Final Review - ch4

Sunday, 6 December 2020 3:08 PM

Chapter 4: Linear functions and differentiation

Section 4.1: Linear Functions

Let $f: V \to \mathbb{R}$ be a function on a vector space V.
f is a linear function if
$$f(\alpha x+\beta y) = \alpha f(x) + \beta f(y) \quad \forall\, \alpha,\beta \in \mathbb{R} \text{ and } x,y \in V.$$

Example for linear function:
Mean function, $a^T x$, $F(f)=f(x)$, $F(f) = \int_a^b f(x)\,dx$, $F: L^2(a,b) \to \mathbb{R}$,
$f(x)=\langle x,z\rangle$, $z \in V$

Example for non-linear function:
$f(x)=\max\{x_1, x_2, \cdots, x_n\}$, $g(x)=\|x\|$ (norm)

Properties of linear functions

Properties of linear functions
- Homogeneity: $f(\alpha x) = \alpha f(x) \quad \forall\, \alpha \in \mathbb{R}$ and $x \in V$.
  (Because $f(\alpha x) = f(\alpha x + 0x) = \alpha f(x) + 0\, f(x) = \alpha f(x)$)
  It implies $f(0) = 0$, because $f(0) = f(0 \cdot x) = 0 \cdot f(x) = 0 \quad \forall\, x \in V$.
- Additivity: $f(x+y) = f(x) + f(y) \quad \forall\, x, y \in V$.
- $f(\alpha_1 x_1 + \cdots + \alpha_k x_k) = \alpha_1 f(x_1) + \cdots + \alpha_k f(x_k), \quad \forall \alpha_1, \cdots, \alpha_k \in \mathbb{R},\ x_1, \cdots, x_k \in V$.
  To see this, we note that
$$\begin{aligned} f(\alpha_1 x_1 + \cdots + \alpha_k x_k) &= \alpha_1 f(x_1) + f(\alpha_2 x_2 + \cdots + \alpha_k x_k) \\ &= \alpha_1 f(x_1) + \alpha_2 f(x_2) + f(\alpha_3 x_3 + \cdots + \alpha_k x_k) \\ &\ \ \vdots \\ &= \alpha_1 f(x_1) + \cdots + \alpha_k f(x_k). \end{aligned}$$

Riesz representation theorm - Preliminary:

a linear function $f: \mathbb{R}^n \to \mathbb{R}$ if and only if $f(x) = \langle a, x\rangle$ for some unique $a \in \mathbb{R}^n$

Pf14: any linear function can be written as inner product
Pf15: the inner product is unique

Riesz representation theorm - extend to hilbert space:

Let H be a Hilbert space, and f be a function: $H \to \mathbb{R}$. Then
f is linear and bounded if and only if $f(x) = \langle a, x\rangle$ for some unique $a \in H$.

Let H be a Hilbert space, and f be a function: $H \to \mathbb{K}$. Then
f is linear and bounded if and only if $f(x) = \langle a, x \rangle$ for some unique $a \in H$.

Examples for riesz representation theorem:

1. Mean(x) = <a,x> where a = $(1/n, 1/n, \ldots, 1/n)^T$
2. $F(f) = \int_a^b f(t)dt = \langle f, h \rangle$, where $h(t) = 1 \quad \forall t \in [a,b]$

Examples for not a riesz representation theorem:

1. Norm (since norm is non linear)
2. $F : C[a,b] \to \mathbb{R}$ defined by $F(f) = f(t_0)$ for a fixed $t_0 \in \mathbb{R}$.
   F is linear on C[a,b]

Pf16: a fixed t0 function cannot have h(t) in <f,h>

Hyperplanes

The set $\{x \in V \mid \langle a, x \rangle = b\}$, where $a \in V$ and $b \in \mathbb{R}$ are given, is called a Hyperplane in V.

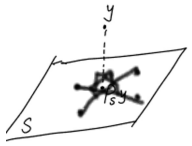
Pf17: $S_{a,0} = \{x \in V \mid \langle a, x \rangle = 0\}$ is a hyperplane.

Pf18: $S_{a,b} = \{x \in V \mid \langle a, x \rangle = b\}$ is also a hyperplane.

Projection onto hyperplanes

The vector on S that is the closest to y is called the projection of y on S, denoted by $P_S y$, i.e., $P_S y = \arg\min_{x \in S} \|x - y\|$.

Theorem: z is a solution of $\min_{x \in S} \|x - y\|$ if and only if $z \in S$ and $\langle z - y, z - x \rangle = 0 \quad \forall x \in S$.



← solution of $P_S y$ should $\perp$ to all $x \in S$.

($(y - P_S y)$ vector should be orthogonal to x)

Pf 19: proof of the above theorem

Theorem: The solution of $\min_{x \in S} \|x - y\|^2$ exists and unique, which is given by $y - \left(\frac{\langle a, y \rangle - b}{\|a\|^2}\right) a$

Pf20: $P_S y$ should be the above formula

- In summary, the projection $P_S y$ of $y \in V$ onto the hyperplane $S = \{x \in V \mid \langle a, x \rangle = b\}$ exists and is unique. Furthermore, $P_S y = y - \left(\frac{\langle a, y \rangle - b}{\|a\|^2}\right) a$ and it satisfies $\langle P_S y - y, P_S y - x \rangle = 0$.

Affine functions:

Affine functions

A linear function plus a constant is called an affine function. That is, a function $f: V \to \mathbb{R}$ is affine if $f(x) = g(x) + b$, where $g: V \to \mathbb{R}$ is linear and $b \in \mathbb{R}$ is a constant.

Properties of affine functions:

Properties:

- If $f: V \to \mathbb{R}$ is affine, then linearity

Properties:

- If $f: V \to \mathbb{R}$ is affine, then linearity
  $f(\alpha x + \beta y) = \alpha f(x) + \beta f(y) \quad \forall x, y \in V$ and $\alpha, \beta \in \mathbb{R}$ s.t. $\alpha + \beta = 1$.

  To see this, (g linear, $\alpha+\beta=1$)
  $$f(\alpha x+\beta y) = g(\alpha x+\beta y) + b = \alpha g(x) + \beta g(y) + (\alpha+\beta) b$$
  $$= \alpha(g(x)+b) + \beta(g(y)+b) = \alpha f(x) + \beta f(y).$$
- If $f: V \to \mathbb{R}$, where V is a Hilbert space, then
  f must be in the form of Can be written as inner product
  $$f(x) = \langle a, x \rangle + b, \quad \text{where } a \in V \text{ and } b \in \mathbb{R}.$$

Section 4.2: Case Studies: Regression and classification

Section 4.2.1: Linear Regression:

- Mathematically, we need to find a function $f: \mathbb{R}^n \to \mathbb{R}$ such that
  $$f(x_i) \simeq y_i, \quad i = 1, 2, \cdots, N$$

Linear model:

We search f in the class of all affine functions,
i.e., $f(x) = \langle a, x \rangle + b$ for some $a \in \mathbb{R}^n$, $b \in \mathbb{R}$.

Therefore, we find $a \in \mathbb{R}^n$, $b \in \mathbb{R}$ by solving
$$\min_{\substack{a \in \mathbb{R}^n, \\ b \in \mathbb{R}}} \sum_{i=1}^{N} \left(\langle a, x_i \rangle + b - y_i \right)^2$$

再轉做下面的形式:

Write
$$X = \begin{bmatrix} x_1^T & 1 \\ x_2^T & 1 \\ \vdots & \vdots \\ x_N^T & 1 \end{bmatrix} \in \mathbb{R}^{N \times (n+1)}, \qquad \beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1}$$
and
$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N.$$

Then LS problem becomes
$$\min_{\beta \in \mathbb{R}^{n+1}} \| X\beta - y \|_2^2.$$

Since we have N linear equations to fit and $n+1$ unknowns,
$N \geq n+1$ is required to have a unique solution.

Regularization:

好多時候我地都不會有足夠的data, 即是N<n+1

所以我們不會在R^{n+1}裏面找beta,而是在一個subset裏面找

而這個subset的選定就有下列不同選擇:

1. Ridge regression

We choose $S = \{\beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1} \mid \|a\|_2 \le C\}$ for some $C > 0$.

The we solve $\min_{\beta \in S} \|X\beta - y\|_2^2$

$\Updownarrow$ ← by convex optimization theory

(Lagrangian)

$\min_{\beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1}} \frac{1}{2}\|X\beta - y\|_2^2 + \lambda \|a\|_2^2$,

where $\lambda > 0$ depends on C and others.

Here $\|a\|_2^2$ is the regularization term.

and $\frac{1}{2}\|X\beta - y\|_2^2$ is the data-fitting term.

2. LASSO regression except above $||a||_2$ is replaced by $||a||_1$

Small $\|a\|_1$ tends to give a sparse vector $a \in \mathbb{R}^n$

(i.e., many entries of a are zeros)

Consequently, LASSO gives a solution $\beta = \begin{bmatrix} a \\ b \end{bmatrix}$ such that

$X\beta \approx y$ and a is sparse

Thus, given $x \in \mathbb{R}^n$, the prediction is given by

$\langle x, a \rangle + b = \sum_{i=1}^{n} a_i x_i + b = \sum_{i \in S} a_i x_i + b$

Let $I = \{i \mid a_i \neq 0\}$

Since I is a small set, the prediction depends only on a small portion entries of x.

This is preferred because the prediction is interpretable.

Section 4.2.2 Kernel Regression

We solve

$$\min_{a \in H} \frac{1}{2} \sum_{i=1}^{N} \left(\langle a, \phi(x_i) \rangle - y_i\right)^2 + \lambda \|a\|_H^2$$

Representer Theorem:

The solution must be in the form of $a = \sum_{i=1}^{N} c_i \phi(x_i)$ for some $c = \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \in \mathbb{R}^N$.

Pf21: proof of representer theorem

After some operation, we have the following result

Thus, the solution of the original minimization is

$$a = \sum_{i=1}^{N} c_i \phi(x_i)$$

where $c \in \mathbb{R}^N$ is a solution of $\min_{c \in \mathbb{R}^N} F_1(c)$.

$$F_1(c) = \frac{1}{2}\|Kc - y\|_2^2 + \lambda c^T K c$$

上述Function 等於下面function

$$\min_{c \in \mathbb{R}^N} \sum_{i=1}^{N} \left(\sum_{j=1}^{N} c_j K(x_i, x_j) - y_i \right)^2 + \lambda \sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j K(x_i, x_j)$$

等於先solve c再solve a, 但係minimize兩個問題都係等價的

Then the predicted output y for input $x \in \mathbb{R}^n$ is

$$\forall x \in \mathbb{R}^n, \quad f(\phi(x)) = y = \langle a, \phi(x) \rangle = \left\langle \sum_{i=1}^{N} c_i \phi(x_i), \phi(x) \right\rangle = \sum_{i=1}^{N} c_i K(x_i, x)$$

Summary: 整個kernel ridge regression algorithm等於

Kerel ridge regression.

Input: (x_i, y_i), $i = 1, \cdots N$, $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$

Algorithm:

① Choose a kernel function $K(\cdot, \cdot)$,

e.g., $K(x, z) = e^{-\frac{\|x - z\|_2^2}{\sigma^2}}$, where $\sigma > 0$

② Solve $c \in \mathbb{R}^N$ by

$$\min_{c \in \mathbb{R}^N} \sum_{i=1}^{N} \left(\sum_{j=1}^{N} c_j K(x_i, x_j) - y_i \right)^2 + \lambda \sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j K(x_i, x_j)$$

Prediction: Given a new $x \in \mathbb{R}^n$, the predicted response is

$$\sum_{i=1}^{N} c_i K(x_i, x)$$

Pf22: prove how to change objective function to inner product only in kernel regression problem

Section 4.2.3 Classification

Problem Definition

- Classification: Giving training data

$$(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N), \quad x_i \in \mathbb{R}^n, \ y_i \in \{-1, +1\}, \ i = 1, \cdots, N,$$

find a classifier (a function) f such that

$$y_i = \begin{cases} 1, & \text{if } f(x_i) \geq 1 \\ -1, & \text{if } f(x_i) \leq -1 \end{cases}.$$

We use hyperplanes to separate the points

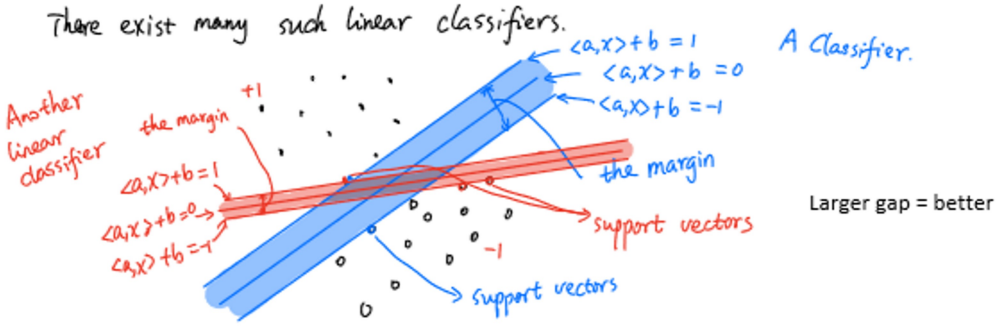
$$f(x) = \langle a, x \rangle + b, \quad \text{where } a \in \mathbb{R}^n, \ b \in \mathbb{R}.$$

The weights $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ are noralized such that

$$\langle a, x_i \rangle + b_i \begin{cases} \geq 1 & \text{if } y_i = +1 \\ \leq -1 & \text{if } y_i = -1 \end{cases}$$

- Support Vector Machine (SVM)

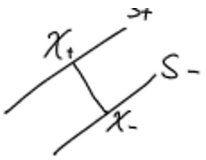
There exist many such linear classifiers.



Larger gap = better

SVM-1:

Let us calculate the margin in terms of a and b.



The margin is the distance between the two hyperplanes

$S_+ = \{x \mid \langle a, x \rangle + b = 1\}$ and $S_- = \{x \mid \langle a, x \rangle + b = -1\}$

Let $x_+ \in S_+$ and $x_- \in S_-$ such that

$$\|x_+ - x_-\|_2 = \text{dist}(S_+, S_-).$$

個margin嘅寬度就係呢個

經過推論之後

$$\min_{a \in \mathbb{R}^n, \, b \in \mathbb{R}} \ \frac{1}{2}\|a\|_2^2 \qquad \text{(SVM-1)}$$
$$\text{s.t} \quad y_i(\langle a, x_i \rangle + b) \geq 1$$

Pf22: SVM的推論過程

Soft SVM:

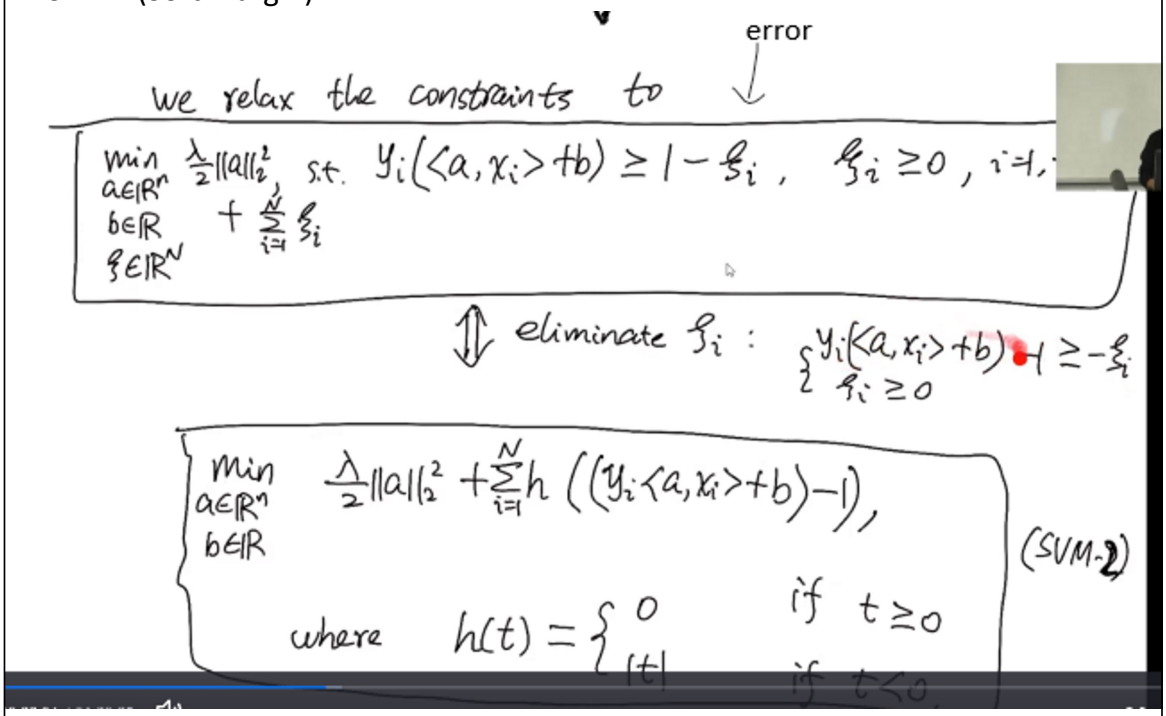
We consider a "soft" version of (SVM-1) as follows.

We minimize the error to the separation

$$\sum_{i=1}^{N} h\left(\left(y_i \langle a, x_i \rangle + b\right) - 1\right),$$

where the function $h(t) = \begin{cases} 0 & \text{if } t \geq 0 \\ |t| & \text{if } t < 0 \end{cases}$

SVM-2(Soft margin)



Pf24: SVM-2邏輯推論

Kernel SVM Problem

Therefore, (SVM-2) becomes

$$\min_{a \in H} h\left(y_i \langle a, \phi(x_i)\rangle - 1\right) + \frac{\lambda}{2}\|a\|_H^2 \qquad \text{(K-SVM)}$$

經過一陣推論後，(主要都係將a拆開a_s + sum{c_i phi(x_i)})　Pf25: Kernel SVM Algo推論

Thus, (K-SVM) becomes

$$\min_{c \in \mathbb{R}^N} h\left(y_i\left(\sum_{j=1}^{N} K(x_i, x_j)\, c_j\right) - 1\right) + \frac{\lambda}{2} c^T K c,$$

where $K = [K(x_i, x_j)]_{i=1,j=1}^{N,N} \in \mathbb{R}^{N \times N}$.

If we define $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j)\rangle$ for some $K: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$.

The prediction of the input x is given by

$\operatorname{sgn}\left(\sum_{j=1}^{N} K(x, x_j) c_j\right)$ =》就係用呢個黎predict,睇佢大過1定細過1

Kernel SVM Algo Summary:

Kernel SVM:

① Choose a kernel function $K: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$
(e.g., $K(x, z) = e^{-\frac{\|x-z\|_2^2}{\sigma^2}}$)

② Solve c from

$$\min_{c \in \mathbb{R}^N} \frac{\lambda}{2} c^T K c + \sum_{i=1}^{N} h\left(y_i \cdot (Kc)_i - 1\right)$$

③ Then the classify function

Sections 4.3 Linear Approximation and differentiation

Definition of differention

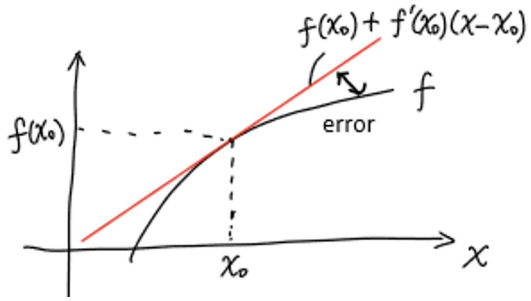
Recall that for a function $f: \mathbb{R} \to \mathbb{R}$, the derivative at x_0 is

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0},$$

which is the same as

$$\lim_{x \to x_0} \left| \frac{f(x) - f(x_0) - f'(x_0)(x - x_0)}{x - x_0} \right| = 0$$

Notice that $f(x_0) + f'(x_0)(x - x_0)$ is an affine function in $\mathbb{R}$ that passes through $(x_0, f(x_0))$.



Differentation 定義的轉換：

用f(x0) + f`(x0)(x-x0) 去approximate f(x), 同埋從differentation個定義來看，可以得知

$$\lim_{x \to x_0} \left| \frac{\text{error}}{x - x_0} \right| \to 0 = \text{little o of } (|x - x_0|)$$

Using this differentiation to extend to hilbert space, we can define Frechet differentation

Definition: Let V be a Hilbert space. Let $f: V \to \mathbb{R}$. Then f is said Frechet differentiable if there exists a $v \in V$ such that

$$\lim_{x \to x^{(0)}} \frac{|f(x) - f(x^{(0)}) - \langle v, x - x^{(0)} \rangle|}{\|x - x^{(0)}\|} = 0.$$

(Note that $x \to x^{(0)}$ is the same as $\|x - x^{(0)}\| \to 0$)

If f is differentiable at $x^{(0)}$, v is called the gradient of f at $x^{(0)}$, denoted by $\nabla f(x^{(0)})$.

Example of how to calculate gradient using definition

Example 1: $f(x) = \|x\|^2$, where $\|x\|$ is the norm on V.

At any $x^{(0)} \in V$,

$$\|x\|^2 = \|(x-x^{(0)})+x^{(0)}\|^2 = \langle (x-x^{(0)})+x^{(0)}, (x-x^{(0)})+x^{(0)} \rangle$$
$$= \|x-x^{(0)}\|^2 + \|x^{(0)}\|^2 + 2\langle x^{(0)}, x-x^{(0)} \rangle$$

Therefore, $\|x\|^2 - \underbrace{\left(\|x^{(0)}\|^2 + 2\langle x^{(0)}, x-x^{(0)} \rangle\right)}_{\text{affine approximation}} = \|x-x^{(0)}\|^2$

So $$\lim_{x \to x^{(0)}} \frac{\left|\|x\|^2 - \|x^{(0)}\|^2 - \langle 2x^{(0)}, x-x^{(0)} \rangle\right|}{\|x-x^{(0)}\|} = \lim_{x \to x_0} \frac{\|x-x^{(0)}\|^k}{\|x-x^{(0)}\|} = 0.$$

Thus, $\nabla f(x^{(0)}) = 2x^{(0)}$.

目標就是要搬到<grad(x) , x-x$^{(0)}$>這個形式

Properties: For frechet

① Frechet differentiation is linear in f, i.e.,

$$\nabla(\alpha f + \beta g)(x) = \alpha \nabla f(x) + \beta \nabla g(x).$$

② Chain rule: Let $f: V \to \mathbb{R}$ and $g: \mathbb{R} \to \mathbb{R}$. Then $g \circ f: V \to \mathbb{R}$ and

$$\nabla(g \circ f)(x) = g'(f(x)) \cdot \nabla f(x)$$

if f and g are differentiable at x and $f(x)$ respectively.

Taylor expression

Taylor's expansion

From the definition, we see that

$$f(x) \approx f(x^{(0)}) + \langle \nabla f(x^{(0)}), x-x^{(0)} \rangle$$

Or, more precisely,

$$f(x) = f(x^{(0)}) + \langle \nabla f(x^{(0)}), x-x^{(0)} \rangle + o(\|x-x^{(0)}\|)$$

$$f(x) = f(x^{(0)}) + \langle \nabla f(x^{(0)}), x - x^{(0)} \rangle + o(\|x - x^{(0)}\|)$$

This is a generalization of Taylor's expansion.

In paticular, if $f : \mathbb{R}^n \to \mathbb{R}$

$$f(x) \simeq f(x^{(0)}) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^{(0)}) \, (x_i - x_i^{(0)}) \quad .$$

Differentiation on normed vector spaces

Since no riesz representation on normed spaces, only use linear function L in differentiation

Differentiation on normed vector spaces

Let V be a normed vector space.

Let $f: V \to \mathbb{R}$ be a function. Let $x^{(0)} \in V$.

To define differentiation, we still use an affine function approximation, and the affine function passes thru $(x^{(0)}, f(x^{(0)}))$.

Thus, we use $f(x^{(0)}) + L(x - x^{(0)})$, where $L: V \to \mathbb{R}$ is a linear function, to approximate $f(x)$.

However, because there is no Riesz representation on normed spaces, we keep the linear function L in the definition of differentiation.

Definition: f is differentiable at $x^{(0)} \in V$, if:

$\exists$ a linear function $L: V \to \mathbb{R}$ such that

$$\lim_{\|x - x^{(0)}\| \to 0} \frac{|f(x) - (f(x^{(0)}) + L(x - x^{(0)}))|}{\|x - x^{(0)}\|} = 0 .$$

The linear function L is called the differentiation of f at $x^{(0)}$.