solución de la ecuación del calor tenemos que se cumple $u_\tau - \Delta_y u = 0$. Aplicando la regla de la cadena, es fácil ver que se cumple

$$\begin{aligned} u_\tau &= v_\tau(t,x)\cdot(1+2\tau)^{-\frac{N}{2}} + v(t,x)\cdot\frac{\partial}{\partial\tau}\left[(1+2\tau)^{-\frac{N}{2}}\right] \\ &= (1+2\tau)^{-\frac{N}{2}-1}\left(-x\nabla_x v(t,x) + v_t(t,x) - Nv(t,x)\right), \end{aligned}$$

$$u_{y_i} = (1+2\tau)^{-\frac{N}{2}} v_{y_i} = (1+2\tau)^{-\frac{N}{2}} \frac{\partial v}{\partial x_i}\frac{\partial x_i}{\partial y_i} = (1+2\tau)^{-\frac{N}{2}} \frac{v_{x_i}}{\sqrt{1+2\tau}},$$

$$u_{y_i y_i} = \frac{(1+2\tau)^{-\frac{N}{2}}}{\sqrt{1+2\tau}} \frac{\partial v_{x_i}}{\partial x_i}\frac{\partial x_i}{\partial y_i} = (1+2\tau)^{-\frac{N}{2}-1} v_{x_i x_i}$$

y, por tanto, tenemos que

$$\Delta_y u = \sum_{i=1}^{N} \frac{\partial^2 u}{\partial y^2} = (1+2\tau)^{-\frac{N}{2}-1} \sum_{i=1}^{N} v_{x_i x_i} = (1+2\tau)^{-\frac{N}{2}-1} \Delta_x v.$$

Ahora igualando según la ecuación del calor $u_\tau - \Delta_y u = 0$

$$(1+2\tau)^{-\frac{N}{2}-1} \left(-x\nabla_x v + v_t - Nv\right) = (1+2\tau)^{-\frac{N}{2}-1} \Delta_x v.$$

Cancelando términos, obtenemos

$$v_t = \Delta_x v + x\nabla_x v + Nv$$

y utilizando que N puede expresarse como $N = div(x)$, obtenemos

$$v_t = \Delta_x v + x\nabla_x v + div(x)v = \Delta_x v + div(xv) = div(\nabla_x v + xv) \tag{4.4}$$

que se trata de la ecuación de Fokker-Planck descrita en 4.1 para el caso más sencillo donde $f(x) = x$.

□

Este último resultado nos ofrece la oportunidad de profundizar en el análisis de la ecuación de Fokker-Planck y de encontrar alguna solución estacionaria y su comportamiento asintótico.

Lema 4.2. *(Soluciones estacionarias de la ecuación de Fokker-Plank). Las soluciones estacionarias de la ecuación de Fokker-Plank para el caso donde $f(w) = w$ vienen dadas por las funciones*

$$G(w) = M\frac{e^{-\frac{\|w\|^2}{2}}}{(2\pi)^{\frac{N}{2}}}$$

cuyo valor de la integral es M.

Demostración. Queremos comprobar si $G(w)$ satisface que

$$0 = \frac{dG}{dt} = div(\nabla_w G + wG)$$

La función G cumple $\nabla_w G = -M\frac{e^{-\frac{\|w\|^2}{2}}}{(2\pi)^{\frac{N}{2}}} w$ y, por tanto, se tiene $\nabla_w G + Gw = 0$. □

4.2. Interpretación de la ecuación de Fokker-Planck como flujo gradiente

Esta sección del capítulo se enfocará en la adaptación de la ecuación de Fokker-Planck para que pueda ser interpretada como un flujo gradiente de un funcional de energía específico, comúnmente conocido como la entropía del sistema. Para lograr esto, el primer paso consistirá en considerar que en la ecuación descrita en 4.1, la función f puede expresarse como el gradiente de una función convexa y positiva, la cual llamaremos potencial y que coincidirá con nuestra función de coste descrita en 3.3. Es decir, podremos expresar f como $f = \nabla V$ para un potencial Vconcreto.

Siguiendo con lo anterior e interpretando la función f como el gradiente de un potencial V, podemos reescribir la ecuación 4.1 como

$$\frac{du}{dt} = div(\nabla u + u\nabla V) \tag{4.5}$$

Proposición 4.3. Sea ahora una función $V : \mathbb{R}^n \to \mathbb{R}$ localmente Lipschitz con $\int_{\mathbb{R}^n} e^{-V} dx < \infty$. Consideramos la ecuación de Fokker-Plank descrita en 4.5, y haciendo el cambio de variables $w = ue^V$ podemos transformar la ecuación de Fokker-Planck como

$$e^{-V}\frac{dw}{dt} = div(e^{-V}\nabla w) \tag{4.6}$$

Demostración. Aplicando la regla de la cadena es fácil comprobar que

$$\frac{dw}{dt} = \frac{d}{dt}(ue^V) = \frac{du}{dt}e^V + \frac{d}{dt}(e^V)u = \frac{du}{dt}e^V = e^V div(\nabla u + u\nabla V)$$

Por otro lado, tenemos que $\frac{\partial}{\partial x_i}(ue^V) = u_{x_i}e^V + uV_{x_i}e^V$ y, por tanto,

$$\begin{aligned} div(e^{-V}\nabla w) &= \sum_i \frac{\partial}{\partial x_i}\left(e^{-V}(u_{x_i}e^V + uV_{x_i}e^V)\right) = \sum_i \frac{\partial}{\partial x_i}\left(u_{xi} + uV_{x_i}\right) \\ &= \sum_i \left(u_{x_ix_i} + u_{x_i}V_{x_i} + uV_{x_ix_i}\right) = \Delta u + \nabla u\nabla V + u\Delta V \\ &= div(\nabla u + u\nabla V) \end{aligned}$$

Igualando ambos términos obtenemos

$$\frac{dw}{dt} = e^V div(e^{-V}\nabla w)$$

tal y como queríamos. □

Esto sugiere que la interpretemos como el flujo gradiente sobre el espacio de Hilbert $H := L^2(\mathbb{R}^n, \gamma = e^{-V}\mu)$ del funcional $D_\gamma : H \to \mathbb{R}_+ \cup \{\infty\}$ definido como

$$D_\gamma(w) = \begin{cases} \frac{1}{2}\int_{\mathbb{R}^n} |\nabla w|^2 \, d\gamma(w) & si\ \nabla w \in L^2(\mathbb{R}^n, \gamma) \\ +\infty \text{ en otro caso} \end{cases} \tag{4.7}$$

Hay que destacar que, pese a que estamos trabajando sobre $\mathcal{H} := L^2(\mathbb{R}^n, \gamma = e^{-V}\mu)$ donde a priori las funciones no son derivables, podemos trabajar con el gradiente ∇w en el sentido de las distribuciones, ya que $L^2(\gamma) \subset L^1_{Loc}(\mathbb{R}^n)$.

Nuestro funcional D_γ es convexo y semicontinuo inferior y, por tanto, podemos aplicar la teoría de flujos gradientes desarrollada en el capítulo dos, en concreto, aplicando el teorema de Brezis-Komura presentado en 2.12 nos da la familia de operadores $S_t\bar{x}$ definida en todo H, que corresponde al flujo dado por

$$e^{-V}\frac{dw}{dt} = div(e^{-V}\nabla w) \tag{4.8}$$

interpretando el lado derecho en el sentido de las distribuciones, y el lado izquierdo como la derivada del operador $w : \mathbb{R}_+ \to H,\ t \to w(t,\cdot)$. Para justificar esto, necesitamos comprobar varias cosas.

La primera de ellas es comprobar que verdaderamente la solución de la ecuación de Fokker-Plank reescrita como en 4.6 es un flujo gradiente del funcional $D_\gamma(w)$ que habíamos definido en 4.7. Para ello, vamos a aplicar uno de los resultados más fuertes que habíamos presentado en el segundo tema, que se trata del teorema de caracterización de flujos gradientes presentado en 2.9. Este teorema nos dice que para comprobar que nuestra solución es un flujo gradiente tenemos que comprobar si se cumple la propiedad 2.4.

Sea $\bar{v} \in \mathcal{H} = L^2_\gamma$ y sea $v(t)$ la solución de 4.6, entoces

$$\begin{aligned}
\frac{d}{dt}\left(\frac{1}{2}\left\|v(t)-\bar{v}\right\|^2_{L^2_\gamma}\right) &= \frac{d}{dt}\left(\frac{1}{2}\int_{\mathbb{R}^n}|v(t)-\bar{v}|^2\,d\gamma\right)\\
&= \int_{\mathbb{R}^n}(v(t)-\bar{v})\,v_t d\gamma\\
&= \int_{\mathbb{R}^n}(v-\bar{v})e^V div(e^{-V}\nabla v)d\gamma\\
&= \int_{\mathbb{R}^n} ve^V div(e^{-V}\nabla v)e^{-V}d\mu - \int_{\mathbb{R}^n}\bar{v}e^V div(e^{-V}\nabla v)e^{-V}d\mu\\
&= \int_{\mathbb{R}^n}|\nabla v|^2\,e^{-V}d\mu + \int_{\mathbb{R}^n}\nabla\bar{v}\nabla v e^{-V}d\mu\\
&\leq -2D_\gamma(v) + \left\|\nabla\bar{v}\right\|_{L^2_\gamma}\left\|\nabla v\right\|_{L^2_\gamma}\\
&\leq -2D_\gamma(v) + \frac{1}{2}\left(\left\|\nabla\bar{v}\right\|^2_{L^2_\gamma} + \left\|\nabla v\right\|^2_{L^2_\gamma}\right)\\
&= -D_\gamma(v) + D_\gamma(\bar{v})
\end{aligned}$$

La segunda es ver que $\partial D_\gamma(w) \neq \emptyset$. Sin entrar en detalles, se puede ver que se cumple

$$\xi \in \partial D_\gamma(w) \ \Leftrightarrow\ \int_{\mathbb{R}^n}\xi\phi d\gamma(x) = \int_{\mathbb{R}^n}\langle\nabla w, \nabla\phi\rangle d\gamma(x) \quad \forall\phi \in C^\infty_0(\mathbb{R}^n) \tag{4.9}$$

donde $C^\infty_0(\mathbb{R}^n)$ representa nuestro conjunto de funciones test que son las funciones C^∞ de soporte compacto. Este último resultado implica que $e^{-V}\xi$ es igual a $div(e^{-V}\nabla w)$

en el sentido de las distribuciones, que es el problema que estábamos estudiando de partida, lo que significa que el camino de máximo decrecimiento de la función de coste y por el cual deberíamos actualizar los pesos de nuestra red neuronal, es el camino dado por la solución de la ecuación de Fokker-Plank, que es equivalente al camino que minimiza el funcional D_γ dado por el flujo gradiente del teorema de Brézis-Komura.

CAPÍTULO 5
Conclusiones

Como se ha señalado en la introducción de este trabajo, la teoría de flujos gradientes emerge como una herramienta fundamental para investigar sistemas dinámicos que vienen determinados por ciertas ecuaciones en derivadas parciales específicas. El objetivo principal de este estudio consiste en la expansión y comprensión de la teoría de flujos gradientes, para luego contextualizarla en una aplicación concreta y relevante, en nuestro caso, la minimización de una función de coste que desempeña un papel esencial en el proceso de aprendizaje de las redes neuronales.

En el primer capítulo de este trabajo se presentan una serie de resultados sobre la teoría de las funciones convexas en los espacios $\mathbb{R}^n$ y sobre la minimización de las mismas, para luego poder generalizarlos en espacios de Hilbert más generales, en concreto, a espacios de funciones como pueden ser $L^2(\mathbb{R}^n)$ o $W^{1,2}(\mathbb{R}^n)$.

En el segundo capítulo, nos sumergimos en la teoría de flujos gradientes, comenzando por la definición del concepto de flujo gradiente y solución EVI para un espacio de Hilbert general. Posteriormente, exploramos uno de los resultados más significativos de este capítulo: la equivalencia entre los flujos gradientes y las soluciones EVI. Esta equivalencia simplifica considerablemente la tarea de verificar si una curva $x(t)$ es un flujo gradiente, ya que solo se requiere comprobar si se cumple la desigualdad mencionada. Después, presentamos y demostramos el resultado fundamental en la teoría de flujos gradientes, conocido como el teorema de Brezis-Komura. Este teorema establece la existencia y unicidad de flujos gradientes para funciones λ-convexas generales. Sin embargo, es importante señalar que en este trabajo nos hemos centrado principalmente en el caso donde $\lambda = 0$ por la simplicidad de los cálculos y porque en las aplicaciones posteriores se asume que los funcionales con los que trabajamos son convexos. Finalmente, examinamos la pendiente descendiente para una función general, que representa la dirección de máximo decrecimiento en cada punto y está determinada por el elemento de mínima norma del flujo gradiente de la función asociada a ese punto. Este es un resultado bastante importante y útil, ya que nos permite identificar las direcciones de decrecimiento de nuestro funcional, las cuales pueden no ser evidentes a priori.

En el tercer capítulo nos enfocamos en la composición y el funcionamiento de una red neuronal, y como podemos expresar los procesos internos como el aprendizaje de

la red con matemáticas de forma rigurosa. Esto nos ha permitido ver que el método de aprendizaje de una red neuronal se centra en el proceso de minimización de una función de coste que viene dada por el error cometido entre los outputs esperados y recibidos. Para minimizar esta función, empleamos una variación del método del descenso del gradiente conocida como el método del descenso del gradiente estocástico. Este método, que a priori no es el más óptimo en términos de la dirección de decrecimiento, nos permite reducir drásticamente la cantidad de operaciones realizadas por nuestro ordenador para hallar una buena aproximación del mínimo.

Sin entrar en detalles, se hace una mención sobre un resultado sobre procesos estocásticos que nos dice que podemos aproximar el método del descenso del gradiente estocástico por un proceso continuo que tiene asociado una ecuación diferencial de tipo Fokker-Planck. Este último resultado nos da pie a estudiar la ecuación de Fokker-Plank y relacionarla con la teoría de flujos gradientes presentada en la primera parte del trabajo. Esto nos permite concluir que las trayectorias dadas por la ecuación de Fokker-Plank son las trayectorias de mayor decrecimiento de un funcional de energía asociado, es decir, son las trayectorias de la arquitectura de la red neuronal que más aceleran su aprendizaje.

Bibliografía

[1] KOT, Mark.: A first course in the calculus of variations. American Mathematical Society, 2014.

[2] BRÈZIS, Haïm.: FUNCIONAL, Análisis. Teoría y aplicaciones. Alianza Editorial, 1984.

[3] Ambrosio, Luigi: Lecture Notes of the Optimal Transport 2016-17 course, 2017.

[4] Bonforte, M and Quirós, F: Notes of the Master Course on PDEs at UAM Sobolev spaces and applications to elliptic PDEs, 2023.

[5] Nesterov, Yurii: Lectures on Convex Optimization, 2018.

[6] Bazaraa, Mokhtar S and Sherali, Hanif D and Shetty, Chitharanjan M: Nonlinear programming: theory and algorithms. John wiley & sons, 2013.

[7] Li, Qianxiao and Tai, Cheng and Weinan, E: Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. Journal of Machine Learning Research, 2019.

[8] Bazaraa, Mokhtar S and Sherali, Hanif D and Shetty, Chitharanjan M: Nonlinear programming: theory and algorithms. John wiley & sons, 2013.

[9] Goodfellow, Ian and Bengio, Yoshua and Courville, Aaron: Deep learning. MIT press, 2016.

[10] Nielsen, Michael A: Neural networks and deep learning. Determination press San Francisco, CA, USA, 2015.

[11] Ibarrondo, P. M: Deep Approximation and Stochastic Gradient Descent. Departamento de Matemáticas, UAM, 2020