# A Proposal on Standard Mathematical Notations of Machine Learning

January 16, 2020

**Summary**

The field of machine learning is evolving rapidly in recent years. Communication between different researchers and research groups becomes increasingly important. A key challenge for communication arises from inconsistent notation usages over different papers. This note suggests a standard for commonly used mathematical notations of machine learning. This standard will be regularly updated based on the progress of the field. We look forward to more suggestions to improve this note in future versions.

**Dataset**

Dataset $S = \{\boldsymbol{z}_i\}_{i=1}^n = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ is sampled from a distribution $\mathcal{D}$ over a domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.

$\mathcal{X}$ is the instance domain (a set), $\mathcal{Y}$ is the label domain (a set), and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ is the example domain (a set).

Usually, $\mathcal{X}$ is a subset of $\mathbb{R}^d$ and $\mathcal{Y}$ is a subset of $\mathbb{R}^{d_o}$, where $d$ is the input dimension, $d_o$ is the output dimension.

$n = \#S$ is the number of samples. Without specification, $S$ and $n$ are for the training set.

**Function**

Hypothesis space is denoted by $\mathcal{H}$. Hypothesis function is denoted by $f_{\boldsymbol{\theta}}(\boldsymbol{x}) \in \mathcal{H}$ or $f(\boldsymbol{x}; \boldsymbol{\theta}) \in \mathcal{H}$ with $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$.

$\boldsymbol{\theta}$ denotes the set of parameters of $f_{\boldsymbol{\theta}}$.

If there exists a target function, it is denoted by $f^*$or $f : \mathcal{X} \to \mathcal{Y}$ satisfying $\boldsymbol{y}_i = f^*(\boldsymbol{x}_i)$ for $i = 1, \ldots, n$.

**Loss function**

Loss function, denoted by $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+ := [0, +\infty)$, measures the difference between a predicted label and a true label, e.g., $L^2$ loss:

$$\ell(f_{\boldsymbol{\theta}}, \boldsymbol{z}) = (f_{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{y})^2,$$

where $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$. $\ell(f_{\boldsymbol{\theta}}, \boldsymbol{z})$ can also be written as

$$\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{y})$$

for convenience. Empirical risk or training loss for a set $S = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ is denoted by $L_S(\boldsymbol{\theta})$ or $L_n(\boldsymbol{\theta})$ or $\mathcal{R}_n(\boldsymbol{\theta})$,

$$L_S(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \boldsymbol{y}_i), \tag{1}$$

Without ambiguity, $L$ is also used for $L_S$.

The population risk or expected loss is denoted by

$$L_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{D}}\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{y}), \tag{2}$$

where $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ follows the distribution $\mathcal{D}$.

**Activation function** Activation function is denoted by $\sigma(x)$.

**Example 1.** *Some commonly used activation functions are*

1. $\sigma(x) = \mathrm{ReLU}(x) = \max(0, x)$;
2. $\sigma(x) = \mathrm{sigmoid}(x) = \frac{1}{1+\mathrm{e}^{-x}}$;
3. $\sigma(x) = \tanh x$;
4. $\sigma(x) = \cos x, \sin x$.

**Two-layer neural network**
The neuron number of the hidden layer is denoted by $m$. The two-layer neural network is

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{m} a_j \sigma(\boldsymbol{w}_j \cdot \boldsymbol{x} + b_j), \tag{3}$$

where $\sigma$ is the activation function, $\boldsymbol{w}_j$ is the input weight, $a_j$ is the output weight, $b_j$ is the bias term.

**General deep neural network**
The counting of the layer number excludes the input layer. The $L$-layer neural network is denoted by

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{W}^{[L-1]}\sigma \circ (\boldsymbol{W}^{[L-2]}\sigma \circ (\cdots(\boldsymbol{W}^{[1]}\sigma \circ (\boldsymbol{W}^{[0]}\boldsymbol{x} + \boldsymbol{b}^{[0]}) + \boldsymbol{b}^{[1]})\cdots) + \boldsymbol{b}^{[L-2]}) + \boldsymbol{b}^{[L-1]}, \tag{4}$$

where $\boldsymbol{W}^{[l]} \in \mathbb{R}^{m_{l+1}\times m_l}$, $\boldsymbol{b}^{[l]} = \mathbb{R}^{m_{l+1}}$, $m_0 = d_{\mathrm{in}} = d$, $m_L = d_{\mathrm{o}}$, $\sigma$ is a scalar function and "$\circ$" means entry-wise operation. We denote the set of parameters by

$$\boldsymbol{\theta} = (\boldsymbol{W}^{[0]}, \boldsymbol{W}^{[1]}, \ldots, \boldsymbol{W}^{[L-1]}, \boldsymbol{b}^{[0]}, \boldsymbol{b}^{[1]}, \ldots, \boldsymbol{b}^{[L-1]}),$$

and an entry of $\boldsymbol{W}^{[l]}$ by $\boldsymbol{W}_{ij}^{[l]}$. This can also be defined recursively.

$$f_{\boldsymbol{\theta}}^{[0]}(\boldsymbol{x}) = \boldsymbol{x}, \tag{5}$$

$$f_{\boldsymbol{\theta}}^{[l]}(\boldsymbol{x}) = \sigma \circ (\boldsymbol{W}^{[l-1]} f_{\boldsymbol{\theta}}^{[l-1]}(\boldsymbol{x}) + \boldsymbol{b}^{[l-1]}) \quad 1 \le l \le L-1, \tag{6}$$

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = f_{\boldsymbol{\theta}}^{[L]}(\boldsymbol{x}) = \boldsymbol{W}^{[L-1]} f_{\boldsymbol{\theta}}^{[L-1]}(\boldsymbol{x}) + \boldsymbol{b}^{[L-1]}. \tag{7}$$

**Complexity**
The VC-dimension of a hypothesis class $\mathcal{H}$ is denoted $\mathrm{VCdim}(\mathcal{H})$.

The Rademacher complexity of a hypothesis space $\mathcal{H}$ on a sample set $S$ is denoted by $\mathrm{Rad}(\mathcal{H} \circ S)$ or $\mathrm{Rad}_S(\mathcal{H})$. The complexity $\mathrm{Rad}_S(\mathcal{H})$ is random because of the randomness of $S$. The expectation of the empirical Rademacher complexity over all samples of size $n$ is denoted by $\mathrm{Rad}_n(\mathcal{H}) = \mathbb{E}_S \mathrm{Rad}_S(\mathcal{H})$.

**Training**
The Gradient Descent is often denoted by GD. The Stochastic Gradient Descent is often denoted by SGD.

A batch set is denoted by $B$ and the batch size is denoted by $b$.

The learning rate is denoted by $\eta$.

**Gram matrix**
The Gram matrix is denoted by $K_n$.

**Fourier Frequency**
The discretized frequency is denoted by $\boldsymbol{k}$, and the continuous frequency is denoted by $\boldsymbol{\xi}$.

**Convolution**
The convolution operation is denoted by $*$.

**More notations**
More notations about GAN, RNN, CNN, Resnet etc. are left to be done.

**Notation table**

| symbol | meaning | LaTeX | simplied |
|---|---|---|---|
| $\boldsymbol{x}$ | input | `\bm{x}` | `\vx` |
| $\boldsymbol{y}$ | output, label | `\bm{y}` | `\vy` |
| $d$ | input dimension | `d` | |
| $d_{\text{o}}$ | output dimension | `d_{\rm o}` | |
| $n$ | number of samples | `n` | |
| $\mathcal{X}$ | instances domain (a set) | `\mathcal{X}` | `\fX` |
| $\mathcal{Y}$ | labels domain (a set) | `\mathcal{Y}` | `\fY` |
| $\mathcal{Z}$ | $= \mathcal{X} \times \mathcal{Y}$ examples domain (a set) | `\mathcal{Z}` | `\fZ` |
| $\mathcal{H}$ | hypothesis space (a set) | `\mathcal{H}` | `\fH` |
| $\boldsymbol{\theta}$ | a set of parameters | `\bm{\theta}` | `\vtheta` |
| $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$ | hypothesis function with parameters $\boldsymbol{\theta}$ | `\f_{\bm{\theta}}` | `f_{\vtheta}` |
| $f$ or $f^* : \mathcal{X} \to \mathcal{Y}$ | target function | `f,f^*` | |
| $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$ | loss function | `\ell` | |
| $\mathcal{D}$ | a distribution over some set (usually $\mathcal{Z}$) | `\mathcal{D}` | `\fD` |
| $S = \{\boldsymbol{z}_i\}_{i=1}^n$ | $= \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ is sampled from $\mathcal{D}$ over $\mathcal{Z}$ | | |
| $L_S(\boldsymbol{\theta})$, $L_n(\boldsymbol{\theta})$, $\mathcal{R}(\boldsymbol{\theta})$ | $= \sum_{i=1}^n \ell(f_{\boldsymbol{\theta}}, \boldsymbol{z}_i)$ empirical risk or training loss | | |
| $L_{\mathcal{D}}(\boldsymbol{\theta})$ | $= \mathbb{E}_{\boldsymbol{z} \sim \mathcal{D}} \ell(f_{\boldsymbol{\theta}}, \boldsymbol{z})$ population risk or expected loss | | |
| $\sigma : \mathbb{R} \to \mathbb{R}^+$ | activation function | `\sigma` | |
| $\boldsymbol{w}_j$ | input weight | `\bm{w}_j` | `\vw_j` |
| $a_j$ | output weight | `a_j` | |
| $b_j$ | bias term | `b_j` | |
| $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ or $f(\boldsymbol{x}; \boldsymbol{\theta})$ | $= \sum_{j=1}^m a_j \sigma(\boldsymbol{w}_j \cdot \boldsymbol{x} + b_j)$ two-layer neural network | `f_{\bm{\theta}}` | `f_{\vtheta}` |
| $\text{VCdim}(\mathcal{H})$ | the VC-dimension of a hypothesis space $\mathcal{H}$ | | |
| $\text{Rad}(\mathcal{H} \circ S)$ | the Rademacher complexity of $\mathcal{H}$ on $S$ | | |
| GD | gradient descent | | |
| SGD | stochastic gradient descent | | |
| $B$ | a batch set | `B` | |
| $b$ | batch size | `b` | |
| $\eta$ | learning rate | `\eta` | |
| $K_n$ | Gram matrix | `K_n` | |
| $\boldsymbol{k}$ | discretized frequency | `\bm{k}` | `\vk` |
| $\boldsymbol{\xi}$ | continuous frequency | `\bm{\xi}` | `\vxi` |
| $*$ | convolution operation | `*` | |

**$L$-layer neural network**

| symbol | meaning | LaTeX | simplied |
|---|---|---|---|
| $d$ | input dimension | `d` | |
| $d_{\rm o}$ | output dimension | `d_{\rm o}` | |
| $m_l$ | the number of $l$th layer neuron, $m_0 = d$, $m_L = d_{\rm o}$ | `m_l` | |
| $\boldsymbol{W}^{[l]}$ | the $l$th layer weight | `\bm{W}^{[l]}` | `\mW^{[l]}` |
| $\boldsymbol{b}^{[l]}$ | the $l$th layer bias term | `\bm{b}^{[l]}` | `\vb^{[l]}` |
| $\circ$ | entry-wise operation | `\circ` | |
| $\sigma : \mathbb{R} \to \mathbb{R}^+$ | activation function | `\sigma` | |
| $\boldsymbol{\theta}$ | $= (\boldsymbol{W}^{[0]}, \boldsymbol{W}^{[1]}, \ldots, \boldsymbol{W}^{[L-1]}, \boldsymbol{b}^{[0]}, \boldsymbol{b}^{[1]}, \ldots, \boldsymbol{b}^{[L-1]})$, parameters | `\bm{\theta}` | `\vtheta` |
| $f_{\boldsymbol{\theta}}^{[0]}(\boldsymbol{x})$ | $= \boldsymbol{x}$ | | |
| $f_{\boldsymbol{\theta}}^{[l]}(\boldsymbol{x})$ | $= \sigma \circ (\boldsymbol{W}^{[l-1]} f_{\boldsymbol{\theta}}^{[l-1]}(\boldsymbol{x}) + \boldsymbol{b}^{[l-1]})$, $l$-th layer output | | |
| $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ | $= f_{\boldsymbol{\theta}}^{[L]}(\boldsymbol{x}) = \boldsymbol{W}^{[L-1]} f_{\boldsymbol{\theta}}^{[L-1]}(\boldsymbol{x}) + \boldsymbol{b}^{[L-1]}$, $L$-layer NN | | |