Mathematical Approach of Artificial Intelligence

Jong Won

University of Seoul, Mathematics

Contents

0 Preliminaries **3**

0.1 Notation . . . 3

1 Artificial Intelligence **4**

2 Feedforward Neural Network **5**

2.1 Structure of FNN . . . 5

2.2 Forward Propagation . . . 6

2.3 Memorization Capacity . . . 6

2.4 Universal approximation theorem . . . 7

2.5 Derivative of FNN . . . 7

2.6 Back Propagation . . . 7

2.7 Optimization . . . 7

3 Convolutional Neural Network **8**

4 Architecture **9**

4.1 Convolutional Neural Network . . . 9

4.2 Transformer . . . 9

5 Optimization **10**

5.1 Gradient Descent . . . 11

A Definition **14**

A.1 Linear Algebra . . . 14

A.2 Calculus . . . 14

A.3 Algorithm . . . 14

0 Preliminaries

0.1 Notation

We use the following notation conventions:
- Bold style $\mathbf{x}$ represents a vector.
- Normal style x represents a scalar.
- Unless otherwise stated, $\mathbf{x} \in \mathbb{R}^n$ is assumed to be a column vector.
- n-tuple $(x_1, x_2, \ldots, x_n)$ denotes that n^{th} column vector.
- Unless otherwise stated, for matrix W, its elements denote that $w_{i,j}$.
- We follow **left product convention** in linear transformation.

Definition 0.1. Define **interval** of $\mathbf{F}$:

$$[a, b]_{\mathbb{Z}} \overset{\text{def}}{=} \{n \in \mathbb{Z} \mid a \leq n \leq b\} \tag{1}$$

1 Artificial Intelligence

2 Feedforward Neural Network

In this section, I analyze the Feedforward Neural Network (FNN), a fundamental model in Artificial Intelligence.

2.1 Structure of FNN

Definition 2.1. Let an ordered tuple of natural numbers $L \overset{\text{def}}{=} (l_0, l_1, \dots, l_{n-1}, l_n)$ be given.
This tuple is referred to as the **formation of an FNN**.
The quantity $\#L$ is called the **number of layers** in the FNN, and
l_i represents the **number of nodes** in the i^{th} layer.
For each $i = 1, \dots, n-1$, let $\sigma_i : \mathbb{R} \to \mathbb{R}$ denote the **activation function** of the i^{th} layer.
Let $\mathcal{F} : \mathbb{R}^{l_n} \to \mathbb{R}^{l_n}$ denote the **output function**.
Define the **weight matrix** of the i^{th} layer for each $i = 1, \dots, n$ as:

$$W_i = \left[w^i_{j,k}\right]_{\substack{1 \le j \le l_i \\ 1 \le k \le l_{i-1}}} = \begin{bmatrix} w^i_{1,1} & \cdots & w^i_{1,l_{i-1}} \\ \vdots & \ddots & \vdots \\ w^i_{l_i,1} & \cdots & w^i_{l_i,l_{i-1}} \end{bmatrix} \in \mathbb{R}^{l_i \times l_{i-1}}.$$

Define the **bias matrix** as:

$$B_i = \begin{bmatrix} b^i_1 \\ \vdots \\ b^i_{l_i} \end{bmatrix} \in \mathbb{R}^{l_i}.$$

Now, define the **Layer** of the FNN inductively as follows:

$$\begin{cases} L_0 \overset{\text{def}}{=} X \in \mathbb{R}^{l_0}, \\ L_1 \overset{\text{def}}{=} W_1 \cdot L_0 + B_1 \in \mathbb{R}^{l_1}, \\ L_i \overset{\text{def}}{=} W_i \cdot \mathcal{M}(\sigma_i; L_{i-1}) + B_i \in \mathbb{R}^{l_i} \quad (2 \le i \le n). \end{cases}$$

With all the components defined above, the **Feedforward Neural Network** (FNN) is defined as:

$$\mathcal{N}(W_1, \dots, W_n, B_1, \dots, B_n, X) \overset{\text{def}}{=} \mathcal{F}(L_n) = Y.$$

We will denote Parameter tuple $\theta \overset{\text{def}}{=} (W_1, \dots, W_n, B_1, \dots, B_n)$. i.e., we can write

$$\mathcal{N}_\theta(X) = \mathcal{N}(W_1, \dots, W_n, B_1, \dots, B_n, X)$$

$$\underset{\text{Input}}{\underline{X = \begin{bmatrix} x_1 \\ \vdots \\ x_{l_0} \end{bmatrix}}} \overset{\substack{W_1 \cdot \\ +B_1}}{\longmapsto} \underset{1^{\text{th}}\text{layer}}{\underline{\begin{bmatrix} L_{1,1} \\ \vdots \\ L_{l_1,1} \end{bmatrix}}} \overset{\mathcal{M}(\sigma_1;\ \cdot\ )}{\longmapsto} \underset{1^{\text{th}}\text{active layer}}{\underline{\begin{bmatrix} \sigma_i(L_{1,1}) \\ \vdots \\ \sigma_i(L_{l_1,1}) \end{bmatrix}}} \overset{\substack{W_2 \cdot \\ +B_2}}{\longmapsto} \underset{2^{\text{th}}\text{layer}}{\underline{\begin{bmatrix} L_{1,2} \\ \vdots \\ L_{l_2,2} \end{bmatrix}}} \overset{\mathcal{M}(\sigma_2;\ \cdot\ )}{\longmapsto} \cdots \overset{\substack{W_n \cdot \\ +B_n}}{\longmapsto} \underset{n^{\text{th}}\text{layer}}{\underline{\begin{bmatrix} L_{1,n} \\ \vdots \\ L_{l_n,n} \end{bmatrix}}} \overset{\mathcal{F}}{\longmapsto} \underset{\text{output}}{\underline{\begin{bmatrix} Y_1 \\ \vdots \\ Y_{l_n} \end{bmatrix}}}$$

This is Basic form of Neural Network. we can determine which activation function and output function to use. Our first goal is to prove the Universal Approximation Theorem for some activation and output functions. This theorem states that for all elements in the domain, the mapping by FNN closely approximates the desired values.

2.2 Forward Propagation

In this context, from the perspective of algorithms, we analyze the time complexity of **Forward Propagation**, which means to the number of operations needed to compute the output for a given input vector X.
Suppose that formation of FNN: $L = (l_0, \dots, l_n)$ be given. And, assume that scalar sum and multiplication has $O(1)$, and each $i = 1, 2, \dots, n-1$, compute $\sigma_i(x)$ has also $O(1)$, and compute $\mathcal{F}(L_n)$ has $O(\mathbf{F}(l_n))$.

Analysis Time complexity
For input X, i.e., 0^{th}layer to transformate 1^{th}layer layer: $L_1 = W_{l_1} X + B_{l_1}$, multiplcate $l_1 \times l_0$ matrix with l_0 size vector and sum l_1 size vector, compute cost is: $l_1 \cdot l_0 + l_1$. And active each l_1 nodes, add compute cost is: $l_1 \cdot l_0 + l_1 + l_1 = l_1(l_0 + 2)$. Thus, we can write:

Proposition 1. In forward propagation with formation of FNN: $L = (l_0, \dots, l_n)$,
Time complexity of **Forward propagation** be:

$$F(L) \in O\left(\mathbf{F}(l_n) + \sum_{i=1}^{n} l_i \cdot l_{i-1}\right)$$

2.3 Memorization Capacity

In this context, we will deal **finite sample expressivity**, which means to how many distinct input datas can be mapped to the desired value.

Definition 2.2. Finite Sample Expressivity
Let formation of FNN, $L = (l_0, \dots, l_n)$ be given. Let **data** and **target** as: for $i = 1, 2, \dots, N$,

$$X_i = \begin{pmatrix} x_{1,i} \\ \vdots \\ x_{l_0,i} \end{pmatrix}, \qquad Y_i = \begin{pmatrix} y_{1,i} \\ \vdots \\ y_{l_l,i} \end{pmatrix}$$

Let **data matrix** as:

$$D \overset{\text{def}}{=} \begin{pmatrix} X_1 & X_2 & \cdots & X_N \end{pmatrix} \in \mathbb{R}^{l_0 \times N}$$

And, **target matrix** as:

$$T \overset{\text{def}}{=} \begin{pmatrix} Y_1 & Y_2 & \cdots & Y_N \end{pmatrix} \in \mathbb{R}^{l_n \times N}$$

Denote index set $I = \{1, 2, \dots, N\}$.
For FNN, $\mathcal{N}_\theta$ is said to has **Finite Sample Expressivity** if:

$$\forall D \in \mathbb{R}^{l_0 \times N}, T \in \mathbb{R}^{l_n \times N},\ \exists \theta_{\textbf{parameter}} \text{ s.t. } \forall i \in I,\ \mathcal{N}_\theta(X_i) = Y_i$$

Now, we will prove FNN can has Finite Sample Expressivity in some specific case.

Theorem 2.1. Let formation of FNN $\mathcal{N}_\theta$, $L = (l_0, l_1, l_2)$ and activation function $\sigma_i(x) = \max\{x, 0\}$ and output function $\mathcal{F}(L) = L$ be given. For number of data, $N \in \mathbb{N}$, $\mathcal{N}_\theta$ has **Finite Sample Expressivity** if:

$$l_0 l_1 l_2 \geq N$$

Proof. Let

$$\mathcal{N}_\theta(X) \overset{\text{set}}{=} W_2 \mathcal{M}(\sigma; W_1 X + B_1) + B_2$$

□

2.4 Universal approximation theorem

In this section, we discuss why AI works well. From the perspective that AI is a function, we show that for any given data, there exists a function such that it produces the desired output.

Theorem 2.2.

2.5 Derivative of FNN

Definition 2.3. Let FNN $\mathcal{N}_\theta : \mathbb{R}^{l_0} \to \mathbb{R}^{l_n}$.
Define **error function** as:

$$\mathcal{L} : \mathbb{R}^{l_n} \times \mathbb{R}^{l_n} \to \mathbb{R}^d : (\mathcal{N}_\theta(X), T) \mapsto (l_1, \dots, l_d)$$

Error function need for calculate difference between output value with target value.
For calculate derivative, Error function gives as at least differentiable once, and distant function.

Theorem 2.3. Let $\mathcal{N}_\theta(X)$ be a FNN that defined as section 2.1, and error function $\mathcal{L} : \mathbb{R}^{l_n} \times \mathbb{R}^{l_n} \to \mathbb{R}$ be given.
For input X and target T,
Derivative of $\mathcal{L}((\mathcal{N}_\theta(X), T))$ with respect to i^{th} weight be:

$$\underbrace{\frac{\partial \mathcal{L}}{\partial W_i}}_{\mathbb{R}^{l_i \times l_{i-1}}} = \underbrace{\frac{\partial \mathcal{L}}{\partial \mathcal{F}}}_{\mathbb{R}^{1 \times l_n}} \cdot \underbrace{\frac{\partial \mathcal{F}}{\partial L_n}}_{\mathbb{R}^{l_n \times l_n}} \cdot \underbrace{\frac{\partial L_n}{\partial L_{n-1}}}_{\mathbb{R}^{l_n \times l_{n-1}}} \cdots \underbrace{\frac{\partial L_{i+1}}{\partial L_i}}_{\mathbb{R}^{l_{i+1} \times l_i}} \cdot \underbrace{\frac{\partial L_i}{\partial W_i}}_{(\mathbb{R}^{l_i \times l_{i-1}})^{l_i \times 1}}$$

Derivative of $\mathcal{L}((\mathcal{N}_\theta(X), T))$ with respect to i^{th} bias be:

$$\underbrace{\frac{\partial \mathcal{L}}{\partial B_i}}_{\mathbb{R}^{1 \times l_i}} = \underbrace{\frac{\partial \mathcal{L}}{\partial \mathcal{F}}}_{\mathbb{R}^{1 \times l_n}} \cdot \underbrace{\frac{\partial \mathcal{F}}{\partial L_n}}_{\mathbb{R}^{l_n \times l_n}} \cdot \underbrace{\frac{\partial L_n}{\partial L_{n-1}}}_{\mathbb{R}^{l_n \times l_{n-1}}} \cdots \underbrace{\frac{\partial L_{i+1}}{\partial L_i}}_{\mathbb{R}^{l_{i+1} \times l_i}}$$

2.6 Back Propagation

In this, we will use Gradient Descent.

2.7 Optimization

3 Convolutional Neural Network

4 Architecture

4.1 Convolutional Neural Network

4.2 Transformer

5 Optimization

In this section, we deal with finding the conditions under which the function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies certain properties.

Definition 5.1. Let $f : \mathbb{R}^d \to \mathbb{R} \sqcup \{\infty\}$, and $\mu > 0$. Define f is **μ-strongly convex** if: $\forall x, y \in \mathbb{R}^d, t \in [0, 1]$,

$$\mu \frac{t(1-t)}{2} \|x - y\|^2 + f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \tag{1}$$

If $\mu = 0$, we simply say that f is convex.

Lemma 5.1.1. f is μ-strongly convex $\iff$ For some convex $g : \mathbb{R}^d \to \mathbb{R}$, $f(x) = g(x) + \frac{\mu}{2}\|x\|^2$

Lemma 5.1.2. If $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and differentiable, then, for all $x, y \in \mathbb{R}^d$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2 \tag{2}$$

Lemma 5.1.3. If $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and twice differentiable, then, for all $x \in \mathbb{R}^d$,

$$D^2 f(x) \succeq \mu I$$

Definition 5.2. Let $f : \mathbb{R}^d \to \mathbb{R}$, and $L > 0$. We say that f satisfies the **L-Hölder condition** of order α if: $\forall x, y \in \mathbb{R}^d$,

$$\|f(x) - f(y)\| \leq L\|x - y\|^\alpha$$

If $\alpha = 1$, we say that f is L-Lipschitz.

Lemma 5.2.1. In above case,

1. If $\alpha > 0$, f is continuous uniformly.
2. If $\alpha > 1$, f is constant.

Proof.
Fix $\alpha > 0$, and Let $\epsilon > 0$ be given. Set $\delta = \left(\frac{\epsilon}{L}\right)^{\frac{1}{\alpha}} > 0$, then

$$\|x - y\| < \delta \implies \|f(x) - f(y)\| \overset{\text{Hölder}}{\leq} L\|x - y\|^\alpha < \epsilon$$

Fix $\alpha > 1$, $x \in \mathbb{R}^d$, and Let $\epsilon > 0$ be given. Set $\delta = \left(\frac{\epsilon}{L}\right)^{\frac{1}{\alpha - 1}}$, then

$$\|x - y\| < \delta \implies \frac{\|f(x) - f(y)\|}{\|x - y\|} \overset{\text{Hölder}}{\leq} L\|x - y\|^{\alpha - 1} < \epsilon$$

This means $\nabla f(x) = 0$ for every $x \in \mathbb{R}^d$, that is, f is constant. □

Definition 5.3. Let $f \in \mathbf{D}^1(\mathbb{R}^d, \mathbb{R})$. f is **L-Smooth** if: $\forall x, y \in \mathbb{R}^d$,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

Lemma 5.3.1. If $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth, then,

$$\forall x, y \in \mathbb{R}^d.\ f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$$

Lemma 5.3.2. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable, and L-smooth. Then,

$$\|D^2 f\|_2 \leq L$$

Lemma 5.3.3. If f is L-smooth and $\lambda > 0$ then

$$\forall x, y \in \mathbb{R}^d,\ f(x - \lambda \nabla f(x)) - f(x) \leq -\lambda \left(1 - \frac{\lambda L}{2}\right) \|\nabla f(x)\|^2$$

And, moreover $\inf f > -\infty$, then

$$\forall x \in \mathbb{R}^d,\ \frac{1}{2L}\|\nabla f(x)\|^2 \leq f(x) - \inf f$$

Theorem 5.1. If $f : \mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth, then, for all $x, y \in \mathbb{R}^d$,

$$\begin{aligned}
&\frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle \\
&\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(y) - \nabla f(x), y - x\rangle
\end{aligned}$$

5.1 Gradient Descent

Definition 5.4. Suppose that $f \in \mathbf{D}^1(\mathbb{R}^d, \mathbb{R})$ satisfies $\operatorname{argmin} f \neq \emptyset$.
Given $x_0 \in \mathbb{R}^d$ and $\gamma_t > 0$, Define **Gradient Descent** algorithm:

$$x_{t+1} = x_t - \gamma_t \nabla f(x_t) \tag{$*$}$$

Theorem 5.2. Let $\{x_t\}$ be a sequence generated by the update rule $(*)$, where f satisfies **convexity** and **L-smooth**, and $0 < \gamma \leq \frac{1}{L}$. Then, for any $x^* \in \operatorname{argmin} f$,

$$\forall t \in \mathbb{N},\ f(x_t) - \inf f \leq \frac{\|x_0 - x^*\|^2}{2\gamma t}$$

Definition 5.5. Sum of functions Let $f : \mathbb{R}^d \to \mathbb{R}$ denote as:

$$f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad \text{where } f_i : \mathbb{R}^d \to \mathbb{R}.$$

Set $\operatorname{argmin} f \neq \emptyset$, and f_i's are bounded below.
We want to find the set:

$$\operatorname{argmin} f \overset{\text{def}}{=} \{x^* \in \mathbb{R}^d \mid \forall x \in \mathbb{R}^d,\ f(x^*) \leq f(x)\}$$

We say that **interpolation holds** if: for some a $x^* \in \mathbb{R}^d$, for all $i = 1, 2, \ldots, n$,

$$f_i(x^*) = \inf f_i$$

Lemma 5.5.1. If interpolation holds for f at $x^* \in \mathbb{R}^d$, then $x^* \in \operatorname{argmin} f$.

Observation. In general,

$$\inf_{x \in D} f(x) + \inf_{x \in D} g(x) \leq \inf_{x \in D}[f(x) + g(x)]$$

that is, in our case,

$$\frac{1}{n}\sum_{i=1}^{n} \inf f_i \leq \inf\left(\frac{1}{n}\sum_{i=1}^{n} f_i\right) = \inf f$$

Definition 5.6. From above observation, we obtain that: **function noise**

$$\Delta_f^* \overset{\text{def}}{=} \inf f - \frac{1}{n}\sum_{i=1}^{n} \inf f_i$$

Clearly, $\Delta_f^* \geq 0$.

And, define

Definition 5.7. Gradient noise as:

$$\sigma_f^* \overset{\text{def}}{=} \inf_{x^* \in \operatorname{argmin} f} \mathbb{V}[\nabla f_i(x^*)]$$

Stochastic Gradient Descent

Consider Problem Sum of functions, that is,

$$f = \frac{1}{n}\sum_{i=1}^{n} f_i$$

We want to find the value of $\operatorname{argmin} f$. To this, we use **Stochastic Gradient Descent Method**.
Let $x_0 \in \mathbb{R}^d$, and $\gamma_t > 0$ be a step size at iteration $t = 0, 1, 2, \ldots$.
Define the stochastic gradient descent algorithm updates the parameter as:

$$x_{t+1} = x_t - \gamma_t \nabla f_{i_t}(x_t), \quad i_t \sim \mathcal{U}(n)$$

Lemma 5.7.1. Let $f : \mathbb{R}^d \to \mathbb{R}$ satisfies Sum of $L_{\max}$-Smooth and Sum of Convex hold.
Then, f is $L_{\max}$-Smooth in expectiation. That is: for all $x, y \in \mathbb{R}^d$,

$$\frac{1}{2L_{\max}}\mathbb{E}[\|\nabla f_i(y) - \nabla f_i(x)\|^2] \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle$$

Proof. For each $i = 1, 2, \ldots, n$,

$$\frac{1}{2L_{\max}}\|\nabla f_i(y) - \nabla f_i(x)\|^2 \leq f_i(y) - f_i(x) - \langle \nabla f_i(x), y - x\rangle$$

Summing the above inequality for all $i = 1, 2, \ldots, n$, then we obtain:

$$\begin{aligned}
&\frac{1}{2L_{\max}}\sum_{i=1}^{n}\|\nabla f_i(y) - \nabla f_i(x)\|^2 \leq \sum_{i=1}^{n}[f_i(y) - f_i(x) - \langle \nabla f_i(x), y - x\rangle] \\
\Longrightarrow\ &\frac{1}{2L_{\max}}\sum_{i=1}^{n}\|\nabla f_i(y) - \nabla f_i(x)\|^2 \leq \left[\sum_{i=1}^{n} f_i(y) - \sum_{i=1}^{n} f_i(x) - \langle \nabla \sum_{i=1}^{n} f_i(x), y - x\rangle\right] \\
\Longrightarrow\ &\frac{1}{2L_{\max}}\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(y) - \nabla f_i(x)\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle
\end{aligned}$$

The Right side is:

$$\frac{1}{2L_{\max}} \cdot \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(y) - \nabla f_i(x)\|^2 = \frac{1}{2L_{\max}}\mathbb{E}[\|\nabla f(y) - \nabla f_i(x)\|^2]$$

□

Theorem 5.3. Let **Sum of $L_{\mathbf{max}}$-Smooth** and **Sum of Convex** hold, and $\{x_t\}_{t=0}^{\infty}$ a sequence generated by **Stochastic Gradient Descent** with $\gamma_t \in \left(0, \frac{1}{2L_{\max}}\right)$. Then,

$$\mathbb{E}\left[f(\bar{x}^t) - \inf f\right] \leq \frac{\|x^0 - x^*\|^2}{2\sum_{k=0}^{t-1}\gamma_k(1-2\gamma_k L_{\max})} + \frac{\sum_{k=0}^{t-1}\gamma_k^2}{\sum_{k=0}^{t-1}\gamma_k(1-2\gamma_k L_{\max})}\sigma_f^*,$$

where

$$\bar{x}^t \overset{\text{def}}{=} \sum_{k=0}^{t-1} p_{t,k} x^k, \quad \text{with} \quad p_{t,k} \overset{\text{def}}{=} \frac{\gamma_k(1-2\gamma_k L_{\max})}{\sum_{i=0}^{t-1}\gamma_i(1-2\gamma_i L_{\max})}.$$

Proof. Let $x^* \in \operatorname{argmin} f$, then $\sigma_f^* = \mathbb{V}[\nabla f_i(x^*)]$. And, denote $\mathbb{E}_k[\ \cdot\ ] \overset{\text{def}}{=} \mathbb{E}[\ \cdot \mid x_k]$ for simplicity. Now,

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - \gamma_{i_t} \cdot \nabla f_{i_t}(x_k) - x^*\|^2 \\
&= \|x_k - x^*\|^2 - 2\gamma_{i_t} \cdot \langle x_k - x^*, \nabla f_{i_t}(x_k)\rangle + {\gamma_k}^2 \cdot \|\nabla f_{i_t}(x_k)\|^2
\end{aligned}$$

Taking expectation:

$$\begin{aligned}
\mathbb{E}\left[\|x_{k+1} - x^*\|^2\right] &= \|x_k - x^*\|^2 - 2\gamma_{i_t} \cdot \langle x_k - x^*, \mathbb{E}[\nabla f_{i_t}(x_k)]\rangle + {\gamma_k}^2 \cdot \mathbb{E}\left[\|\nabla f_{i_t}(x_k)\|^2\right] \\
&= \|x_k - x^*\|^2 - 2\gamma_{i_t} \cdot \underbrace{\langle x_k - x^*, \nabla f(x_k)\rangle}_{\leq f(x_k) - f(x^*)} + {\gamma_k}^2 \cdot \underbrace{\mathbb{E}\left[\|\nabla f_{i_t}(x_k)\|^2\right]}_{\leq 4L_{\max}(f(x) - \inf f) + 2\sigma_f^*} \\
&\leq \|x_k - x^*\|^2 + 2\gamma_k(2\gamma_k L_{\max} - 1)(f(x_k) - \inf f) + 2\gamma_k^2\sigma_f^*
\end{aligned}$$

$\square$

Rearranging and taking expectation:

$$2\gamma_k(1-2\gamma_k L_{\max})\mathbb{E}\left[f(x_k) - \inf f\right] \leq \mathbb{E}\left[\ \|x_k - x^*\|^2\right] - \mathbb{E}\left[\|x_{k+1} - x^*\|^2\right] + 2\gamma_k^2\sigma_f^*$$

Summing $k = 1, 2, \dots, t-1$:

$$\begin{aligned}
2\sum_{k=0}^{t-1}\gamma_k(1-2\gamma_k L_{\max})\mathbb{E}\left[f(x_k) - \inf f\right] &\leq \|x_0 - x^*\|^2 - \mathbb{E}\left[\|x_t - x^*\|^2\right] + 2\sigma_f^*\sum_{k=0}^{t-1}\gamma_k^2 \\
&\leq \|x_0 - x^*\|^2 + 2\sigma_f^*\sum_{k=0}^{t-1}\gamma_k^2
\end{aligned}$$

Dividing both side by $2\sum_{i=0}^{t-1}\gamma_i(1-2\gamma_i L_{\max})$:

$$\begin{aligned}
\sum_{k=0}^{t-1}\gamma_k(1-2\gamma_k L_{\max})\mathbb{E}\left[f(x_k) - \inf f\right] &= \sum_{k=0}^{t-1}\mathbb{E}\left[\frac{\gamma_k(1-2\gamma_k L_{\max})}{\sum_{i=0}^{t-1}\gamma_i(1-2\gamma_i L_{\max})}(f(x_k) - \inf f)\right] \\
&\leq \frac{\|x_0 - x^*\|^2}{2\sum_{i=0}^{t-1}\gamma_i(1-2\gamma_i L_{\max})} + \frac{\sigma_f^*\sum_{k=0}^{t-1}\gamma_k^2}{\sum_{i=0}^{t-1}\gamma_i(1-2\gamma_i L_{\max})} \qquad (*)
\end{aligned}$$

Meanwhile,

$$\begin{aligned}
\sum_{k=0}^{t-1}\mathbb{E}\left[\frac{\gamma_k(1-2\gamma_k L_{\max})}{\sum_{i=0}^{t-1}\gamma_i(1-2\gamma_i L_{\max})}(f(x_k) - \inf f)\right] &= \mathbb{E}\left[\sum_{k=0}^{t-1}\frac{\gamma_k(1-2\gamma_k L_{\max})}{\sum_{i=0}^{t-1}\gamma_i(1-2\gamma_i L_{\max})} \cdot f(x_k) - \inf f\right] \\
&\overset{\text{convex}}{\geq} \mathbb{E}\left[f\left(\sum_{k=0}^{t-1}\frac{\gamma_k(1-2\gamma_k L_{\max})}{\sum_{i=0}^{t-1}\gamma_i(1-2\gamma_i L_{\max})} \cdot x_k\right) - \inf f\right] \qquad (**)
\end{aligned}$$

Consequently, combining $(*)$ and $(**)$, we obtain that:

$$\mathbb{E}\left[f\left(\sum_{k=0}^{t-1}\frac{\gamma_k(1-2\gamma_k L_{\max})}{\sum_{i=0}^{t-1}\gamma_i(1-2\gamma_i L_{\max})} \cdot x_k\right) - \inf f\right] \leq \frac{\|x_0 - x^*\|^2}{2\sum_{i=0}^{t-1}\gamma_i(1-2\gamma_i L_{\max})} + \frac{\sigma_f^*\sum_{k=0}^{t-1}\gamma_k^2}{\sum_{i=0}^{t-1}\gamma_i(1-2\gamma_i L_{\max})}$$

A Definition

A.1 Linear Algebra

Definition A.1. Matrix multiplcation of $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{m\times l}$ is defined as:

$$A \cdot B \overset{\text{def}}{=} \left[\left(\sum_{k=1}^{m} a_{i,k}b_{k,j}\right)_{i,j}\right]_{\substack{1\leq i\leq n\\ 1\leq j\leq l}} \in \mathbb{R}^{n\times l}$$

A.2 Calculus

Definition A.2. Let $f : A(\subset \mathbb{R}^n) \to \mathbb{R}$ be a differentiable in A. **Gradient** of f define as:

$$\frac{\partial f}{\partial X} \overset{\text{def}}{=} \begin{pmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{pmatrix}$$

In this case, we denote that $\nabla f = \dfrac{\partial f}{\partial X}$.

Definition A.3. Let $\mathbf{f} : A(\subset \mathbb{R}^n) \to \mathbb{R}^m$ be a differentiable in A. **Jacobian** of $\mathbf{f}$ define as:

$$\frac{\partial \mathbf{f}}{\partial X} \overset{\text{def}}{=} \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{m\times n}$$

Definition A.4. Let $\mathbf{F} : A(\subset \mathbb{R}^{n\times m}) \to \mathbb{R}^{k\times l}$ be a differentiable in A. **Jacobian** of $\mathbf{F}$ define as:

$$\frac{\partial \mathbf{F}}{\partial X} \overset{\text{def}}{=} \begin{pmatrix} \frac{\partial \mathbf{F}^{[1]}}{\partial X_{[1]}} & \cdots & \frac{\partial \mathbf{F}^{[l]}}{\partial X_{[1]}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{F}^{[1]}}{\partial X_{[n]}} & \cdots & \frac{\partial \mathbf{F}^{[l]}}{\partial X_{[n]}} \end{pmatrix} \in \mathbb{B}^{n\times l},\ \mathbb{B} \in \mathbb{R}^{k\times m}$$

Later

Definition A.5. Let $f : \mathbb{R} \to \mathbb{R}$ be a scalar function, and W be a real matrix in $\mathbb{R}^{n\times m}$.
Define **elementwise transformation** of f, $\mathcal{M} : \mathcal{F}(\mathbb{R}) \times \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}$ as:

$$\mathcal{M}(f; W) \overset{\text{def}}{=} \begin{pmatrix} f(w_{1,1}) & \cdots & f(w_{1,m}) \\ \vdots & \ddots & \vdots \\ f(w_{n,1}) & \cdots & f(w_{n,m}) \end{pmatrix}$$

A.3 Algorithm

Definition A.6. Time complexity:

1. $\mathbf{O}(g(n)) \overset{\text{def}}{=} \{f(n) \mid \exists c, n_0 > 0 \text{ s.t. } \forall n \geq n_0,\ 0 \leq f(n) \leq cg(n)\}$ (Upper bound)

2. $\Omega(g(n)) \overset{\text{def}}{=} \{f(n) \mid \exists c, n_0 > 0 \text{ s.t. } \forall n \geq n_0,\ 0 \leq cg(n) \leq f(n)\}$ (Lower bound)

3. $\mathbf{o}(g(n)) \overset{\text{def}}{=} \{f(n) \mid \forall c > 0,\ \exists n_0 > 0 \text{ s.t. } \forall n \geq n_0,\ 0 \leq f(n) < cg(n)\}$

4. $\omega(g(n)) \overset{\text{def}}{=} \{f(n) \mid \forall c > 0,\ \exists n_0 > 0 \text{ s.t. } \forall n \geq n_0,\ 0 \leq cg(n) < f(n)\}$

5. $\Theta(g(n)) \overset{\text{def}}{=} \mathbf{O}(g(n)) \cap \Omega(g(n))$

Probability

Definition A.7. Let X be a set. A subset $\Sigma \subset \mathcal{P}(X)$ is called a **σ-algebra** if:

S1. $X \in \Sigma$.

S2. For any countable union of sets in Σ, $\displaystyle\bigcup_{i=1}^{\infty} A_i \in \Sigma$ where $A_i \in \Sigma$.

S3. For any $A \in \Sigma$, $X \setminus A \in \Sigma$.

Definition A.8. A **probability space** is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where:

- Ω is the sample space, the set of all possible outcomes.
- $\mathcal{F} \subseteq 2^{\Omega}$ is a σ-algebra over Ω.
- $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a probability measure satisfying:

  PS1. $\mathbb{P}(\Omega) = 1$.

  PS2. For any countable collection $\{A_i\}$ of disjoint sets in $\mathcal{F}$,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

Definition A.9. A **random variable** is a measurable function $X : \Omega \to \mathbb{R}$ such that for every Borel set $B \subseteq \mathbb{R}$, the pre-image $X^{-1}(B) \in \mathcal{F}$.

Definition A.10. Let $X : \Omega \to \mathbb{R}$ be an integrable random variable. The **expectation** of X, denoted $\mathbb{E}[X]$, is defined by:

$$\mathbb{E}[X] = \int_{\Omega} X(\omega)\, d\mathbb{P}(\omega)$$

This is a **Lebesgue integral** of X with respect to the measure $\mathbb{P}$.

Definition A.11. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $f : \Omega \to [0, \infty]$ be a measurable function. The **Lebesgue integral** of f over a set $A \in \mathcal{F}$ is defined as:

- First, define the integral for simple functions: if $f = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i}$, where $a_i \geq 0$ and $A_i \in \mathcal{F}$ are disjoint, then

$$\int_A f\, d\mu = \sum_{i=1}^{n} a_i \mu(A_i \cap A)$$

- For a general non-negative measurable function f, define

$$\int_A f\, d\mu = \sup\left\{\int_A s\, d\mu : 0 \leq s \leq f,\ s \text{ simple}\right\}$$

- For an arbitrary measurable function f, write $f = f^+ - f^-$, and define

$$\int_A f\, d\mu = \int_A f^+\, d\mu - \int_A f^-\, d\mu$$

  whenever both terms on the right-hand side are finite.

Definition A.12. Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra. The **conditional expectation** of X given $\mathcal{G}$, denoted $\mathbb{E}[X \mid \mathcal{G}]$, is the unique (up to a.s. equality) $\mathcal{G}$-measurable function Y such that:

$$\forall G \in \mathcal{G}, \quad \int_G Y \, d\mathbb{P} = \int_G X \, d\mathbb{P}$$

If $\mathcal{G} = \sigma(Z)$ for some random variable Z, we often write:

$$\mathbb{E}[X \mid Z] := \mathbb{E}[X \mid \sigma(Z)]$$

.