MathDNN - A deep mathematical understanding of DNNs

James JIANG
Data Engineer / Scientist
France

Alex JIANG
Preparatory class for the Grandes Écoles
France

iLoveDataJjia Github

Version: 0.00

Date: December 28, 2021

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, 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.

Keywords: Dense neural network, Equation, Proof

1 Fundamentals

1.1 Matrices

Notation 1. Let $a_{i,j} \in \mathbb{R}$ for $i \in [\![1, n]\!]$ and $j \in [\![1, m]\!]$. Then a real matrix of dimension $n * m$ will 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

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

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

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

The notation $\mathscr{M}_{n,m}$ means the matrix set of dimension $n \times m$ with coefficients in $\mathbb{R}$.

The notation $\mathscr{M}_{n,m}(E)$ means the matrix set of dimension $n \times m$ with coefficients in $E \subseteq \mathbb{R}$.

Convention 1. A vector is a matrix with only one row. Thus, the real vector set $\mathbb{R}^n$ is equivalent to $\mathscr{M}_{1,n}$.

Notation 2. Let $A \in \mathscr{M}_{n,m}$ and $B \in \mathscr{M}_{m,p}$. Let the product noted $A * B$ be

$$C = A * B$$

where $C \in \mathscr{M}_{n,p}$ with

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

Notation 3. The matrix transpose operation will be noted as A^T.

Notation 4. Let $a \in \mathbb{R}^n$. The eucliean norm on $\mathbb{R}^n$ will be noted as $\|a\|_n$.

$$\|a\|_n = \sqrt{a * a^T}$$

1.2 Differential calculus

Convention 2. All sets considered are not empty.

Notation 5. Let $E \subseteq \mathbb{R}^n$ and $F \subseteq \mathbb{R}^m$.

The notation $\mathring{E}$ means the interior of E.

The notation $\overline{E}$ means the adherence of E.

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

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

The notation $\mathscr{C}(E,F)$ 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 1.1. Let $E \subseteq \mathbb{R}^n$ and $F \subseteq \mathbb{R}^m$. Then f differentiable on E is equivalent to

$$\begin{gathered} \forall a \in \mathring{E}, \exists \frac{\partial f}{\partial \cdot}(a) \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^m), \\ \forall h \in \mathbb{R}^n, f(a+h) = f(a) + \frac{\partial f}{\partial h}(a) + \underset{h \to 0}{o}(\|h\|_n) \end{gathered} \tag{1}$$

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

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

Proposition 1.1. Let $E \subseteq \mathbb{R}^n$, $F \subseteq \mathbb{R}^m$ and $f \in \mathscr{D}(E,F)$. Then $\frac{\partial f}{\partial \cdot}(a)$ is unique and $\mathscr{D}(E,F) \subset \mathscr{C}(E,F)$.

Proof. Suppose ϕ_1 and ϕ_2 two differentiales of f on a.

$$\forall h \in \mathbb{R}^n, \phi_2(h) - \phi_1(h) \underset{(1)}{=} \underset{h\to 0}{o}(\|h\|_n)$$

$$\underset{def}{\Longrightarrow} \forall \epsilon > 0, \exists \eta > 0, \forall h \in \mathbb{R}^n, (\|h\|_n \leq \eta \Rightarrow \left\|\phi_2(h) - \phi_1(h)\right\|_m \leq 2 * \|h\|_n * \epsilon)$$

$$\underset{\phi_2-\phi_1 \in \mathscr{L}(\mathbb{R}^n,\mathbb{R}^m)}{\Longrightarrow} \forall \epsilon > 0, \forall h \in \mathbb{R}^n, \left\|\phi_2(h) - \phi_1(h)\right\|_m \leq 2 * \|h\|_n * \epsilon$$

$$\underset{\epsilon \to 0}{\Longrightarrow} \forall h \in \mathbb{R}^n, \phi_2(h) = \phi_1(h)$$

Let $f \in \mathscr{D}(E,F)$. and $a \in \mathring{E}$.

$$\frac{\partial f}{\partial \cdot}(a) \in \mathscr{L}(\mathbb{R}^n,\mathbb{R}^m) \implies \frac{\partial f}{\partial \cdot}(a) \in \mathscr{C}(\mathbb{R}^n,\mathbb{R}^m), \frac{\partial f}{\partial 0_{\mathbb{R}^n}}(a) = 0_{\mathbb{R}^m}$$

$$\underset{(1)}{\Longrightarrow} f(a+h) \underset{h\to 0}{\to} f(a)$$

□

Definition 1.2. Let $E \subseteq \mathbb{R}^n$, $F \subseteq \mathbb{R}^m$ and $f = (f_1 \dots f_m) \in \mathscr{D}(E,F)$. Then f_i is differentiable on E for all $i \in [\![1,m]\!]$. The jacobian is defined as

$$\begin{aligned} \mathscr{J}_f \quad : \quad \mathring{E} \quad &\longrightarrow \mathscr{M}_{m,n} \\ a \quad &\longmapsto \begin{bmatrix} \frac{\partial f}{\partial e_1}(a) & \cdots & \frac{\partial f}{\partial e_n}(a) \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial e_1}(a) & \cdots & \frac{\partial f_1}{\partial e_n}(a) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial e_1}(a) & \cdots & \frac{\partial f_m}{\partial e_n}(a) \end{bmatrix} \end{aligned} \tag{2}$$

$(e_i)_{i\in[\![1,n]\!]}$ means the matrices $e_i = \begin{bmatrix} 0 & \cdots & \underset{\text{at index 0}}{1} & \cdots & 0 \end{bmatrix}$ corresponding to $\mathbb{R}^n$ standard basis.
$\frac{\partial f}{\partial e_i}$ is named the partial derivative of f according the i^{th} variable.
The jacobian is also named gradient when $m = 1$ and is noted as $\nabla_f = \mathscr{J}_f$.
The jacobian is also named derivative when $m = 1$ with $n = 1$ and is noted as $f' = \nabla_f = \mathscr{J}_f$.

Proof. Suppose $f = (f_1 \dots f_m) \in \mathscr{D}(E,F)$. Let $i \in [\![1,m]\!]$, $a \in \mathring{E}$ and $h \in \mathbb{R}^n$.

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

$$\frac{\partial f}{\partial \cdot}(a)_i \in \mathscr{L}(\mathbb{R}^n,\mathbb{R}) \underset{prop1.1}{\Longrightarrow} \frac{\partial f_i}{\partial h}(a) = \frac{\partial f}{\partial h}(a)_i$$

□

Corollary. Let $E \subseteq \mathbb{R}^n$, $F \subseteq \mathbb{R}^m$ and $f \in \mathscr{D}(E,F)$. The jacobian of f on $a \in \mathring{E}$ fixed is the canonical associated matrix to the differential of f on a.

Notes: It means a function differentiability can also be proved by exhibing its jacobian.

Proof. Let $a \in \mathring{E}$. $\frac{\partial f}{\partial \cdot}(a) \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ and any linear application in finite dimension with values in $\mathbb{R}$ has an unique associated matrix in the standard basis called canonical associated matrix.

□

Notation 6. Let $f \in \mathscr{F}(E,F)$ and $g \in \mathscr{F}(F,G)$. Then the notation $g \circ f$ means the application

$$\begin{aligned} g \circ f \quad : \quad E &\longrightarrow G \\ x &\longmapsto g(f(x)) \end{aligned}$$

Let $f_i \in \mathscr{F}(E_i, E_{i+1})$ for $i \in [\![1,n]\!]$. Then the notation $\overset{n}{\underset{i=1}{\circ}} f_i$ means the application

$$\begin{aligned} \overset{n}{\underset{i=1}{\circ}} f_i \quad : \quad E_1 &\longrightarrow E_{n+1} \\ x &\longmapsto f_n(\dots f_2(f_1(x))) \end{aligned}$$

Theorem 1.2. Let $E \subseteq \mathbb{R}^n$, $F \subseteq \mathbb{R}^m$, $G \subseteq \mathbb{R}^p$, $f \in \mathscr{D}(E,F)$ and $g \in \mathscr{D}(E,F)$. Then $g \circ f \in \mathscr{D}(E,G)$ and

$$\begin{aligned} \mathscr{J}_{g\circ f} \quad : \quad \mathring{E} &\longrightarrow G \\ a &\longmapsto \mathscr{J}_g(f(a)) * \mathscr{J}_f(a) \end{aligned} \tag{3}$$

Proof. Let $E \subseteq \mathbb{R}^n$, $F \subseteq \mathbb{R}^m$, $G \subseteq \mathbb{R}^p$, $f \in \mathscr{D}(E,F)$ and $g \in \mathscr{D}(F,G)$. Let $a \in \mathring{E}$ and $h \in \mathbb{R}^n$.

$$\begin{aligned} (g \circ f)(a+h) &= g(f(a) + \frac{\partial f}{\partial h}(a) + \underset{h\to 0}{o}(\|h\|_n)) \\ &= g(f(a)) + \frac{\partial g}{\partial(\frac{\partial f}{\partial h}(a) + \underset{h\to 0}{o}(\|h\|_n))}(f(a)) + \underset{h\to 0}{o}(\left\| \frac{\partial f}{\partial h}(a) + \underset{h\to 0}{o}(\|h\|_n) \right\|_n) \\ &\underset{\frac{\partial f}{\partial \cdot}(a) \in \mathscr{C}(\mathbb{R}^n,\mathbb{R}^m),\, \frac{\partial f}{\partial 0_{\mathbb{R}^n}}(a) = 0_{\mathbb{R}^m}}{=} g(f(a)) + \frac{\partial g}{\partial(\frac{\partial f}{\partial h}(a) + \underset{h\to 0}{o}(\|h\|_n))}(f(a)) + \underset{h\to 0}{o}(\|h\|_n) \\ &\underset{\frac{\partial g}{\partial \cdot}(a) \in \mathscr{L}(\mathbb{R}^m,\mathbb{R}^p)}{=} g(f(a)) + \frac{\partial g}{\partial(\frac{\partial f}{\partial h}(a))}(f(a)) + \frac{\partial g}{\partial(\underset{h\to 0}{o}(\|h\|_n))}(f(a)) + \underset{h\to 0}{o}(\|h\|_n) \\ &\underset{\frac{\partial g}{\partial \cdot}(a) \in \mathscr{C}(\mathbb{R}^m,\mathbb{R}^p),\, \frac{\partial g}{\partial 0_{\mathbb{R}^m}}(a) = 0_{\mathbb{R}^p}}{=} g(f(a)) + \frac{\partial g}{\partial(\frac{\partial f}{\partial h}(a))}(f(a)) + \underset{h\to 0}{o}(\|h\|_n) \end{aligned}$$

$$\frac{\partial g}{\partial(\frac{\partial f}{\partial \cdot}(a))}(f(a)) = \frac{\partial g}{\partial \cdot}(f(a)) \circ \frac{\partial f}{\partial \cdot}(a) \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^p) \underset{prop1.1}{\Longrightarrow} g \circ f \in \mathscr{D}(\mathbb{R}^n, \mathbb{R}^p), \frac{\partial (g\circ f)}{\partial \cdot}(a) = \frac{\partial g}{\partial \cdot}(f(a)) \circ \frac{\partial f}{\partial \cdot}(a)$$

$$\underset{mat}{\Longrightarrow} \mathscr{J}_{g\circ f}(a) = \mathscr{J}_g(f(a)) * \mathscr{J}_f(a)$$

Note: mat indicates in canonical associated matrix way.

□

1.3 Others

Notation 7. The notation $\delta_{\cdot,\cdot}$ means the kronecker delta application

$$\begin{aligned} \delta_{\cdot,\cdot} \quad : \quad \mathbb{Z} \times \mathbb{Z} &\longrightarrow \{0,1\} \\ (i,j) &\longmapsto \begin{matrix} 1 & i = j \\ 0 & i \neq j \end{matrix} \end{aligned}$$

Notation 8. Let $E \subseteq \mathbb{R}^n$. The notation $\mathbb{1}_E$ means the E indicator function on $\mathbb{R}^n$.

$$\begin{array}{rcll} \mathbb{1}_E & : & E & \longrightarrow \{0,1\}^n \\ & & x & \longmapsto \begin{array}{ll} 1 & x \in E \\ 0 & x \notin E \end{array} \end{array}$$

Notation 9. The notation $max(0,\cdot)$ means the application

$$\begin{array}{rcll} max(0,\cdot) & : & \mathbb{R} & \longrightarrow \mathbb{R}^+ \\ & & x & \longmapsto \begin{array}{ll} x & x > 0 \\ 0 & x \leq 0 \end{array} \end{array}$$

Assumption 1. $max(0,\cdot) \in \mathscr{D}(\mathbb{R},\mathbb{R}^+)$ with

$$\begin{array}{rcll} max(0,\cdot)' & : & \mathbb{R} & \longrightarrow \mathbb{R}^+ \\ & & x & \longmapsto \mathbb{1}_{\mathbb{R}^+}(x) \end{array}$$

Note: $max(0,\cdot)$ is actually not differentiable on 0.

Notation 10. Let f an application with n inputs and m outputs.

$$\begin{array}{rcll} f & : & E_1 \times \ldots \times E_n & \longrightarrow F_1 \times \ldots \times F_m \\ & & (x_1,\ldots,x_n) & \longmapsto f(x_1,\ldots,x_n) \end{array}$$

Let $k \in [\![1,n]\!]$. The notation $f(x_1,\ldots,x_{k-1},\cdot,x_{k+1},\ldots,x_n)$ means

$$\begin{array}{rcll} f(x_1,\ldots,x_{k-1},\cdot,x_{k+1},\ldots,x_n) & : & E_k & \longrightarrow F_1 \times \ldots \times F_m \\ & & x_k & \longmapsto f(x_1,\ldots,x_{k-1},x_k,x_{k+1},\ldots,x_n) \end{array}$$

2 Activation functions

Definition 2.1. Let $E \subseteq \mathbb{R}^m$, $F \subseteq \mathbb{R}^m$ and $F_{act} \in \mathscr{D}(E,F)$.

F_{act} is an activation function.

The notation $\mathscr{F}_{act}(E,F)$ means the set of activation functions from E to F.

Definition 2.2. Let the application *ReLU* noted as $\mathscr{R}$ be

$$\begin{array}{rcll} \mathscr{R} & : & \mathbb{R}^m & \longrightarrow \mathbb{R}^m \\ & & z & \longmapsto \begin{bmatrix} max(0,z_1) \\ \vdots \\ max(0,z_m) \end{bmatrix} \end{array}$$

Proposition 2.1. $\mathscr{R} = (\mathscr{R}_1 \ldots \mathscr{R}_m) \in \mathscr{F}_{act}(\mathbb{R}^m,\mathbb{R}^m)$ and its jacobian is

$$\begin{array}{rcll} \mathscr{J}_{\mathscr{R}} & : & \mathbb{R}^m & \longrightarrow \mathscr{M}_{m,m} \\ & & z & \longmapsto \begin{bmatrix} \mathbb{1}_{\mathbb{R}^+}(z_1) & 0 & \cdots & 0 \\ 0 & \mathbb{1}_{\mathbb{R}^+}(z_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbb{1}_{\mathbb{R}^+}(z_m) \end{bmatrix} \end{array} \tag{4}$$

Proof. Let $i \in \llbracket 1, m \rrbracket$, $j \in \llbracket 1, m \rrbracket$ and $z \in \mathbb{R}^m$.

$$\mathscr{R}_i(z) = max(0, z_i) \underset{assump1}{\Longrightarrow} \frac{\partial \mathscr{R}_i}{\partial e_j}(z) = \begin{matrix} \mathbb{1}_{\mathbb{R}^+}(z_i) & i = j \\ 0 & i \neq j \end{matrix}$$

□

Definition 2.3. Let the application *Softmax* noted as $\mathscr{S}$ be

$$\begin{aligned} \mathscr{S} \quad : \quad \mathbb{R}^m \quad & \longrightarrow]0,1[^m \\ z \quad & \longmapsto \begin{bmatrix} \frac{e^{z_1}}{\sum_{k=1}^m e^{z_k}} \\ \vdots \\ \frac{e^{z_m}}{\sum_{k=1}^m e^{z_k}} \end{bmatrix} \end{aligned}$$

Proposition 2.2. $\mathscr{S} = (\mathscr{S}_1 \dots \mathscr{S}_m) \in \mathscr{F}_{act}(\mathbb{R}^m,]0,1[^m)$ and its jacobian is

$$\begin{aligned} \mathscr{J}_{\mathscr{S}} \quad : \quad \mathbb{R}^m \quad & \longrightarrow \mathscr{M}_{m,m} \\ z \quad & \longmapsto \begin{bmatrix} \mathscr{S}_1 * (1 - \mathscr{S}_1) & -\mathscr{S}_1 * \mathscr{S}_2 & \cdots & -\mathscr{S}_1 * \mathscr{S}_m \\ -\mathscr{S}_2 * \mathscr{S}_1 & \mathscr{S}_2 * (1 - \mathscr{S}_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\mathscr{S}_{m-1} * \mathscr{S}_m \\ -\mathscr{S}_m * \mathscr{S}_1 & \cdots & -\mathscr{S}_m * \mathscr{S}_{m-1} & \mathscr{S}_m * (1 - \mathscr{S}_m) \end{bmatrix} (z) \end{aligned} \tag{5}$$

Proof. Let $i \in \llbracket 1, m \rrbracket$, $j \in \llbracket 1, m \rrbracket$ and $z \in \mathbb{R}^m$.

$$\begin{aligned} \mathscr{S}_i(z) &= \frac{e^{z_i}}{\sum_{k=1}^m e^{z_k}} \\ \Longrightarrow \frac{\partial \mathscr{S}_i}{\partial e_j} &= \frac{(\delta_{i,j} * e^{z_i}) * \sum_{k=1}^m e^{z_k} - e^{z_j} * e^{z_i}}{(\sum_{k=1}^m e^{z_k})^2} \\ &= \delta_{i,j} * \mathscr{S}_i(z) - \mathscr{S}_j(z) * \mathscr{S}_i(z) \\ &= \mathscr{S}_i(z) * (\delta_{i,j} - \mathscr{S}_j(z)) \end{aligned}$$

□

3 Loss

Definition 3.1. Let $E \subseteq \mathbb{R}^m$, $F \subseteq \overline{E}$, $F_{loss} \in \mathscr{F}(E \times F, \mathbb{R})$ with $\forall y^* \in F, F_{loss}(\cdot, y^*) \in \mathscr{D}(E, \mathbb{R})$. F_{loss} is a loss function is equivalent to

$$\forall y^* \in F,$$

$$\exists g \in \mathscr{F}(\{\epsilon \in \mathbb{R}^m | y^* + \epsilon \in E\}, E),$$

$$\begin{aligned} F_{loss}(\cdot, y^*) \circ g \quad : \quad \{\epsilon \in \mathbb{R}^m | y^* + \epsilon \in E\} \quad & \longrightarrow \mathbb{R} \\ \epsilon \quad & \longmapsto F_{loss}(y^* + \epsilon, y^*) \end{aligned}$$

is an inscreasing function according $\|\epsilon\|_m$.

y^* is named the ground truth matrix.

The notation $\mathscr{F}_{loss}(E)$ means the set of loss functions from $E \times F$ (with $F \subseteq \overline{E}$) to $\mathbb{R}$.

Proposition 3.1. Let $E \subseteq \mathbb{R}^m$, $F \subseteq \overline{E}$, $F_{loss} \in \mathscr{F}_{loss}(E)$. F_{loss} is a loss function is equivalent to

$$\forall y^* \in F,$$

$$\begin{array}{rcll} \phi_{F_{loss}}(\cdot, y^*) & : & \{\epsilon \in \mathbb{R}^m | y^* + \epsilon \in E\} & \longrightarrow \mathbb{R} \\ & & \epsilon & \longmapsto F_{loss}(y^* + \epsilon, y^*) \end{array}$$

is an inscreasing function according $\|\epsilon\|_m$.

Note: It means only the increasing aspect as to be proved.

Proof. Let $E \subseteq \mathbb{R}^m$, $F \subseteq \overline{E}$, $y \in E$ and $y^* \in F$.

$$\begin{aligned} \epsilon = y - y^* & \implies y^* + \epsilon \in \{\epsilon \in \mathbb{R}^m | y^* + \epsilon \in E\} \\ & \implies \{\epsilon \in \mathbb{R}^m | y^* + \epsilon \in E\} \neq \emptyset \\ & \implies \exists g \in \mathscr{F}(\{\epsilon \in \mathbb{R}^m | y^* + \epsilon \in E\}, E) \end{aligned}$$

□

Definition 3.2. Let the application *Categorical cross-entropy* noted as ξ be

$$\begin{array}{rcll} \xi & : &]0,1[^m \times \{0,1\}^m & \longrightarrow \mathbb{R}^m \\ & & (y, y^*) & \longmapsto -\sum_{k=1}^m y_k^* * \log(y_k) \end{array}$$

Proposition 3.2. $\xi \in \mathscr{F}_{loss}(]0,1[^m)$ and $\forall y^* \in \{0,1\}^m$, $\xi(\cdot, y^*)$ gradient is

$$\begin{array}{rcll} \nabla_{\xi(\cdot,y^*)} & : &]0,1[^m & \longrightarrow \mathbb{R}^m \\ & & y & \longmapsto -\begin{bmatrix} \frac{y_1^*}{y_1} & \dots & \frac{y_m^*}{y_m} \end{bmatrix} \end{array} \tag{6}$$

Proof. Suppose $E =]0,1[^m$, $F = \{0,1\}^m$. Then $F \subseteq \overline{E}$.

Let $y^* \in F$ and suppose $A = \{\epsilon \in \mathbb{R}^m | y^* + \epsilon \in E\}$. Let $(\epsilon_1, \epsilon_2) \in A^2$ with $\|\epsilon_1\|_m \leq \|\epsilon_2\|_m$. □