A deep mathematical understanding of DNNs

Jiang J.
Data Engineer, Data Scientist
ENSEEIHT Computer Science Engineering Degree
INP Toulouse Dual MSc Research Degree in AI, Big Data and Ops
France

Abstract

Frameworks such as TensorFlow or PyTorch make deep learning developments easy. They have made this field wide spread for every enthusiast. Implementations only needs an instinctive understanding of deep learning. The proper math aspect is little by little forgotten.

Matrix, Normalized vector space, Limit plus continuity, Finite dimensional linear algebra and Linear application matrix theories are supposed known. The objective is to do a collection of the important propositions explaining dense neural network (DNN) theories. These propositions will be mathematically proven. The subject used as reference is a multi-class classification problem with – dense layers, $ReLU$ and $SoftMax$ activation layers, Categorical cross-entropy loss and Stochastic gradient descent optimizer. But all the elements below can be easily re-used or re-defined to cover regressions.

1 – Prerequisites and Notations

1.1 – Matrices

Convention – A vector is a matrix with only one row. Thus, the real vector set $\Re^m$ is equivalent to $M_{1,m}$.

Notation – Let $a_{i,j} \in \Re$ for $i \in [\![1, n]\!]$ and $j \in [\![1, m]\!]$. Then a real matrix of dimension $n \times m$ will be noted as

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} \end{bmatrix}$$

The following notations are also considered

$$\forall i \in [\![1, n]\!] \ , \ \forall j \in [\![1, m]\!] \ , \ A_{i,j} = a_{i,j}$$

$$\forall j \in [\![1,m]\!] \quad , \quad A_{:,j} = \begin{bmatrix} a_{1,j} \\ a_{2,j} \\ \vdots \\ a_{n,j} \end{bmatrix}$$

$$\forall i \in [\![1,n]\!] \quad , \quad A_{i,:} = \begin{bmatrix} a_{i,1} & a_{i,2} & \cdots & a_{i,n} \end{bmatrix}$$

The notation $M_{n,m}$ means the matrix set of dimension $n \times m$ with coefficients in $\Re$.

The notation $M_{n,m}(E)$ means the matrix set of dimension $n \times m$ with coefficients in $E \subseteq \Re$.

Notation – Let $A \in M_{n,m}$, and $B \in M_{m,p}$, and let the product noted $A \times B$ or AB be

$$C = A \times B = AB$$

where C is in $M_{n,p}$ with

$$\forall i \in [\![1,n]\!] \quad , \quad \forall j \in [\![1,p]\!] \quad , \quad C_{i,j} = \sum_{k=1}^{m} A_{i,k} \times B_{k,j}$$

Notation – Let $a \in \Re$ and $B \in M_{n,m}$. Let the scalar wise product noted as $a \times B$ be

$$C = a \times B = B \times a$$

where C is in $M_{n,m}$ with

$$\forall i \in [\![1,n]\!] \quad , \quad \forall j \in [\![1,m]\!] \quad , \quad C_{i,j} = a \times B_{i,j}$$

Notation – The matrix transpose operation will be noted as A^T .

Notation – Let $a \in \Re^n$ and $b \in \Re^n$. Let the scalar product on $\Re^n$ between two vectors noted as ${}_n(a|b)$ be

$${}_n(a|b) = a \times b^T = b \times a^T = {}_n(b|a)$$

Let $c \in \Re^n$. Let the Euclidean norm on $\Re^n$ noted as ${}_n\|c\|$ be

$${}_n\|c\| = \sqrt{(c|c)} = \sqrt{c \times c^T}$$

1.2 – Functions

Notation – Let $E \subseteq \Re^n$ and $F \subseteq \Re^m$.

The notation $f : E \to F$ means the application from E to F .

The notation $\zeta(E, F)$ means the set of continuous applications from E to F .

The notation $\mathscr{L}(E, F)$ means the set of linear applications from E to F .

Definition – Let $E \subseteq \Re^n$ and $F \subseteq \Re^m$. Let $f : E \to F$. Then f differentiable on E is equivalent to

$$\forall\, a \in \Omega \quad , \quad \exists \frac{\partial f}{\partial \cdot}(a) : \begin{cases} \Re^n \to \Re^m \\ x \mapsto \dfrac{\partial f}{\partial x}(a) \end{cases} \in \mathscr{L}(\Re^n, \Re^m) \quad , \quad f(a+h) = f(a) + \frac{\partial f}{\partial x}(a) + \underset{h \to 0}{o}\left({}_n\|h\|\right)$$

The notation $D(E, F)$ means the set of differentiable applications from E to F .

Notation – Let $E \subseteq \Re^n$, $F \subseteq \Re^m$ and $f = (f_1 \ \cdots \ f_m) \in D(E, F)$. Then its Jacobian matrix is noted as the application

$$J_f : \begin{cases} \Omega \to M_{m', m} \\ x \mapsto \begin{bmatrix} \dfrac{\partial f_1}{\partial e_1}(x) & \cdots & \dfrac{\partial f_1}{\partial e_n}(x) \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_m}{\partial e_1}(x) & \cdots & \dfrac{\partial f_m}{\partial e_n}(x) \end{bmatrix} \end{cases}$$

$(e_i)_{i \in [\![1, n]\!]}$ means the matrices $e_i = [0 \ \cdots \ 0 \ \underset{\text{at index } i}{1} \ 0 \ \cdots \ 0] \in \Re^n$ corresponding to the $\Re^n$ standard basis.

<HERE>

Definition – Let $\Omega \neq \emptyset$ subset of $\Re^m$, ${}_m\|\cdot\|$ a norm on $\Re^m$, ${}_{m'}\|\cdot\|$ a norm on $\Re^{m'}$, and $f : \Omega \to \Re^{m'}$. Then f continuous function is equivalent to

$$\forall a \in \Omega \quad ,$$

$$\forall \epsilon > 0 \quad , \quad \exists \eta > 0 \quad , \quad \forall x \in \Omega \quad , \quad {}_m\|x-a\| < \eta \Rightarrow {}_{m'}\|f(x) - f(a)\| < \epsilon$$

The notation $\zeta(\Omega, \Re^{m'})$ means the set of continuous functions from Ω to $\Re^{m'}$.

The notation $\zeta(\Re^m)$ means the set of continuous functions from $\Re^m$ to $\Re^m$.

Notation – Let $\Omega \neq \emptyset$ subset of $\Re^m$ **.** The notation $\mathscr{L}(\Omega, \Re^{m'})$ means the set of linear applications from Ω to $\Re^{m'}$.

Definition – Let $\Omega \neq \emptyset$ subset of $\Re^m$, ${}_m\|\cdot\|$ a norm on $\Re^m$, ${}_{m'}\|\cdot\|$ a norm on $\Re^{m'}$, and $f : \Omega \rightarrow \Re^{m'}$. Then f differentiable is equivalent to

$$\forall a \in \Omega \quad , \quad \exists \frac{\partial f}{\partial \cdot}(a) : \begin{cases} \Omega \rightarrow \Re \\ x \mapsto \frac{\partial f}{\partial x}(a) \end{cases} \in \mathscr{L}(\Omega, \Re^{m'}) \quad , \quad \forall x \in \Omega \quad ,$$

$$\forall \epsilon > 0 \quad , \quad \exists \eta > 0 \quad , \quad \forall \tau \in \Re \quad , \quad |\tau| < \eta \Rightarrow {}_{m'}\left\|\frac{f(\tau \times x + a) - f(a)}{\tau} - \frac{\partial f}{\partial x}(a)\right\| < \epsilon$$

In the notation $\frac{\partial f}{\partial x}(a)$, x and a mean the direction and point of differentiation respectively.

The notation $\frac{\partial f}{\partial x}(a)$ can also be wrote as $f'(a)$ when $m=1$.

The notation $D(\Omega, \Re^{m'})$ means the set of differentiable functions from Ω to $\Re^{m'}$.

The notation $D(\Re^m)$ means the set of differentiable functions from $\Re^m$ to $\Re^m$.

Definition – Let $\Omega \neq \emptyset$ subset of $\Re^m$ and $f \in D(\Omega, \Re^{m'})$. Then its Jacobian matrix J_f is defined as the application

$$J_f : \begin{cases} \Omega \rightarrow M_{m',m} \\ x \mapsto \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \cdots & \frac{\partial f_1}{\partial x_m}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_{m'}}{\partial x_1}(x) & \cdots & \frac{\partial f_{m'}}{\partial x_m}(x) \end{bmatrix} \end{cases}$$

The notation f_i means the application $f_i:\begin{cases} \Omega \to \Re \\ x \mapsto f(x)_i \end{cases}$ and x_i means the matrices $x_i = [0 \ \cdots \ 0 \ \underset{at\,index\,i}{1} \ 0 \ \cdots \ 0] \in \Re^m$.

The Jacobian matrix is also named Gradient matrix when $m'=1$.

Notation – Let $f:\Omega \to \Omega'$ and $g:\Omega' \to \Omega''$. Then the notation $g \circ f$ means the application $g \circ f:\begin{cases} \Omega \to \Omega'' \\ x \mapsto g(f(x)) \end{cases}$.

Let $(f_i)_{i \in [\![1,n]\!]}$ with $f_i:\Omega_i \to \Omega_{i+1}$ for $i \in [\![1,n]\!]$. Then the notation $\overset{n}{\underset{i=1}{\circ}} f_i$ means the application $\overset{n}{\underset{i=1}{\circ}} f_i:\begin{cases} \Omega_1 \to \Omega_{n+1} \\ x \mapsto f_n(f_{n-1}(\ldots f_2(f_1(x)))) \end{cases}$.

Proposition – Let $U \neq \emptyset$ and $V \neq \emptyset$ subsets of $\Re^n$ and $\Re^m$. Let $f:\begin{cases} U \to V \\ x \mapsto f(x) \end{cases}$ and $g:\begin{cases} V \to \Re^p \\ y \mapsto g(y) \end{cases}$ such as $f \in \zeta(U,V)$ and $g \in \zeta(U,V)$. Then $g \circ f:\begin{cases} U \to \Re^p \\ x \mapsto g(f(x)) \end{cases}$ is in $\zeta(U,\Re^p)$.

Proof: TO DO.

Theorem – Let $U \neq \emptyset$ and $V \neq \emptyset$ subsets of $\Re^n$ and $\Re^m$. Let $f:\begin{cases} U \to V \\ x \mapsto f(x) \end{cases}$ and $g:\begin{cases} V \to \Re^p \\ y \mapsto g(y) \end{cases}$ such as $f \in \zeta(U,V) \cap D(U,V)$ and $g \in \zeta(U,V) \cap D(V,\Re^p)$. Then $g \circ f:\begin{cases} U \to \Re^p \\ x \mapsto g(f(x)) \end{cases}$ is in $D(U,\Re^p)$ and its Jacobian is

$$J_{g \circ f}:\begin{cases} U \to M_{p,n} \\ x \mapsto J_g(f(x)) \times J_f(x) \end{cases}$$

Proof: TO DO.

1.3 – Function convexity and smoothness

Definition – Let $\Omega \neq \emptyset$ convex of $\Re^m$ and $f : x \mapsto f(x)$ such as $f \in \zeta(\Omega, \Re)$. Then f convex on Ω is equivalent to

$$\forall (y,z) \in \Omega^2 \quad , \quad \forall t \in [0,1] \quad ,$$
$$f(t \times y + (1-t) \times z) \leq t \times f(y) + (1-t) \times f(z)$$

Proposition – Let $\Omega \neq \emptyset$ convex of $\Re^m$ and $f \in D(\Omega, \Re)$ convex on Ω . Then f convex on Ω is equivalent to

$$\forall (y,z) \in \Omega^2 \quad ,$$
$$f(y) + \frac{df}{dx}(y) \times (z-y)^T \leq f(z)$$

<u>Proof:</u> TO DO.

Proposition – Let $\Omega \neq \emptyset$ convex of $\Re^m$ and $f : x \mapsto f(x)$ such as $f \in D(\Omega, \Re)$ and convex. Then

$\exists X^* \subset \Omega_{\setminus \{\emptyset\}} \quad , \quad \forall x^* \in X^* \quad , \quad f : x \mapsto f(x^*) \leq f(x)$ <TODO existing and that's all>

Definition – Let $\Omega \neq \emptyset$ subset of $\Re^m$, ${}_m\|\cdot\|$ a norm on $\Re^m$, and $f : x \mapsto f(x)$ such as $f \in D(\Omega, \Re)$. Let $L > 0$. Then f $L-smooth$ on Ω is equivalent to

$$\forall (y,z) \in \Omega^2 \quad , \quad {}_m\|\frac{df}{dx}(y) - \frac{df}{dx}(z)\| \leq L \times {}_m\|y-z\|$$

Proposition – Let $\Omega \neq \emptyset$ convex of $\Re^m$ and $f : x \mapsto f(x)$ such as $f \in D(\Omega, \Re)$. Then f first order integral form Taylor expansion is

$$\forall (y,z) \in \Omega^2 \quad ,$$
$$f(y) = f(z) + \int_0^1 \frac{df}{dx}(z + \tau(y-z))(y-z)^T d\tau$$

Proof: TO DO.

Proposition – Let $\Omega \neq \emptyset$ convex of $\Re^m$, $L>0$, and $f: x \mapsto f(x)$ such as $f \in D(\Omega, \Re)$ plus $L-smooth$ on Ω . Then

$$\forall (y,z) \in \Omega^2 \ ,$$

$$f(z) \leq f(y) + \frac{df}{dx}(y) \times (z-y)^T + \frac{L}{2} \times {}_m\|z-y\|$$

$$\forall x \in \Omega \ ,$$

$$f(x - \frac{1}{L} \times \frac{df}{dx}(x)) - f(x) \leq -\frac{1}{2L} \times {}_m\|\frac{df}{dx}(x)\|^2$$

Proof: TO DO.

Proposition – Let $\Omega \neq \emptyset$ convex of $\Re^m$, ${}_m\|\cdot\|$ a norm on $\Re^m$, and $f: x \mapsto f(x)$ such as $f \in D(\Omega, \Re)$, convex and $L-smooth$. Then f is co-coercive

$$\forall (y,z) \in \Omega^2 \ ,$$

$$\frac{1}{L} \times {}_m\|\frac{df}{dx}(y) - \frac{df}{dx}(z)\|^2 \leq (\frac{df}{dx}(y) - \frac{df}{dx}(z)) \times (x-y)^T$$

Proof: TO DO.

1.3 – Others

Notation – Let $\Omega \in \Re^m$. The notation 1_Ω means the Ω indicator function on $\Re^m$.

$$1_\Omega : \begin{cases} \Omega \rightarrow \Re^m \\ x \mapsto \begin{cases} 1 & x \in \Omega \\ 0 & x \notin \Omega \end{cases} \end{cases}$$

Notation – Let $f:\begin{cases}\Omega_1\times\ldots\times\Omega_n\rightarrow\Omega'_1\times\ldots\times\Omega'_m\\(x_1,\ldots,x_n)\mapsto f(x_1,\ldots,x_n)\end{cases}$ an application with n parameters and m outputs. Then for $k\in[\![1,n]\!]$ the notation $f(x_1,\ldots,x_{k-1},\cdot,x_{k+1},\ldots,x_n)$ means the application $f(x_1,\ldots,x_{k-1},\cdot,x_{k+1},\ldots,x_n):\begin{cases}\Omega_k\rightarrow\Omega'_1\times\ldots\times\Omega'_m\\x_k\mapsto f(x_1,\ldots,x_{k-1},x_k,x_{k+1},\ldots,x_n)\end{cases}$.

2 – Activation functions

Definition – Let $F_{act}\in D(\Re^m)$. Then the vector wise application

$$F_{act}:\begin{cases}\Re^m\rightarrow\Re^m\\z\mapsto f(z)\end{cases}$$

is an activation function.

Definition – $ReLU$ is the following vector wise application

$$ReLU:\begin{cases}\Re^m\rightarrow\Re^m\\z\mapsto max(0_{\Re^m},z)\end{cases}$$

with max the element-wise maximum operation between two vectors.

Hypothesis – The notation $ReLU_j$ means the application corresponding to the coefficient j of the function $ReLU$. Let $z\in\Re^m$ then

$$\forall\, j\in[\![1,m]\!]\quad,\quad ReLU_j(z_j)=max(0,z_j)=ReLU(z)_j$$

$ReLU$ is supposed derivable on every coefficients at 0

$$\forall\, j\in[\![1,m]\!]\quad,\quad ReLU_j'(0)=0$$

Proposition – $ReLU$ is an activation function. Its Jacobian matrix is

$$\frac{d\,ReLU}{d\,z}:\begin{cases}\Re^m \to M_{m,m}\\ z \mapsto \begin{bmatrix} 1_{\Re^+_{\setminus\{0\}}}(z_1) & 0 & \cdots & 0\\ 0 & 1_{\Re^+_{\setminus\{0\}}}(z_2) & \ddots & \vdots\\ \vdots & \ddots & \ddots & 0\\ 0 & \cdots & 0 & 1_{\Re^+_{\setminus\{0\}}}(z_m)\end{bmatrix}\end{cases}$$

<u>Proof:</u> TO DO.

Proposition – The following vector wise application is an activation function

$$SoftMax:\begin{cases}\Re^m \to]0,1[^m\\ z \mapsto e^{z_j} \Big/ \sum_{j'=1}^{m} e^{z_{j'}}\end{cases}$$

with e the element-wise exponential operation.

The *SoftMax* function will be denoted as S for simplicity.

Its Jacobian matrix is

$$\frac{d\,S}{d\,z}:\begin{cases}\Re^m \to M_{m,m}\\ z \mapsto \frac{d\,S}{d\,z}(z)\end{cases}$$

where $\forall z \in \Re^m$, $\forall (j, j') \in \{1, 2, \ldots, m\}^2$,

$$\frac{d\,S}{d\,z}(z)_{j,j'} = S(z)_j \times (\delta_{j,j'} - S(z)_{j'})$$

with $\delta_{j,j'}$ the Kronecker delta.

<u>Proof:</u> TO DO.

3 – Loss

Definition – Let $\hat{\Omega} \in M_{n,m}$ and $\Omega \subseteq M_{n,m}$ non empty subsets. Let $\hat{y} \in \hat{\Omega}$ and $F^{\hat{y}}_{loss} \in D(\Omega, \Re)$

. Then $F_{loss}^{\hat{y}}$ is a loss function is equivalent to the application

$$F_{loss}^{\hat{y}} \circ g : \begin{cases} E \rightarrow \Re \\ \epsilon \mapsto (F_{loss}^{\hat{y}} \circ g)(\epsilon) = F_{loss}^{\hat{y}}(\hat{y} + \epsilon) \end{cases}$$

is an increasing function according each coefficient with $E \subseteq M_{n,m}$ such as $F_{loss}^{\hat{y}} \circ g$ is always defined.

The $\hat{y}$ matrix is named the ground truth.

Proposition – Let $\hat{y} \in \{0,1\}^m$ a ground truth matrix. Then the application

$$\xi^{\hat{y}} : \begin{cases}]0,1[^m \rightarrow \Re \\ y \mapsto -\sum_{j=1}^{m} \hat{y}_j \log(y_j) \end{cases}$$

is a loss function. The application is named Categorical cross-entropy loss.

Its Gradient matrix is

$$\frac{d\xi^{\hat{y}}}{dy} : \begin{cases}]0,1[^m \rightarrow \Re^m \\ y \mapsto -\begin{bmatrix} \frac{\hat{y}_1}{y_1} & \dots & \frac{\hat{y}_m}{y_m} \end{bmatrix} \end{cases}$$

<u>Proof:</u> TO DO.

Proposition – Let $\hat{y} \in \{0,1\}^m$ a ground truth matrix. Let $S : \Re^m \rightarrow]0,1[^m$ and $\xi^{\hat{y}} :]0,1[^m \rightarrow \Re$ the *SoftMax* activation and Categorical cross-entropy loss functions. Then $\xi^{\hat{y}} \circ S : \Re^m \rightarrow \Re$ is derivable on $\Re^m$ and its Gradient matrix is

$$\frac{d(\xi^{\hat{y}} \circ S)}{dz} : \begin{cases} \Re^m \rightarrow \Re^m \\ z \mapsto S(z) - \hat{y} \end{cases}$$

<u>Proof:</u> TO DO.

4 – Dense layers

Definition – The application

$$L: \begin{cases} \Re^m \times M_{m',m} \times \Re^{m'} \to \Re^{m'} \\ (y, W, b) \mapsto y \times W^T + b \end{cases}$$

defines a dense layer with y named the input vector, W named the weight matrix and b named the bias matrix.

The notation L_j means the application $L_j : \Re^m \times \Re^m \times \Re \to \Re$ corresponding to the row j of the second matrix component of the dense layer L . Let $y \in \Re^m$ an input vector, $W \in M_{m',m}$ a weight matrix and $b \in \Re^{m'}$ a bias matrix then

$$\forall j \in [\![1, m']\!] \quad , \quad L_j(y, W_{j,:}, b_j) = y \times (W_{j,:})^T + b_j = L(y, W, b)_j$$

Proposition – Let $L : \Re^m \times M_{m',m} \times \Re^{m'} \to \Re^{m'}$ a dense layer function. Then L is derivable according the first and third variables on $\Re^m$ and $\Re^{m'}$ respectively.

Let $y \in \Re^m$ an input vector and $b \in \Re^{m'}$ a bias matrix. Then $L_{j'} : \Re^m \times \Re^m \times \Re \to \Re$ is also derivable according the second variable for all $j \in [\![1, m']\!]$.

Its Gradient or Jacobian matrices are

$$\frac{\partial L}{\partial y} : \begin{cases} \Re^m \to M_{m',m} \\ y \mapsto W \end{cases}$$

$$\forall j \in [\![1, m']\!] \quad , \quad \frac{\partial L_j}{\partial w} : \begin{cases} \Re^m \to \Re^m \\ w \mapsto y \end{cases}$$

$$\frac{\partial L}{\partial b} : \begin{cases} \Re^{m'} \to M_{m',m'} \\ b \mapsto I_{m'} \end{cases}$$

with $I_{m'}$ the identity matrix of size $m' \times m'$.

<u>Proof:</u> TO DO.

Proposition – Let $L : \Re^m \times M_{m',m} \times \Re^{m'} \to \Re^{m'}$ and $ReLU : \Re^{m'} \to \Re^{m'}$ the dense layer and $ReLU$ activation functions. Let $F^{upstream} : \begin{cases} \Re^{m'} \to \Re \\ y' \mapsto F^{upstream}(y') \end{cases}$ such as $F^{upstream} \in D(\Re^{m'}, \Re)$.

Then $F_{upstream} \circ ReLU \circ L(\cdot, W, b): \Re^m \times M_{m',m} \times \Re^{m'} \to \Re$ is derivable according the first and third variables on $\Re^m$ and $\Re^{m'}$ respectively.

The notation $F_{j'}^{upstream}$ means the application corresponding to the coefficient j' of $F^{upstream}$.
Let $y' \in \Re^{m'}$ then

$$\forall j' \in [\![1, m']\!] \ , \ F_{j'}^{upstream}(y'_j) = F^{upstream}(y')_{j'}$$

Let $y \in \Re^m$ an input vector and $b \in \Re^{m'}$ a bias matrix. Then
$F_{j'}^{upstream} \circ ReLU_{j'} \circ L_{j'}(\cdot, w, b_{j'}): \Re \to \Re$ is also derivable for all $j' \in [\![1, m']\!]$.

Its Gradient matrices are

$$\frac{\partial (F^{upstream} \circ ReLU \circ L(\cdot, W, b))}{\partial y} : \begin{cases} \Re^m \to \Re^m \\ y \mapsto \dfrac{\partial (F^{upstream} \circ ReLU \circ L(\cdot, W, b))}{\partial y}(y) \end{cases}$$

where $\forall y \in \Re^m$, $\forall j \in [\![1, m]\!]$,

$$\frac{\partial (F^{upstream} \circ ReLU \circ L(\cdot, W, b))}{\partial y}(y)_j = \sum_{j'=1}^{m'} \frac{d F^{upstream}}{d y'}(ReLU(L(y, W, b)))_{j'} \times 1_{\Re^+_{\setminus\{0\}}}(L(y, W, b)_{j'}) \times W_{j',j}$$

with $W \in M_{m',m}$ a weight matrix, $b \in \Re^{m'}$ a bias matrix.

$$\forall j' \in [\![1, m']\!] \ ,$$
$$\frac{\partial (F_{j'}^{upstream} \circ ReLU_{j'} \circ L_{j'}(\cdot, w, b_{j'}))}{\partial w} : \begin{cases} \Re^m \to \Re^m \\ w \mapsto \dfrac{\partial (F_{j'}^{upstream} \circ ReLU_{j'} \circ L_{j'}(\cdot, w, b_{j'}))}{\partial w}(w) \end{cases}$$

where $\forall w \in \Re^m$, $\forall j \in [\![1, m]\!]$,

$$\frac{d (F_{j'}^{upstream} \circ ReLU_{j'} \circ L_{j'}(\cdot, w, b_{j'}))}{d w}(w)_j = F_{j'}^{upstream}{}'(ReLU_{j'}(L_{j'}(y, w, b_{j'}))) \times 1_{\Re^+_{\setminus\{0\}}}(L_{j'}(y, w, b_{j'})) \times y_j$$

with $y \in \Re^m$ an input matrix, $b \in \Re^{m'}$ a bias matrix.

$$\frac{\partial\left(F^{upstream}\circ ReLU\circ L(\cdot,W,b)\right)}{\partial b}:\begin{cases}\Re^{m'}\rightarrow\Re^{m'}\\ b\mapsto\dfrac{\partial\left(F^{upstream}\circ ReLU\circ L(\cdot,W,b)\right)}{\partial b}(b)\end{cases}$$

where $\forall b\in\Re^{m'}$, $\forall j'\in[\![1,m']\!]$,

$$\frac{\partial\left(F^{upstream}\circ ReLU\circ L(\cdot,W,b)\right)}{\partial b}(b)_{j'}=\frac{d\,F^{upstream}}{d\,y'}\left(ReLU\left(L(y,W,b)\right)\right)_{j'}\times 1_{\Re^{+}_{\setminus\{0\}}}\left(L(y,W,b)_{j'}\right)$$

with $y\in\Re^m$ a weight matrix and $W\in M_{m',m}$ a weight matrix.

<u>Proof:</u> TO DO.

5 – Neural Networks

Definition – A training data set is defined as couples of vectors $(X^i,\hat{Y}^i)\in\Re^m\times\Re^l$ for $i\in[\![1,n]\!]$. The X^i are named input or feature matrices and the $\hat{Y}^i$ target or label matrices.

Definition – Let p dense layers with activation functions $F^k_{act}\circ L^k(\cdot,W^k,b^k):\Re^{m_k}\rightarrow\Re^{m_{k+1}}$ for $k\in[\![1,p]\!]$ with $W^k\in M_{m_{k+1},m_k}$ and $b^k\in\Re^{m_{k+1}}$ the L^k weight and bias matrices. Let a training data set $(X^i,\hat{Y}^i)\in\Re^{m_1}\times\Re^{m_{p+1}}$ for $i\in[\![1,n]\!]$. Let $F^{\hat{Y}^i}_{loss}:\Re^{m_{p+1}}\rightarrow\Re$ loss functions for $i\in[\![1,n]\!]$ with $(\hat{Y}^i)_{i\in[\![1,n]\!]}$ as ground truth matrices respectively.

Then a neural network is defined as the application $N:\begin{cases}\Re^{m_1}\rightarrow\Re^{m_{p+1}}\\ y\mapsto\overset{n}{\underset{k=1}{\circ}}\left(F^k_{act}\circ L^k(\cdot,W^k,b^k)\right)(y)\end{cases}$.

The optimization problem is $\underset{(W^k_{1,:},\dots,W^k_{m_{k+1},:},b^k)_{k\in[\![1,p]\!]}}{min}\sum_{i=1}^{n}F^{\hat{Y}^i}_{loss}\left(N(X^i)\right)$ and $F^{global}_{loss}:\left((W^k_{1,:},\dots,W^k_{m_{k+1},:},b^k)_{k\in[\![1,p]\!]}\right)\mapsto\sum_{i=1}^{n}F^{\hat{Y}^i}_{loss}\left(N(X^i)\right)$ is named the objective function or global loss.

Theorem – Let p dense layers with activation functions – $ReLU^k\circ L^k(\cdot,W^k,b^k):\Re^{m_k}\rightarrow\Re^{m_{k+1}}$

for $k\in[\![1,p-1]\!]$ and $S\circ L^p(\cdot,W^p,b^p):\Re^{m_p}\to\Re^{m_{p+1}}$. $W^k\in M_{m_{k+1},m_k}$ and $b^k\in\Re^{m_{k+1}}$ are defined as the L^k weight and bias matrices. Let a training data set $(X^i,\hat{Y}^i)\in\Re^{m_1}\times\Re^{m_{p+1}}$ for $i\in[\![1,n]\!]$. Let $\xi^{\hat{Y}^i}:\Re^{m_{p+1}}\to\Re$ Categorical cross-entropy losses for $i\in[\![1,n]\!]$ with $(\hat{Y}^i)_{i\in[\![1,n]\!]}$ as ground truth matrices respectively.

Then the following application $N:\Re^{m_1}\to\Re^{m_{p+1}}$ with

$$N=S\circ L^p(\cdot,W^p,b^p)\circ\mathop{\circ}_{k=1}^{p-1}(ReLU^k\circ L^k(\cdot,W^k,b^k))$$

is a neural network and its objective function is

$$\xi_{loss}^{global}:((W_{1,:}^k,\ldots,W_{m_{k+1},:}^k,b^k)_{k\in[\![1,p]\!]})\mapsto\sum_{i=1}^{n}\xi_{loss}^{\hat{Y}^i}(N(X^i))$$

Let $k\in[\![1,p]\!]$. For all $i\in[\![1,n]\!]$, let

$$y^{downstream(k),X^i}=\begin{cases}\mathop{\circ}_{l=1}^{k-1}(ReLU^l\circ L^l(\cdot,W^l,b^l))(X^i) & k\geq 2\\ X^i & k=1\end{cases}$$

$$F^{upstream(k),\hat{Y}^i}:\begin{cases}\Re^{m_{k+1}}\to\Re\\ y\mapsto\begin{cases}\xi^{\hat{Y}^i}\circ S(y) & k=p\\ \xi^{\hat{Y}^i}\circ S\circ L^p(\cdot,W^p,b^p)(y) & k=p-1\\ \xi^{\hat{Y}^i}\circ S\circ L^p(\cdot,W^p,b^p)\circ\mathop{\circ}_{l=k+1}^{p-1}(ReLU^l\circ L^l(\cdot,W^l,b^l))(y) & k\leq p-2\end{cases}\end{cases}$$

then $$\frac{dF^{upstream(k),\hat{Y}^i}}{dy}:\begin{cases}\Re^{m_{k+1}}\to\Re\\ y\mapsto\begin{cases}S(y)-\hat{Y}^i & k=p\\ (S(y)-\hat{Y}^i)\times W^p & k=p-1\\ \dfrac{\partial(F^{upstream(k+1),\hat{Y}^i}\circ ReLU^{k+1}\circ L^{k+1}(\cdot,W,b))}{\partial y}(y) & k\leq p-2\end{cases}\end{cases}$$ where

$\forall\, k\in[\![1,p-2]\!]$, $\forall\, y\in\Re^{m_{k+1}}$, $\forall\, j\in[\![1,m]\!]$,

$$\frac{\partial(F^{upstream(k+1),\hat{Y}^i}\circ ReLU^{k+1}\circ L^{k+1}(\cdot,W^{k+1},b^{k+1}))}{\partial y}(y)_j$$

$$=\sum_{j'=1}^{m'}\frac{dF^{upstream(k+1),\hat{Y}^i}}{dy'}\left(ReLU^{k+1}\left(L^{k+1}\left(y,W^{k+1},b^{k+1}\right)\right)\right)_{j'}\times 1_{\Re^+_{\setminus\{0\}}}\left(L^{k+1}\left(y,W^{k+1},b^{k+1}\right)_{j'}\right)\times W^{k+1}_{j',j}$$

with $W^{k+1}\in M_{m',m}$ a weight matrix, $b^{k+1}\in\Re^{m'}$ a bias matrix.

Let $k=p$. Then ξ^{global}_{loss} Gradient matrices are

$$\forall j'\in[\![1,m_{k+1}]\!] \quad , \quad \frac{\partial\xi^{global}_{loss}}{\partial W^k_{j',:}}:\begin{cases}\Re^{m_k}\rightarrow\Re^{m_k}\\ w\mapsto\sum_{i=1}^{n}\frac{\partial\left(F^{upstream(k),\hat{Y}^i}_{j'}\circ L^k_{j'}\left(\cdot,w,b^k_{j'}\right)\right)}{\partial w}(w)\end{cases}$$

where $\forall i\in[\![1,n]\!]$, $\forall w\in\Re^{m_k}$, $\forall j\in[\![1,m_k]\!]$,

$$\frac{\partial\left(F^{upstream(k),\hat{Y}^i}_{j'}\circ L^k_{j'}\left(\cdot,w,b^k_{j'}\right)\right)}{\partial w}(w)_j$$

$$=F^{upstream(k),\hat{Y}^i}_{j'}{}'\left(L^k_{j'}\left(y^{downstream(k),X^i},w,b^k_{j'}\right)\right)\times y^{downstream(k),X^i}_j$$ with $b^k\in\Re^{m_{k+1}}$ a bias matrix.

$$\frac{\partial\xi^{global}_{loss}}{\partial b^k}:\begin{cases}\Re^{m_{k+1}}\rightarrow\Re^{m_{k+1}}\\ b^k\mapsto\sum_{i=1}^{n}\frac{\partial\left(F^{upstream(k),\hat{Y}^i}\circ L^k\left(\cdot,W^k,b^k\right)\right)}{\partial b^k}\left(b^k\right)\end{cases}$$

where $\forall i\in[\![1,n]\!]$, $\forall b^k\in\Re^{m_k}$, $\forall j'\in[\![1,m_{k+1}]\!]$,

$$\frac{\partial\left(F^{upstream(k),\hat{Y}^i}\circ L^k\left(\cdot,W^k,b^k\right)\right)}{\partial b^k}\left(b^k\right)_{j'}$$

$$=\frac{dF^{upstream(k),\hat{Y}^i}}{dy}\left(L^k\left(y^{downstream(k),X^i},W^k,b^k\right)\right)_{j'}\times 1_{\Re^+_{\setminus\{0\}}}\left(L^k\left(y^{downstream(k),X^i},W^k,b^k\right)_{j'}\right)$$

with $W^k\in M_{m_{k+1},m_k}$ a weight matrix.

Let $k\in[\![1,p-1]\!]$. Then ξ^{global}_{loss} Gradient matrices are

$$\forall j'\in[\![1,m_{k+1}]\!] \quad , \quad \frac{\partial\xi^{global}_{loss}}{\partial W^k_{j',:}}:\begin{cases}\Re^{m_k}\rightarrow\Re^{m_k}\\ w\mapsto\sum_{i=1}^{n}\frac{\partial\left(F^{upstream(k),\hat{Y}^i}_{j'}\circ ReLU^k_{j'}\circ L^k_{j'}\left(\cdot,w,b^k_{j'}\right)\right)}{\partial w}(w)\end{cases}$$

where $\forall i \in [\![1,n]\!]$, $\forall w \in \Re^{m_k}$, $\forall j \in [\![1,m_k]\!]$,

$$\frac{\partial\left(F_{j'}^{upstream(k),\hat{Y}^i} \circ ReLU_{j'}^k \circ L_{j'}^k(\cdot, w, b_{j'}^k)\right)}{\partial w}(w)_j$$

$$= F_{j'}^{upstream(k),\hat{Y}^i}{}'\left(ReLU_{j'}^k\left(L_{j'}^k\left(y^{downstream(k),X^i}, w, b_{j'}^k\right)\right)\right) \times 1_{\Re^+_{\setminus\{0\}}}\left(L_{j'}^k\left(y^{downstream(k),X^i}, w, b_{j'}^k\right)\right) \times y_j^{downstream(k),X^i}$$

with $b^k \in \Re^{m_{k+1}}$ a bias matrix.

$$\frac{\partial \xi_{loss}^{global}}{\partial b^k} : \begin{cases} \Re^{m_{k+1}} \rightarrow \Re^{m_{k+1}} \\ b^k \mapsto \sum_{i=1}^{n} \frac{\partial\left(F^{upstream(k),\hat{Y}^i} \circ ReLU^k \circ L^k(\cdot, W^k, b^k)\right)}{\partial b^k}(b^k) \end{cases}$$

where $\forall i \in [\![1,n]\!]$, $\forall b^k \in \Re^{m_k}$, $\forall j' \in [\![1,m_{k+1}]\!]$,

$$\frac{\partial\left(F^{upstream(k),\hat{Y}^i} \circ ReLU^k \circ L^k(\cdot, W^k, b^k)\right)}{\partial b^k}(b^k)_{j'}$$

$$= \frac{d F^{upstream(k),\hat{Y}^i}}{d y}\left(ReLU^k\left(L^k\left(y^{downstream(k),X^i}, W^k, b^k\right)\right)\right)_{j'} \times 1_{\Re^+_{\setminus\{0\}}}\left(L^k\left(y^{downstream(k),X^i}, W^k, b^k\right)_{j'}\right)$$

with $W^k \in M_{m_{k+1}, m_k}$ a weight matrix.

<u>Proof:</u> TO DO.

6 – Optimizations

Definition – Let $f \in D(\Omega, \Re)$ with $\Omega \in \Re^m$. <TODO>

7 – References