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: January 6, 2022

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, Riemann integral theory, 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. All the propositions will be mathematically proven as far as possible and under assumptions if necessary. 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 with decay and momentum. But all the elements below can be easily re-used or re-defined to cover regressions.

Keywords: Dense neural network, Differentiability, Continuous optimization

1 Fundamentals

1.1 Matrices

Convention 1. All sets considered are not empty.

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 2. Let $E \subseteq \mathbb{R}$.

A vector is a matrix with only one row. Thus, the vector set E^n is equivalent to $\mathscr{M}_{1,n}(E)$.

A m-tuple of vectors is a matrix with m rows. Thus, the cartesian products of vectors $(E^n)^m$ is equivalent to $\mathscr{M}_{m,n}(E)$.

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. The notation I_n means the identity matrix of size $n \times n$.

$$I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix}$$

Notation 5. 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

Notation 6. Let $E \subseteq \mathbb{R}^n$ and $F \subseteq \mathbb{R}^m$.

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

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

$\frac{\partial f}{\partial \cdot}(a)$ 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 $E \subseteq \mathbb{R}^n$, $F \subseteq \mathbb{R}^m$, $f \in \mathscr{D}(E,F)$ and $a \in \mathring{E}$. 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 differentials of f on a.

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

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

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

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

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

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

$(e_i)_{i\in[\![1,n]\!]}$ means the matrices $e_i = \begin{bmatrix}0 & \cdots & \underset{\text{at column } i}{1} & \cdots & 0\end{bmatrix}$ 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. Let $i \in [\![1,m]\!]$, $a \in \mathring{E}$ and $h \in \mathbb{R}^n$.

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

$$\frac{\partial f}{\partial \cdot}(a)_i \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}) \underset{prop 1.1}{\Longrightarrow} \frac{\partial f_i}{\partial h}(a) = \frac{\partial f}{\partial h}(a)_i$$

□

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 exhibiting 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.

□

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

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

Proof. Let $a \in \mathring{E}$ and $h \in \mathbb{R}^n$.

$$\begin{aligned} (g+f)(a+h) &= g(a+h) + f(a+h) \\ &\underset{(1)}{=} g(a) + f(a) + \frac{\partial g}{\partial h}(a) + \frac{\partial f}{\partial h}(a) + 2 * \underset{h \to 0}{o}(\|h\|_n) \\ &= (g+f)(a) + \frac{\partial g}{\partial h}(a) + \frac{\partial f}{\partial h}(a) + \underset{h \to 0}{o}(\|h\|_n) \end{aligned}$$

$$\frac{\partial g}{\partial \cdot}(a) + \frac{\partial f}{\partial \cdot}(a) \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^m) \underset{prop 1.1}{\Longrightarrow} g + f \in \mathscr{D}(E,F), \frac{\partial (g+f)}{\partial \cdot}(a) = \frac{\partial g}{\partial \cdot}(a) + \frac{\partial f}{\partial \cdot}(a)$$

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

Note: mat indicates in canonical associated matrix way.

□

Notation 7. 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 $\underset{i=1}{\overset{n}{\circ}} f_i$ means the application

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

Theorem 1.3. Let $E \subseteq \mathbb{R}^n$, $F \subseteq \mathbb{R}^m$, $G \subseteq \mathbb{R}^p$, $f \in \mathscr{D}(E,F)$ and $g \in \mathscr{D}(F,G)$. Then $g \circ f \in \mathscr{D}(E,G)$ and

$$\begin{aligned} \mathscr{J}_{g \circ f} \quad : \quad \mathring{E} &\longrightarrow G \\ a &\longmapsto \mathscr{J}_g(f(a)) * \mathscr{J}_f(a) \end{aligned} \tag{4}$$

Note: This theorem is named the chain rule.

Proof. Let $a \in \mathring{E}$ and $h \in \mathbb{R}^n$.

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

$$\frac{\partial g}{\partial(\frac{\partial f}{\partial\cdot}(a))}(f(a))=\frac{\partial g}{\partial\cdot}(f(a))\circ\frac{\partial f}{\partial\cdot}(a)\in\mathscr{L}(\mathbb{R}^n,\mathbb{R}^p)\underset{prop 1.1}{\Longrightarrow} g\circ f\in\mathscr{D}(E,G),\frac{\partial(g\circ f)}{\partial\cdot}(a)=\frac{\partial g}{\partial\cdot}(f(a))\circ\frac{\partial f}{\partial\cdot}(a)$$

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

Note: mat indicates in canonical associated matrix way.

□

1.3 Others

Notation 8. Let E and F two sets and $(E_i)_{i\in[\![1,n]\!]}$ n sets.

The notation $E\times F$ means the cartesian product between E and F.

The notation $\overset{n}{\underset{i=1}{\circ}} E_i$ means the cartesian product $E_n\times\ldots\times E_1$.

Notation 9. The notation $\delta_{\cdot,\cdot}$ means the kronecker delta application

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

Notation 10. Let $E\subseteq\mathbb{R}^n$. The notation $\mathbb{1}_E$ means the indicator function of E on $\mathbb{R}^n$.

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

Notation 11. The notation $max(0,\cdot)$ means the application

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

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

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

Note: $max(0,\cdot)$ is actually not *differentiable* on 0 and the notation $\mathbb{R}^*$ means $\mathbb{R}_{\backslash\{0\}}$.

Notation 12. Let f an application with n inputs and m outputs.

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

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

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

2 Activation functions

Notation 13. Let $E \subseteq \mathbb{R}^m \times (\mathbb{R}^{\cdot})^p$ (p parameter vectors of any sizes) and $F \subseteq \mathbb{R}^m$.

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

Note: An activation function is an application defined in this section.

Definition 2.1. Let the activation function *ReLU* noted as $\mathscr{R}$ be

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

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

$$\begin{aligned} \mathscr{J}_{\mathscr{R}} \quad : \quad \mathbb{R}^m &\longrightarrow \mathscr{M}_{m,m} \\ z &\longmapsto \begin{bmatrix} \mathbb{1}_{\mathbb{R}^{+*}}(z_1) & 0 & \cdots & 0 \\ 0 & \mathbb{1}_{\mathbb{R}^{+*}}(z_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbb{1}_{\mathbb{R}^{+*}}(z_m) \end{bmatrix} \end{aligned} \tag{5}$$

Proof. Let $i \in [\![1, m]\!]$, $j \in [\![1, m]\!]$ and $z \in \mathbb{R}^m$.

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

□

Definition 2.2. Let the activation function *Softmax* noted as $\mathscr{S}$ be

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

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

$$\begin{array}{rcll} \mathscr{J}_{\mathscr{S}} & : & \mathbb{R}^m & \longrightarrow \mathscr{M}_{m,m} \\ & & z & \longmapsto \begin{bmatrix} \mathscr{S}_1 * (1-\mathscr{S}_1) & -\mathscr{S}_1 * \mathscr{S}_2 & \cdots & -\mathscr{S}_1 * \mathscr{S}_m \\ -\mathscr{S}_2 * \mathscr{S}_1 & \mathscr{S}_2 * (1-\mathscr{S}_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\mathscr{S}_{m-1} * \mathscr{S}_m \\ -\mathscr{S}_m * \mathscr{S}_1 & \cdots & -\mathscr{S}_m * \mathscr{S}_{m-1} & \mathscr{S}_m * (1-\mathscr{S}_m) \end{bmatrix} (z) \end{array} \tag{6}$$

Proof. Let $i \in [\![1,m]\!]$, $j \in [\![1,m]\!]$ and $z \in \mathbb{R}^m$.

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

□

3 Loss

Notation 14. Let $E \subseteq \mathbb{R}^m \times \mathbb{R}^m$, $F \subseteq \mathbb{R}$.

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

Note: A loss function is an application defined in this section.

Definition 3.1. Let the loss function *Categorical cross-entropy* noted as ξ be

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

Proposition 3.1. Let $y^* \in \{0,1\}^m$. $\xi(\cdot, y^*) \in \mathscr{D}(]0,1[^m, \mathbb{R})$ and its gradient is

$$\begin{array}{rcll} \nabla_{\xi(\cdot,y^*)} & : &]0,1[^m & \longrightarrow \mathbb{R} \\ & & y & \longmapsto -\begin{bmatrix} \frac{y_1^*}{y_1} & \ldots & \frac{y_m^*}{y_m} \end{bmatrix} \end{array} \tag{7}$$

Proof. Let $j \in [\![1,m]\!]$ and $y \in]0,1[^m$.

$$\xi(y, y^*) = -\sum_{k=1}^m y_k^* * \log(y_k) \quad \Longrightarrow \frac{\partial \xi(\cdot, y^*)}{\partial e_j}(y) \quad = -\frac{y_j^*}{y_j}$$

□

4 Layers

Notation 15. Let $E \subseteq \mathbb{R}^n \times (\mathbb{R}^{\cdot})^p$ (p parameter vectors of any sizes) and $F \subseteq \mathbb{R}^m$.

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

Note: A layer function is an application defined in this section.

Definition 4.1. Let the layer function *Dense layer* noted as $\mathbb{L}$ be

$$\begin{array}{llll} \mathbb{L} & : & \mathbb{R}^n \times \mathcal{M}_{m,n} \times \mathbb{R}^m & \longrightarrow \mathbb{R}^m \\ & & (y, W, b) & \longmapsto y * W^T + b \end{array}$$

Note: $\mathbb{L} : \mathbb{R}^n \times \mathcal{M}_{m,n} \times \mathbb{R}^m \longrightarrow \mathbb{R}^m$ is equivalent to $\mathbb{L} : \mathbb{R}^n \times (\mathbb{R}^n)^m \times \mathbb{R}^m \longrightarrow \mathbb{R}^m$.

Proposition 4.1. Let $W \in \mathcal{M}_{m,n}$ and $b \in \mathbb{R}^m$. $\mathbb{L}(\cdot, W, b) = (\mathbb{L}_1(\cdot, W, b) \ldots \mathbb{L}_m(\cdot, W, b)) \in \mathcal{D}(\mathbb{R}^n, \mathbb{R}^m)$ and its gradient is

$$\begin{array}{llll} \mathcal{J}_{\mathbb{L}(\cdot, W, b)} & : & \mathbb{R}^n & \longrightarrow \mathcal{M}_{m,n} \\ & & y & \longmapsto W \end{array} \tag{8}$$

Proof. Let $i \in [\![1, m]\!]$, $j \in [\![1, n]\!]$ and $y \in \mathbb{R}^n$.

$$\begin{array}{lll} \mathbb{L}(y, W, b) = y * W^T + b & \Longrightarrow \mathbb{L}_i(y, W, b) & = y * W_{i,:}^T + b_i \\ & \Longrightarrow \frac{\partial \mathbb{L}_i(\cdot, W, b)}{\partial e_j}(y) & = W_{i,j} \end{array}$$

□

Proposition 4.2. Let $y \in \mathbb{R}^n$, $(w^{(k)})_{k \in [\![1, m-1]\!]} \in (\mathbb{R}^n)^{m-1}$, $b \in \mathbb{R}^m$.

$\forall i^* \in [\![1, m]\!]$, $\mathbb{L}(y, w^{(1)}, \ldots, w^{(i^*-1)}, \underset{\text{at index } i^*}{\cdot}, w^{(i^*)}, \ldots, w^{(m-1)}, b) \in \mathcal{D}(\mathbb{R}^n, \mathbb{R}^m)$ and jacobians are

$$\forall i^* \in [\![1, m]\!],$$

$$\begin{array}{llll} \mathcal{J}_{\mathbb{L}(y, w^{(1)}, \ldots, w^{(i^*-1)}, \cdot, w^{(i^*)}, \ldots, w^{(m-1)}, b)} & : & \mathbb{R}^n & \longrightarrow \mathcal{M}^{m,n} \\ & & w & \longmapsto \begin{bmatrix} & (0) & \\ y_1 & \cdots & y_n \\ & (0) & \end{bmatrix} \text{at row } i^* \end{array} \tag{9}$$

Note: For $i^* = 1$ and $i^* = m$, the applications $\mathbb{L}(y, \cdot, w^{(1)}, \ldots, w^{(m-1)}, b)$ and $\mathbb{L}(y, w^{(1)}, \ldots, w^{(m-1)}, \cdot, b)$ are meant respectively.

Proof. Let $i^* \in [\![1, m]\!]$, $i \in [\![1, m]\!]$, $j \in [\![1, n]\!]$ and $w \in \mathbb{R}^n$.

$$\begin{aligned} & \mathbb{L}(y, w^{(1)}, \ldots, w^{(i^*-1)}, \cdot, w^{(i^*)}, \ldots, w^{(m-1)}, b)(w) \\ = & \left[y * w^{(1)T} \quad \cdots \quad y * w^{(i^*-1)T} \quad y * w^T \quad y * w^{(i^*)T} \quad \cdots \quad y * w^{(m-1)T} \right] + b \\ \Longrightarrow & \frac{\partial \mathbb{L}_i(y, w^{(1)}, \ldots, w^{(i^*-1)}, \cdot, w^{(i^*)}, \ldots, w^{(m-1)}, b)}{\partial e_j}(w) = \begin{array}{ll} y_j & i = i^* \\ 0 & i \neq i^* \end{array} \end{aligned}$$

□

Proposition 4.3. Let $y \in \mathbb{R}^n$ and $W \in \mathcal{M}_{m,n}$. $\mathbb{L}(y, W, \cdot) = (\mathbb{L}_1(y, W, \cdot) \ldots \mathbb{L}_m(y, W, \cdot)) \in \mathcal{D}(\mathbb{R}^m, \mathbb{R}^m)$ and its gradient is

$$\begin{array}{llll} \mathcal{J}_{\mathbb{L}(y, W, \cdot)} & : & \mathbb{R}^m & \longrightarrow \mathcal{M}_{m,m} \\ & & b & \longmapsto I_m \end{array} \tag{10}$$

Proof. Let $i \in [\![1,m]\!]$, $j \in [\![1,n]\!]$ and $b \in \mathbb{R}^m$.

$$\begin{aligned} \mathbb{L}(y,W,b) = y * W^T + b \quad &\implies \mathbb{L}_i(y,W,b) &&= y * W_{i,:}^T + b_i \\ &\implies \frac{\partial \mathbb{L}_i(y,W,\cdot)}{\partial e_j}(b) &&= \delta_{i,j} \end{aligned}$$

□

5 Neural networks

5.1 Simplified jacobian matrices

Proposition 5.1. Let $\mathscr{F}^{(upstream)} \in \mathscr{D}(\mathbb{R}^m, \mathbb{R})$ and $\mathscr{R} \in \mathscr{F}_{act}(\mathbb{R}^m, \mathbb{R}^m)$. $\mathscr{F}^{(upstream)} \circ \mathscr{R} \in \mathscr{D}(\mathbb{R}^m, \mathbb{R})$ and its gradient is

$$\begin{aligned} \nabla_{\mathscr{F}^{(upstream)} \circ \mathscr{R}} \quad : \quad & \mathbb{R}^m && \longrightarrow \mathbb{R}^m \\ & z && \longmapsto \begin{bmatrix} \nabla_{\mathscr{F}^{(upstream)}}(y)_1 * \mathbb{1}_{\mathbb{R}^{+*}}(z_1) & \cdots & \nabla_{\mathscr{F}^{(upstream)}}(y)_m * \mathbb{1}_{\mathbb{R}^{+*}}(z_m) \end{bmatrix} \end{aligned} \tag{11}$$

where

$$y = \mathscr{R}(z)$$

Note: It means such a gradient can be implemented without matrix multiplication.

Proof. Let $j \in [\![1,m]\!]$ and $z \in \mathbb{R}^m$.

$$\nabla_{\mathscr{F}^{(upstream)} \circ \mathscr{R}} \underset{(4)}{=} \nabla_{\mathscr{F}^{(upstream)}}(\mathscr{R}(z)) * \mathscr{J}_{\mathscr{R}}(z) \underset{(5)}{\implies} \frac{\partial \mathscr{F}^{(upstream)} \circ \mathscr{R}}{\partial e_j}(z) = \nabla_{\mathscr{F}^{(upstream)}}(\mathscr{R}(z))_j * \mathbb{1}_{\mathbb{R}^{+*}}(z_j)$$

□

Proposition 5.2. Let $\mathscr{F}^{(upstream)} \in \mathscr{D}(\mathbb{R}^m, \mathbb{R})$, $\mathbb{L} \in \mathscr{F}_{layer}(\mathbb{R}^n \times (\mathbb{R}^n)^m \times \mathbb{R}^m, \mathbb{R}^m)$, $y \in \mathbb{R}^n$, $(w^{(k)})_{k \in [\![1,m-1]\!]} \in (\mathbb{R}^n)^{m-1}$ and $b \in \mathbb{R}^m$.

$\forall i^* \in [\![1,m]\!]$, $\mathscr{F}^{(upstream)} \circ \mathbb{L}(y, w^{(1)}, \ldots, w^{(i^*-1)}, \underset{\text{at index } i^*}{\cdot}, w^{(i^*)}, \ldots, w^{(m-1)}, b) \in \mathscr{D}(\mathbb{R}^n, \mathbb{R})$ and gradients are

$$\forall i^* \in [\![1,m]\!],$$

$$\begin{aligned} \nabla_{\mathscr{F}^{(upstream)} \circ \mathbb{L}(y, w^{(1)}, \ldots, w^{(i^*-1)}, \cdot, w^{(i^*)}, \ldots, w^{(m-1)}, b)} \quad : \quad & \mathbb{R}^n && \longrightarrow \mathbb{R}^n \\ & w && \longmapsto \nabla_{\mathscr{F}^{(upstream)}}(z)_{i^*} * y \end{aligned} \tag{12}$$

where

$$z = \mathbb{L}(y, w^{(1)}, \ldots, w^{(i^*-1)}, w, w^{(i^*)}, \ldots, w^{(m-1)}, b)$$

Note: It means these gradients for $i^* \in [\![1,m]\!]$ can be implemented with $\nabla_{\mathscr{F}^{(upstream)}}(z)^T * y$.

Proof. Let $i^* \in [\![1,m]\!]$, $j \in [\![1,n]\!]$ and $w \in \mathbb{R}^n$. Let $z = \mathbb{L}(y, w^{(1)}, \ldots, w^{(i^*-1)}, w, w^{(i^*)}, \ldots, w^{(m-1)}, b)$.

$$\begin{aligned} \nabla_{\mathscr{F}^{(upstream)} \circ \mathbb{L}(y, w^{(1)}, \ldots, w^{(i^*-1)}, \cdot, w^{(i^*)}, \ldots, w^{(m-1)}, b)}(w) &\underset{(4)}{=} \nabla_{\mathscr{F}^{(upstream)}}(z) * \mathscr{J}_{\mathbb{L}(y, w^{(1)}, \ldots, w^{(i^*-1)}, \cdot, w^{(i^*)}, \ldots, w^{(m-1)}, b)}(w) \\ &\underset{(9)}{\implies} \frac{\partial \nabla_{\mathscr{F}^{(upstream)} \circ \mathbb{L}(y, w^{(1)}, \ldots, w^{(i^*-1)}, \cdot, w^{(i^*)}, \ldots, w^{(m-1)}, b)}}{\partial e_j}(w) = \nabla_{\mathscr{F}^{(upstream)}}(z)_{i^*} * y_j \end{aligned}$$

□

Proposition 5.3. Let $\mathscr{S} \in \mathscr{F}_{act}(\mathbb{R}^m,]0,1[^m)$ and $\xi \in \mathscr{F}_{loss}(]0,1[^m \times \{0,1\}^m, \mathbb{R})$.

Let $y^* \in \{0,1\}^m$ with $\|y^*\|_m = 1$. $\xi(\cdot, y^*) \circ \mathscr{S} \in \mathscr{D}(\mathbb{R}^m, \mathbb{R})$ and its gradient is

$$\begin{array}{rccl} \nabla_{\xi(\cdot,y^*)\circ\mathscr{S}} & : & \mathbb{R}^m & \longrightarrow \mathbb{R} \\ & & z & \longmapsto \mathscr{S}(z) - y^* \end{array} \tag{13}$$

Proof. Let $j \in [\![1, m]\!]$ and $z \in \mathbb{R}^m$.

$$\begin{aligned} \nabla_{\xi(\cdot,y^*)\circ\mathscr{S}}(z) &\underset{(4)}{=} \nabla_{\xi(\cdot,y^*)}(\mathscr{S}(z)) * \mathscr{J}_{\mathscr{S}}(z) \\ \Longrightarrow \frac{\partial \xi(\cdot,y^*)\circ\mathscr{S}}{\partial e_j}(z) &\underset{(6),(7)}{=} -y_j^* + \sum_{k=1}^m y_k^* * \mathscr{S}_j(z) \\ &\underset{y^*\in\{0,1\}^m, \|y\|_m=1}{=} \mathscr{S}_j(z) - y_j^* \end{aligned}$$

□

5.2 Definitions

Notation 16. Let $E \subseteq \mathbb{R}^n \times (\mathbb{R}^{\cdot})^p$ (p parameter vectors of any sizes) and $F \subseteq \mathbb{R}^m$.

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

Note: A neural network is an application defined in this section.

Definition 5.1. Let $(m_k)_{k\in[\![0,p]\!]} \in (\mathbb{N}^*)^p$. Let the neural network *Multi-class dense neural network* noted as $\mathscr{N}_{c^+}$ be

$$\begin{array}{rcll} \mathscr{N}_{c^+} & : & \mathbb{R}^{m_0} \times (\overset{p}{\underset{k=1}{\times}} \mathscr{M}_{m_k,m_{k-1}}) \times (\overset{p}{\underset{k=1}{\times}} \mathbb{R}^{m_k}) & \longrightarrow \mathbb{R}^{m_p} \\ & & (x, (W^{(k)})_{k\in[\![1,p]\!]}, (b^{(k)})_{k\in[\![1,p]\!]}) & \longmapsto (\mathscr{S} \circ \mathbb{L}^{(p)}(\cdot, W^{(p)}, b^{(p)})) \circ (\overset{p-1}{\underset{k=1}{\bigcirc}} \mathscr{R}^{(k)} \circ \mathbb{L}^{(k)}(\cdot, W^{(k)}, b^{(k)}))(x) \end{array}$$

where

$$\begin{aligned} &(\mathbb{L}^{(k)})_{k\in[\![1,p]\!]} \in \overset{p}{\underset{k=1}{\times}} \mathscr{F}_{layer}(\mathbb{R}^{m_{k-1}} \times \mathscr{M}_{m_k,m_{k-1}} \times \mathbb{R}^{m_k}, \mathbb{R}^{m_k}) \\ &(\mathscr{R}^{(k)})_{k\in[\![1,p-1]\!]} \in \overset{p-1}{\underset{k=1}{\times}} \mathscr{F}_{act}(\mathbb{R}^{m_k}, \mathbb{R}^{m_k}) \\ &\mathscr{S} \in \mathscr{F}_{act}(\mathbb{R}^{m_p},]0,1[^{m_p}) \end{aligned}$$

Note: $\mathscr{N}_{c^+} : \mathbb{R}^{m_0} \times (\overset{p}{\underset{k=1}{\times}} \mathscr{M}_{m_k,m_{k-1}}) \times (\overset{p}{\underset{k=1}{\times}} \mathbb{R}^{m_k}) \longrightarrow \mathbb{R}^{m_p}$ is equivalent to $\mathscr{N}_{c^+} : \mathbb{R}^{m_0} \times (\overset{p}{\underset{k=1}{\times}} (\mathbb{R}^{m_{k-1}})^{m_k}) \times (\overset{p}{\underset{k=1}{\times}} \mathbb{R}^{m_k}) \longrightarrow \mathbb{R}^{m_p}$.

Corollary. $\mathscr{N}_{c^+} \in \mathscr{D}(\mathbb{R}^{m_0} \times (\overset{p}{\underset{k=1}{\times}} (\mathbb{R}^{m_{k-1}})^{m_k}) \times (\overset{p}{\underset{k=1}{\times}} \mathbb{R}^{m_k}), \mathbb{R}^{m_p})$ and the total number of parameter is

$$\sum_{k=1}^p m_k * (m_{k-1} + 1)$$

Proof. $\mathscr{N}_{c^+}$ is a composition of *differentiable* applications so it is *differentiable* by the chain rule theorem. Let $a \in (\overset{p}{\underset{k=1}{\times}} (\mathbb{R}^{m_{k-1}})^{m_k}) \times (\overset{p}{\underset{k=1}{\times}} \mathbb{R}^{m_k})$ then a has $\sum_{k=1}^p m_k * m_{k-1} + \sum_{k=1}^p m_k$ coefficients.

□

Definition 5.2. Let $(m_k)_{k\in[\![0,p]\!]}$, $\mathscr{N}_{c^+} \in \mathscr{F}_{net}(\mathbb{R}^{m_0} \times (\underset{k=1}{\overset{p}{\times}} \mathscr{M}_{m_k,m_{k-1}}) \times (\underset{k=1}{\overset{p}{\times}} \mathbb{R}^{m_k}), \mathbb{R}^{m_p})$, $X = (x^{(i)})_{i\in[\![1,n]\!]} \in (\mathbb{R}^{m_0})^n$ and $Y^* = (y^{*(i)})_{i\in[\![1,n]\!]} \in (\{0,1\}^{m_p})^n$ with $\forall i \in [\![1,n]\!]$, $\left\|y^{*(i)}\right\|_{m_p} = 1$.

Let the *Multi-class optimization problem* noted as $(\mathscr{P}_{c^+})$ be

$$(\mathscr{P}_{c^+}) : \min_{(W^{(k)})_{k\in[\![1,p]\!]}, (b^{(k)})_{k\in[\![1,p]\!]}} \sum_{i=1}^{n} \xi(y^{(i)}, y^{*(i)}) \tag{14}$$

where

$$\xi \in \mathscr{F}_{loss}(]0,1[^{m_p} \times \{0,1\}^{m_p}, \mathbb{R})$$
$$y^{(i)} = \mathscr{N}_{c^+}(x^{(i)}, (W^{(k)})_{k\in[\![1,p]\!]}, (b^{(k)})_{k\in[\![1,p]\!]})$$

$\sum_{i=1}^{n} \xi(\cdot, y^{*(i)}) \circ \mathscr{N}_{c^+}(x^{(i)}, \cdot, \cdot)$ is named the objective function and will be noted as $\mathscr{O}_{c^+}(X, Y^*, \cdot, \cdot)$.

Corollary. $\mathscr{O}_{c^+}(X, Y^*, \cdot, \cdot) \in \mathscr{D}((\underset{k=1}{\overset{p}{\times}} \mathscr{M}_{m_k,m_{k-1}}) \times (\underset{k=1}{\overset{p}{\times}} \mathbb{R}^{m_k}), \mathbb{R})$.

Note: Its gradients for each variable can be computed recursively through each composition using (1.2), (4), (8), (12), (10), (11) and (13).

Proof. $\mathscr{O}_{c^+}(X, Y^*, \cdot, \cdot)$ is a sum and composition of *differentiable* applications so it is *differentiable* by the proposition 1.2 and chain rule theorem.

□

6 Gradient descent

6.1 Optimization fundamentals

Definition 6.1. Let $f \in \mathscr{D}(\mathbb{R}^m, \mathbb{R})$. f *convex* means

$$\begin{gathered} \forall x \in \mathbb{R}^m, \forall y \in \mathbb{R}^m, \\ \forall \tau \in [0,1], f(\tau * x + (1-\tau) * y) \leq \tau * f(x) + (1-\tau) * f(y) \end{gathered} \tag{15}$$

Proposition 6.1. Let $f \in \mathscr{D}(\mathbb{R}^m, \mathbb{R})$. f *convex* is equivalent to

$$\begin{gathered} \forall x \in \mathbb{R}^m, \forall y \in \mathbb{R}^m, \\ f(x) + \nabla_f(x) * (y-x)^T \leq f(y) \end{gathered} \tag{16}$$

Proof. Suppose f *convex* (15). Let $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^m$.

$$\forall \tau \in]0,1[, f(x + \tau * (y-x)) \underset{\|\cdot\|_m \in \mathscr{C}(\mathbb{R}^m,\mathbb{R}), f \in \mathscr{D}(\mathbb{R}^m,\mathbb{R}), (1)}{=} f(x) + \frac{\partial f}{\partial(\tau * (y-x))}(x) + \underset{\tau\to 0}{o}(\left\|\tau * (y-x)\right\|_m)$$

$$\begin{aligned}
&\underset{(15)}{\Longrightarrow} \forall \tau \in]0,1[, f(x) + \tfrac{\partial f}{\partial(\tau*(y-x))}(x) + \underset{\tau\to 0}{o}(\left\|\tau * (y-x)\right\|_m) \leq \tau * f(y) + (1-\tau) * f(x) \\
&\underset{\frac{\partial f}{\partial \cdot} \in \mathscr{L}(\mathbb{R}^m,\mathbb{R})}{\Longrightarrow} \forall \tau \in]0,1[, \tfrac{\partial f}{\partial(y-x)}(x) + \underset{\tau\to 0}{o}(\left\|y-x\right\|_m) \leq f(y) - f(x) \\
&\underset{\tau\to 0}{\Longrightarrow} \tfrac{\partial f}{\partial(y-x)}(x) \leq f(y) - f(x) \\
&\underset{mat}{\Longrightarrow} f(x) + \nabla_f(x) * (y-x)^T \leq f(y)
\end{aligned}$$

Suppose (16). Let $x \in \mathbb{R}^m$, $y \in \mathbb{R}^m$ and $\tau \in [0,1]$. Let $z = \tau * x + (1-\tau) * y$.

$$\begin{array}{lll} (a): & f(z) - (1-\tau) * \nabla_f(z) * (y-x)^T & \underset{(16)}{=} f(z) + \nabla_f(z) * (x-z)^T \leq f(x) \\ (b): & f(y) + \tau * \nabla_f(z) * (y-x)^T & \underset{(16)}{=} f(z) + \nabla_f(z) * (y-z)^T \leq f(y) \\ & \underset{\tau*(a)+(1-\tau)*(b)}{\Longrightarrow} f(z) \leq \tau * f(x) + (1-\tau) f(y) & \end{array}$$

□

Definition 6.2. Let $f \in \mathscr{D}(\mathbb{R}^m, \mathbb{R})$ and $L \in \mathbb{R}^{+*}$. f *L-smooth* means

$$\begin{gathered} \forall x \in \mathbb{R}^m, \forall y \in \mathbb{R}^m, \\ \left\| \nabla_f(x) - \nabla_f(y) \right\|_m \leq L * \left\| x - y \right\|_m \end{gathered} \tag{17}$$

Proposition 6.2. Let $f \in \mathscr{D}(\mathbb{R}^m, \mathbb{R})$ and $L \in \mathbb{R}^{+*}$. If f *L-smooth* then

$$\begin{gathered} \forall x \in \mathbb{R}^m, \forall y \in \mathbb{R}^m, \\ f(y) \leq f(x) + \nabla_f(x) * (y-x)^T + \frac{L}{2} * \left\| y - x \right\|_m^2 \end{gathered} \tag{18}$$

Proof. Let $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^m$. Let

$$\begin{array}{rcll} g & : & [0,1] & \longrightarrow \mathbb{R}^m \\ & & \tau & \longrightarrow x + \tau * (y-x) \end{array}$$

$$\begin{array}{ll} \forall \tau \in [0,1], (f \circ g)(\tau) = f(x + \tau * (y-x)) & \underset{(4)}{\Longrightarrow} \forall \tau \in [0,1], (f \circ g)'(\tau) = \nabla_f(g(\tau)) * (y-x)^T \\ & \underset{\int}{\Longrightarrow} f(y) - f(x) = \int_0^1 \nabla_f(g(\tau)) * (y-x)^T d\tau \end{array}$$

$$\begin{array}{ll} f(y) & = f(x) + \int_0^1 \nabla_f(g(\tau)) * (y-x)^T d\tau \\ & = f(x) + \nabla_f(x) * (y-x)^T + \int_0^1 (\nabla_f(g(\tau)) - \nabla_f(x)) * (y-x)^T d\tau \\ & \underset{Cauchy-Schwarz}{\leq} f(x) + \nabla_f(x) * (y-x)^T + \int_0^1 \left\| \nabla_f(g(\tau)) - \nabla_f(x) \right\|_m * \left\| y - x \right\|_m d\tau \\ & \underset{(17)}{\leq} f(x) + \nabla_f(x) * (y-x)^T + L * \left\| y - x \right\|_m^2 * \int_0^1 \tau d\tau \end{array}$$

□

Proposition 6.3. Let $L \in \mathbb{R}^{+*}$ and $f \in \mathscr{D}(\mathbb{R}^m, \mathbb{R})$. If f *convex* and *L-smooth* then

$$\begin{gathered} \forall x \in \mathbb{R}^m, \forall y \in \mathbb{R}^m, \\ \frac{1}{L} * \left\| \nabla_f(y) - \nabla_f(x) \right\|_m^2 \leq (\nabla_f(y) - \nabla_f(x)) * (y-x)^T \end{gathered} \tag{19}$$

Notes: This proposition is named the gradient co-coercivity.

Proof. Let $x \in \mathbb{R}^m$, $y \in \mathbb{R}^m$ and $z = x - \frac{1}{L}(\nabla_f(y) - \nabla_f(x))$.

$$\begin{array}{ll} f(y) - f(x) & = f(y) - f(z) + f(z) - f(x) \\ & \underset{(16),(18)}{\leq} \nabla_f(y) * (y-z)^T + \nabla_f(x) * (z-x) + \frac{L}{2} * \| z - x \|_m^2 \\ & \leq \nabla_f(y) * (y-x)^T - \frac{1}{2L} * \left\| \nabla_f(x) - \nabla_f(y) \right\|_m^2 \end{array}$$

It means the inequality is true for all $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^m$.

Let $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^m$. The previous inequality gives

$$\begin{aligned}
(a): \quad & f(y) - f(x) \leq \nabla_f(y) * (y-x)^T - \tfrac{1}{2L} * \left\|\nabla_f(x) - \nabla_f(y)\right\|_m^2 \\
(b): \quad & f(x) - f(y) \leq \nabla_f(x) * (x-y)^T - \tfrac{1}{2L} * \left\|\nabla_f(y) - \nabla_f(x)\right\|_m^2 \\
\underset{(a)+(b)}{\Longrightarrow} & 0 \leq (\nabla_f(y) - \nabla_f(x)) * (y-x)^T - \frac{1}{L} * \left\|\nabla_f(y) - \nabla_f(x)\right\|_m^2
\end{aligned}$$

□

6.2 Algorithms

7 References