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.

Topology, Normalized vector space, Limit plus continuity, Taylor series expansion, Matrix, 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 – Fundamentals and Notations

1.1 – Matrices

Convention – All sets considered are non empty.

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

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

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

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

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

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

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

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

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 – Differential calculus

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

The notation $\overset{o}{E}$ means the set with only the interior point of E .
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

$$\forall a \in \overset{o}{E} \ , \ \exists \frac{\partial f}{\partial \cdot}(a) : \begin{cases} \Re^n \to \Re^m \\ h \mapsto \frac{\partial f}{\partial h}(a) \end{cases} \in \mathscr{L}(\Re^n, \Re^m) \ ,$$

$$\forall h \in \Re^n \ , \ f(a+h) = f(a) + \frac{\partial f}{\partial h}(a) + \underset{h \to 0}{o}\left({}_n\|h\|\right)$$

$\frac{\partial f}{\partial \cdot}(a)$ is named differential on a of f .

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

Proposition – Let $E \subseteq \Re^n$, $F \subseteq \Re^m$ and $f \in D(E,F)$. Then the differential $\frac{\partial f}{\partial \cdot}(a)$ is unique and $D(E,F) \subset \zeta(E,F)$.

<u>Proof:</u> Suppose ϕ_1 and ϕ_2 two differential of f on a . Then

$$\forall h \in \Re^n \ , \ \phi_2(h) - \phi_1(h) = \underset{h \to 0}{o}\left(2 \times {}_n\|h\|\right)$$

$$\text{so} \ \ \forall \epsilon > 0 \ , \ \exists \eta > 0 \ , \ \forall h \in \Re^n \ , \ {}_n\|h\| \leq \eta \Rightarrow {}_m\|\phi_2(h) - \phi_1(h)\| \leq 2 \times {}_n\|h\| \times \epsilon$$

$$\phi_2 - \phi_1 \in \mathscr{L}(\Re^n, \Re^m)$$

$$\text{so} \ \ \forall \epsilon > 0 \ , \ \forall h \in \Re^n \ , \ {}_m\|\phi_2(h) - \phi_1(h)\| < 2 \times {}_n\|h\| \times \epsilon$$

Then Squeeze theorem with $\epsilon \to 0$ *gives the differential uniqueness*

$$\forall h \in \Re^n \quad , \quad \phi_2(h) = \phi_1(h)$$

Let $a \in \overset{o}{E}$ *. Then*

$$\frac{\partial f}{\partial \cdot}(a) \in \mathscr{L}(\Re^n, \Re^m)$$

$$so \quad \frac{\partial f}{\partial \cdot}(a) \in \xi(\Re^n, \Re^m) \quad and \quad \frac{\partial f}{\partial 0_{\Re^n}}(a) = 0_{\Re^m}$$

Then f *differentiable definition gives the continuity*

$$f(a+h) = f(a) + \frac{\partial f}{\partial h}(a) + \underset{h \to 0}{o}(\|h\|)_n \underset{h \to 0}{\to} f(a)$$

Definition & Proposition – Let $E \subseteq \Re^n$, $F \subseteq \Re^m$ and $f = (f_1 \ \cdots \ f_m) \in D(E,F)$. Then f_i is differentiable on E for all $i \in [\![1,m]\!]$ and its Jacobian matrix is noted as the application

$$J_f : \begin{cases} \overset{o}{E} \to M_{m,n} \\ a \mapsto \left[\frac{\partial f}{\partial e_1}(a) \ \cdots \ \frac{\partial f}{\partial e_n}(a) \right] = \begin{bmatrix} \frac{\partial f_1}{\partial e_1}(a) & \cdots & \frac{\partial f_1}{\partial e_n}(a) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial e_1}(a) & \cdots & \frac{\partial f_m}{\partial e_n}(a) \end{bmatrix} \end{cases}$$

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

The Jacobian matrix of f on a fixed is also the canonical matrix associated to the differential of f on a .

$\frac{\partial f}{\partial e_i}(a)$ is named the partial derivative according the i^{th} variable.

The Jacobian is also named Gradient when $m = 1$ and will be noted as $\nabla_f = J_f$.

<u>Proof:</u> Let $f = (f_1 \ \cdots \ f_m) \in D(E,F)$ *. Then*

$$\forall i\in[\![1,m]\!]\ ,\ \forall h\in\Re^n\ ,$$

$$\begin{aligned} f_i(a+h) &= f(a+h)_i \\ &= f(a)_i+\frac{\partial f}{\partial h}(a)_i+\underset{h\to 0}{o}(_n\|h\|) \\ &= f_i(a)+\frac{\partial f_i}{\partial h}(a)+\underset{h\to 0}{o}(_n\|h\|) \end{aligned}$$

with $\forall i\in[\![1,m]\!]\ ,\ \frac{\partial f_i}{\partial .}(a)=\frac{\partial f}{\partial .}(a)_i\in\mathscr{L}(\Re^n,\Re)$

Notation – Let $E\subseteq\Re^n$, $F\subseteq\Re^m$, $G\subseteq\Re^p$, $f:E\to F$ and $g:F\to G$. Then the notation $g\circ f$ means the application $g\circ f:\begin{cases}E\to F\\ x\mapsto g(f(x))\end{cases}$.

Let $(f_i)_{i\in[\![1,n]\!]}$ with $f_i:E_i\to E_{i+1}$ for $i\in[\![1,n]\!]$. Then the notation $\overset{n}{\underset{i=1}{\circ}}f_i$ means the application $\overset{n}{\underset{i=1}{\circ}}f_i:\begin{cases}E_1\to E_{n+1}\\ x\mapsto f_n(f_{n-1}(\dots f_2(f_1(x))))\end{cases}$.

Theorem – Let $E\subseteq\Re^n$, $F\subseteq\Re^m$, $G\subseteq\Re^p$, $f\in D(E,F)$ and $g\in D(F,G)$. Then $g\circ f\in D(E,G)$ and its Jacobian is

$$J_{g\circ f}:\begin{cases}\overset{o}{E}\to M_{p,n}\\ a\mapsto J_g(f(a))\times J_f(a)\end{cases}$$

<u>Proof:</u> Let $E\subseteq\Re^n$, $F\subseteq\Re^m$, $G\subseteq\Re^p$, $f\in D(E,F)$ *and* $g\in D(F,G)$. *Let* $a\in\overset{o}{E}$. *Then*

$$\forall h\in\Re^n\ ,$$

$$\begin{aligned} g\circ f(a+h) &= g(f(a)+\frac{\partial f}{\partial h}(a)+\underset{h\to 0}{o}(_n\|h\|)) \\ &= g(f(a))+\frac{\partial g}{\partial(\frac{\partial f}{\partial h}(a)+\underset{h\to 0}{o}(_n\|h\|))}(f(a))+\underset{h\to 0}{o}(\ \|\frac{\partial f}{\partial h}(a)+\underset{h\to 0}{o}(_n\|h\|)\|) \end{aligned}$$

Then $\frac{\partial g}{\partial .}(a)\in\mathscr{L}(\Re^m,\Re^p)$, $\frac{\partial g}{\partial .}(a)\in\zeta(\Re^m,\Re^p)$, $\frac{\partial g}{\partial 0_{\Re^m}}(a)=0_{\Re^p}$, $\frac{\partial f}{\partial .}(a)\in\zeta(\Re^n,\Re^m)$ *and* $\frac{\partial f}{\partial 0_{\Re^n}}(a)=0_{\Re^m}$ *gives*

$$\forall h \in \Re^n \ ,$$

$$g \circ f(a+h) = g(f(a)) + \frac{\partial g}{\partial (\frac{\partial f}{\partial h}(a))}(f(a)) + \underset{h \to 0}{o}(\|h\|_n)$$

It means $g \circ f \in D(E,G)$ *and* $\frac{\partial g \circ f}{\partial \cdot}(a) = \frac{\partial g}{\partial \cdot}(f(a)) \circ \frac{\partial f}{\partial \cdot}(a)$.

It also gives with canonical associated matrices

$$\forall h \in \Re^n \ , \ J_{g \circ f}(a) \times h = J_g(f(a)) \times (J_f(a) \times h) = (J_g(f(a)) \times J_f(a)) \times h$$

$$so \quad J_{g \circ f}(a) = J_g(f(a)) \times J_f(a)$$

1.3 – Function convexity and smoothness

Definition – Let $E \subseteq \Re^n$ convex, $F \subseteq \Re$ and $f \in \xi(E,F)$. Then f convex is equivalent to

$$\forall (x,y) \in E^2 \ , \ \forall t \in [0,1] \ ,$$

$$f(t \times x + (1-t) \times y) \leq t \times f(x) + (1-t) \times f(y)$$

Proposition – Let $E \subseteq \Re^n$ convex, $F \subseteq \Re$ and $f \in D(E,F)$ convex. Then f convex is equivalent to

$$\forall (x,y) \in E^2 \ ,$$

$$f(x) + \nabla_f(x) \times (y-x)^T \leq f(y)$$

Proof: 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

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

Definition – Let $\Omega \neq \emptyset$ subset of $\Re^m$, $\|\cdot\|_m$ 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

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

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

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

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

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

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

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

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

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

Proof: TO DO.

1.3 – Others

Notation – Let $E \subseteq \Re^n$. The notation 1_E means the E indicator function on $\Re^n$.

$$1_E : \begin{cases} E \to \Re^n \\ x \mapsto \begin{cases} 1 & x \in E \\ 0 & x \notin E \end{cases} \end{cases}$$

Notation – Let $f : \begin{cases} E_1 \times \ldots \times E_n \to F_1 \times \ldots \times F_m \\ (x_1, \ldots, x_n) \mapsto f(x_1, \ldots, x_n) \end{cases}$ 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 $f(x_1, \ldots, x_{k-1}, \cdot, x_{k+1}, \ldots, x_n) : \begin{cases} \Omega_k \to \Omega'_1 \times \ldots \times \Omega'_m \\ x_k \mapsto f(x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_n) \end{cases}$.

2 – Activation functions

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

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

is an activation function.

Definition – $ReLU$ is the following vector wise application

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

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

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

$ReLU$ is supposed derivable on every coefficients at 0

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

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

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

Proof: 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.

Proof: 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 $F_{loss}^{\hat{y}} \in D(\Omega, \Re)$. Then $F_{loss}^{\hat{y}}$ is a loss function is equivalent to the application

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

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

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

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

Its Gradient matrix is

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

Proof: TO DO.

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

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

Proof: 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

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

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 $L_{j\cdot}:\Re^m\times\Re^m\times\Re\to\Re$ is also derivable according the second variable for all $j\in[\![1,m']\!]$.

Its Gradient or Jacobian matrices are

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

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

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

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

Proof: 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 $F^{upstream}_{j'}$ means the application corresponding to the coefficient j' of $F^{upstream}$.
Let $y'\in\Re^{m'}$ then

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

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

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

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

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

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

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

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

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

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.

Proof: 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 $F^k_{act} \circ L^k(\cdot, W^k, b^k) : \Re^{m_k} \to \Re^{m_{k+1}}$ 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 $F^{\hat{Y}^i}_{loss} : \Re^{m_{p+1}} \to \Re$ 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 $N : \begin{cases} \Re^{m_1} \to \Re^{m_{p+1}} \\ y \mapsto \overset{n}{\underset{k=1}{\circ}}\left(F^k_{act} \circ L^k(\cdot, W^k, b^k)\right)(y) \end{cases}$.

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

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

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

$$N = S \circ L^p(\cdot, W^p, b^p) \circ \overset{p-1}{\underset{k=1}{\circ}} \left(ReLU^k \circ L^k(\cdot, W^k, b^k) \right)$$

is a neural network and its objective function is

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\forall j' \in [\![1, m_{k+1}]\!] \ , \ \frac{\partial \xi_{loss}^{global}}{\partial W_{j',:}^{k}} : \begin{cases} \Re^{m_k} \to \Re^{m_k} \\ w \mapsto \sum_{i=1}^{n} \frac{\partial \left(F_{j'}^{upstream(k), \hat{Y}^i} \circ ReLU_{j'}^{k} \circ L_{j'}^{k}(\cdot, w, b_{j'}^{k}) \right)}{\partial w}(w) \end{cases}$$

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

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

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

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

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

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.

Proof: TO DO.

6 – Optimizations

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

7 – References