Machine Learning

Daniel Drake

May 21, 2020

1 Change

Definition 1.1. ***Metric***
Let X be a non-empty set..
Let $d : X \times X \to \mathbb{R}_0^+$ such that:

- $(\forall x \in X) d(x, x) = 0$
- $(\forall x, y \in X) d(x, y) = 0 \Leftrightarrow x = y$
- $(\forall x, y \in X) d(x, y) = d(y, x)$
- $(\forall x, y, z \in X) d(x, z) \leq d(x, y) + d(y, z)$

Then d is called a metric and (X, d) is called a metric space.
Reference

Definition 1.2. ***Limit of a function***
Let $T : X \to Y$ where (X, d_X) and (Y, d_Y) are metric spaces.
Then fix $x_0 \in X$.
If:

$$(\exists L \in Y)(\forall \epsilon > 0)(\exists \delta > 0)(\forall x \in X)(d(x, x_0) < \delta \Rightarrow d(f(x), L) < \epsilon)$$

Then:

$$\lim_{x \to x_0} f(x) = L$$

Reference

Definition 1.3. ***Derivative***
Let $\hat{f} : \mathbb{R} \to \mathbb{R}$
Further let $f = \hat{f}|_U$ where $U \in \tau_{\mathbb{R}}$
Then f is said to be differentiable at $x \in U$ if there exists an L_x such that:

$$L_x = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}$$

If L_x exists for all $x \in U$ then we write:

$$\frac{d}{dx} f(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}$$

Reference

Theorem 1.1. ***Fundamental increment lemma***
Let f be described as above and be differentiable at x.
Then there exists a function $\phi : \mathbb{R} \to \mathbb{R}$ such that:

$$f(x+h) = f(x) + \frac{d}{dx}f(x)h + \phi(x)h$$

and

$$\lim_{h\to 0} \phi(h) = 0$$

Proof:
Define: $\phi(h) = \frac{f(x+h)-f(x)}{h} - \frac{d}{dx}f(x)$
Then: $\phi(h)h = f(x+h) - f(x) - \frac{d}{dx}f(x)h$
Then: $\phi(h)h + f(x) - \frac{d}{dx}f(x)h = f(x+h)$
And so we have property 1.
Next:

$$\lim_{h\to 0} \phi(h) = \lim_{h\to 0}\left[\frac{f(x+h) - f(x) - \frac{d}{dx}f(x)h}{h}\right] = \lim_{h\to 0}\left[\frac{f(x+h)-f(x)}{h} - \frac{d}{dx}f(x)\right]$$
$$= \lim_{h\to 0}\frac{f(x+h)-f(x)}{h} - \lim_{h\to 0}\frac{d}{dx}f(x) = \frac{d}{dx}f(x) - \frac{d}{dx}f(x) = 0$$

Definition 1.4. ***Partial Derivative***
Let $\hat{f} : \mathbb{R}^n \to \mathbb{R}$
Further let $f = \hat{f}|_U$ where $U \in \tau_{\mathbb{R}^n}$
Then f is said to be differentiable at $x \in U$ with respect to the i'th component of x if there exists an L_{x_i} such that:

$$L_{x_i} = \lim_{h\to 0}\frac{f(x_1,...,x_i+h,...,x_n) - f(x_1,...,x_i,...,x_n)}{h}$$

If L_{x_i} exists for all $x \in U$ then we write:

$$\frac{\partial}{\partial x_i}f(x) = \lim_{h\to 0}\frac{f(x_1,...,x_i+h,...,x_n) - f(x_1,...,x_i,...,x_n)}{h}$$

Reference

Theorem 1.2. ***Equivalent characterization***
Let $\hat{f} : \mathbb{R}^n \to \mathbb{R}$
Further let $f = \hat{f}|_U$ where $U \in \tau_{\mathbb{R}^n}$
And let f be differentiable at $x \in U$ with respect to the i'th component of x, then:

$$L_{x_i} = \lim_{h\to 0}\frac{f(x_1,...,x_i+h,...,x_n) - f(x_1,...,x_i,...,x_n)}{h}$$

$$\Leftrightarrow 0 = \lim_{h\to 0}\left[\frac{f(x_1,...,x_i+h,...,x_n) - f(x_1,...,x_i,...,x_n)}{h} - L_{x_i}\right]$$

$$\Leftrightarrow 0 = \lim_{h\to 0}\left[\frac{f(x_1,...,x_i+h,...,x_n) - f(x_1,...,x_i,...,x_n)}{h} - \frac{L_{x_i}\cdot h}{h}\right]$$

$$\Leftrightarrow 0 = \lim_{h\to 0}\left[\frac{f(x_1,...,x_i+h,...,x_n) - f(x_1,...,x_i,...,x_n) - \langle L_{x_i}, h\rangle}{h}\right]$$

Definition 1.5. ***Gradient***
Let $\hat{f}:\mathbb{R}^n \to \mathbb{R}$ *and let* $f:U \to \mathbb{R}$ *such that* $f=\hat{f}|_U$ *where* $U \in \tau_{\mathbb{R}^n}$
f *is said to be differentiable at* $x \in U$ *if* $\exists \nabla f:\mathbb{R}^n \to \mathbb{R}^n$ *such that:*

$$\lim_{h\to 0}\frac{|f(x+h)-f(x)-\langle \nabla f(x), h\rangle|}{||h||}=0$$

Theorem 1.3. ***Form of the Gradient***
Let f *be defined as above.*
Then $\nabla f:\mathbb{R}^n \to \mathbb{R}^n$ *where:*

$$\nabla f(x)=\begin{bmatrix}\frac{\partial}{\partial x_1}f(x)\\ \vdots \\ \frac{\partial}{\partial x_n}f(x)\end{bmatrix}\forall x \in \mathbb{R}^n$$

is the form of ∇f *which satisfies the above statement if* f *is differentiable.*
Reference

Proof:
Suppose ∇f *is defined as above and all the partial derivatives exist.*
Then:

$$\frac{1}{||h||}|f(x+h)-f(x)-\langle \nabla f(x), h\rangle| = \frac{1}{||h||}\left|f(x+h)-f(x)-\sum_{j=1}^{n}\frac{\partial}{\partial x_j}f(x)\cdot h_j\right|$$

Definition 1.6. ***Matrix Functional Differentiability***
Let $\hat{T} : \mathbb{R}^{n\times m} \to \mathbb{R}$ *and let* $T : U \to \mathbb{R}$ *such that* $T = \hat{T}|_U$ *where* $U \in \tau_{\mathbb{R}^{n\times m}}$
T *is said to be differentiable at* $x \in U$ *if* $\exists D : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}$ *such that:*

$$\lim_{h\to 0} \frac{|T(x+h) - T(x) - \langle DT(x), h\rangle|}{||h||} = 0$$

where $\langle \cdot, \cdot \rangle$ *is an inner product defined on* $\mathbb{R}^{n\times m}$

Definition 1.7. ***Frobenius inner product***
The Frobenius inner product is defined as:

$$\langle \cdot, \cdot \rangle_{FB} : \mathbb{R}^{n\times m} \times \mathbb{R}^{n\times m} \to \mathbb{R} \text{ such that: } \langle A, B\rangle_{FB} = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{i,j}b_{i,j} \text{ for all } A, B \in \mathbb{R}^{n\times m}$$

Theorem 1.4. ***Form of Matrix Functional Derivative***

Definition 1.8. ***Differentiability of a multi-variable function.***
Let $\hat{f} : \mathbb{R}^m \to \mathbb{R}^n$ such that:

$$\hat{f}(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix} \text{ and } (\forall j \in \mathbb{N}_n)(f_j : \mathbb{R}^m \to \mathbb{R})$$

Further let $f = \hat{f}|_U$ where $U \in \tau_{\mathbb{R}^m}$
Then f is said to be differentiable at $x \in U$ if there exists a linear operator $J_f : \mathbb{R}^m \to \mathbb{R}^n$ such that:

$$\lim_{h \to \vec{0}} \frac{||f(x+h) - f(x) + J_f(h)||_{\mathbb{R}^m}}{||h||_{\mathbb{R}^n}} = 0$$

Reference

Theorem 1.5. ***If a multi-variable function, f, is differentiable at x then the linear operator J is the Jacobian matrix.***
So our guess is that:

$$J_f = \begin{bmatrix} \frac{\partial}{\partial x_1} f_1(x) & \cdots & \frac{\partial}{\partial x_n} f_1(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} f_m(x) & \cdots & \frac{\partial}{\partial x_n} f_m(x) \end{bmatrix}$$

since this form is a linear operator mapping from the appropriate space to the appropriate space. It should be noted that the transpose of this matrix can not satisfy the definition of differentiability of a multi-variable function and so it is not the correct linear operator.

Definition 1.9. ***Matrix operator differentiability***
Let $T : \mathbb{R}^{n \times m} \to \mathbb{R}^n$ such that:

$$T(A) = \begin{bmatrix} T_1(A) \\ \vdots \\ T_n(A) \end{bmatrix} \forall A \in \mathbb{R}^{n \times m} \text{ and } (\forall j \in \mathbb{N}_n)(T_j : \mathbb{R}^{n \times m} \to \mathbb{R})$$

Then T is said to be differentiable at $A \in \mathbb{R}^{n \times m}$ if there exists a linear operator $D : \mathbb{R}^{n \times m} \to \mathbb{R}^n$ where:

$$\lim_{h \to 0} \frac{||T(A+h) - T(A) + D(h)||_{\mathbb{R}^n}}{||h||_{\mathbb{R}^{n \times m}}} = 0$$

If D exists then it is called the Matrix operator derivative and is written: $D_{\mathbb{R}^{n \times m}} T(A)$

Theorem 1.6. ***The form of the Matrix operator derivative.***
Let T be described as above and differentiable at $A \in \mathbb{R}^{n \times m}$

$$\frac{T(A+h) - T(A)}{||h||} = \begin{bmatrix} \frac{T_1(A+h) - T_1(A)}{||h||} \\ \vdots \\ \frac{T_n(A+h) - T_n(A)}{||h||} \end{bmatrix}$$

and so:

$$\lim_{h \to 0} \frac{T(A+h) - T(A)}{||h||} = \begin{bmatrix} \lim_{h \to 0} \frac{T_1(A+h) - T_1(A)}{||h||} \\ \vdots \\ \lim_{h \to 0} \frac{T_n(A+h) - T_n(A)}{||h||} \end{bmatrix}$$

Definition 1.10. ***Subspace Differentiability***
Let $X = \{X_j\}_{j=1}^{n}$ *be a sequence of finite dimensional vector spaces where* $dim(X_j) = k_j = m_j \times n_j$
Let $T : \prod_{j=1}^{n} X_j \to Y$ *where* Y *is a finite dimensional vector space with* $dim(Y) = k_y$
Let $x_j \in X_j$ *for some* $j \in \mathbb{N}_n$
Where

$$x_j = \begin{bmatrix} x_{1,1} & \cdots & x_{1,n_j} \\ \vdots & \ddots & \vdots \\ x_{m_j,1} & \cdots & x_{m_j,n_j} \end{bmatrix}$$

T *is said to be differentiable at* $x \in X$ *where* $x = (x_0, ..., x_j, ..., x_{n-1})$ *with respect to* X_j *if there exists a linear operator* $D : X_j \to Y$*:*
Given $h \in X_j \setminus \{\vec{0}\}$ *define* $\hat{h} = (0, ..., h, ..., 0) \in X$ *where* h *is in the j'th place of* $\hat{h}$*:*

$$\lim_{h \to 0} \frac{||T(x + \hat{h}) - T(x) + D(h)||_X}{||h||_{X_j}} = 0$$

Then D *is called the subspace derivative of* T *at* x *with respect to* X_j *and is written:* $D_{x_j}T(x)$

Definition 1.11. ***Product space Derivative***

Let $X = \{X_j\}_{j=0}^{n-1}$ *be a sequence of finite dimensional vector spaces where* $dim(X_j) = k_j$

Let $T : \prod_{j=0}^{n-1} X_j \to Y$ *where* Y *is a finite dimensional vector space with* $dim(Y) = k_y$

Let $\{x_j\}_{j=0}^{n-1}$ *be a sequence of vectors such that:* $(\forall j \in \{0, ..., n-1\})(x_j \in X_j)$

The product space derivative at the point $z \in X$ *is:*

$$D_X T(z) = \begin{bmatrix} D_{x_0} T(z) \\ \vdots \\ D_{x_{n-1}} T(z) \end{bmatrix}$$

Definition 1.12. ***Fréchet derivative***
Let V, W be normed vector spaces and $U \subset V$ be an open set.
An operator $f : U \to W$ is said to be Fréchet differentiable if there exists a bounded linear operator $A : V \to W$ such that:

$$\lim_{||h|| \to 0} \frac{||f(x+h) - f(x) + Ah||_W}{||h||_V} = 0$$

Reference

Theorem 1.7. ***Fréchet derivative of a bounded linear operator***
Let V, W be normed vector spaces and $U \subset V$ be an open set.
Let $\hat{f} : V \to W$ be a bounded linear operator.
Then lets look at $f = \hat{f}|_U$
My guess is that $A = \hat{f}$
Let $x \in U$ and $h \in U \pitchfork ||h|| \neq 0$ and $x + h \in U$, Then:

$$\frac{||f(x+h) - f(x) + Ah||_W}{||h||_V} = \frac{||f(x) + f(h) - f(x) + \hat{f}(h)||_W}{||h||_V} = \frac{||f(x) + f(h) - f(x) + f(h)||_W}{||h||_V} = 0$$

Thus let $\epsilon > 0$ and $\delta > 0$
Then if $0 < ||h|| < \delta$ we know that $\frac{||f(x+h)-f(x)+Ah||_W}{||h||_V} = 0 < \epsilon$
Therefore:

$$\lim_{||h|| \to 0} \frac{||f(x+h) - f(x) + Ah||_W}{||h||_V} = 0$$

Thus $A = \hat{f}$ is the Fréchet derivative of f.

1.1 Finite Composition Operator

Definition 1.13. ***Finite Composition Operator***
Let the collection $X = \{X_j\}_{j=0}^{n}$ be a finite sequence of sets.
Further let $\{T_j\}_{j=0}^{n-1}$ be a finite sequence of operators such that $(\forall j \in \mathbb{N}_{n-1})(T_j : X_j \to X_{j+1})$
Then $T^n : X_0 \to X_n$ defined by:

$$T^n := \bigcirc_{j=0}^{n-1} T_j$$

*is called the **Finite Composition Operator defined on X**.*

Definition 1.14. ***Multi-variable Finite Composition Iteration***
Let the collection $X = \{X_j\}_{j=0}^{n}$ and $Y = \{Y_j\}_{j=0}^{n-1}$ be finite sequences of sets.
Further let $\{T_j\}_{j=0}^{n-1}$ be a finite sequence of operators such that: $(\forall j \in \mathbb{N}_{n-1})(T_j : X_j \times Y_j \to X_{j+1})$
Let $T^n : X_0 \times \prod_{j=0}^{n-1} Y_j \to X_n$ where:

$$T^n(x, y) = z_n \text{ where } z_{j+1} = T_j(z_j, \pi_j(y)) \text{ or } z_{j+1} = T_j(z_j) \text{ and } z_0 = x \in X_0$$

Definition 1.15. ***Gradient Descent***
Let $E : \mathbb{R}^n \to \mathbb{R}$ be a differentiable operator.
The method of Gradient Descent says that a local minimum of E can be found using the following iteration:

$$a_{n+1} = a_n - \gamma \nabla E(a_n)$$

Where $\gamma > 0$

Example 1.1. ***Objective Operator for Data Set Defined Operator Approximation***
Let $X \subset \mathbb{R}^n, Y \subset \mathbb{R}^m$ *such that* $X \times Y$ *defines an operator* T*.*

$$E(a) = \sum_{x \in X} ||T(x) - T^n(x, a)||$$

2 Surjective Continuous Non-decreasing Bounded Functionals

Let $B = \{f : \mathbb{R} \to [0,1] | f$ is surjective, continuous, and non-decreasing.$\}$

Theorem 2.1. ***B is convex.***
Let $f, g \in B$ and $h(x) := \lambda f(x) + (1-\lambda)g(x)$ where $\lambda \in [0,1]$
Then h is still continuous since the linear combination of continuous functions is continuous.

Since both f and g are surjective and non-decreasing, then there exists x_0, y_0, x_1, y_1 in $\mathbb{R}$ such that:
$f(x_0) = 0 = g(y_0)$ *and* $f(x_1) = 1 = g(y_1)$
Suppose WLOG that $x_0 \leq y_0$ and $x_1 \leq y_1$
Then we know that:

$$h(x_0) = \lambda f(x_0) + (1-\lambda)g(x_0) = \lambda 0 + (1-\lambda)0 = 0$$

and

$$h(y_1) = \lambda f(y_1) + (1-\lambda)g(y_1) = \lambda 1 + (1-\lambda)1 = 1$$

Now if we pick $\alpha \in [0,1]$ by the intermediate value theorem, we know that there exists an $x_\alpha \in [x_0, y_1]$ such that:

$$h(x_\alpha) = \alpha$$

Since α was arbitrary element, I have shown that h is surjective.

Finally, let $x_0 < x_1$ be elements in $\mathbb{R}$
Then we know that $f(x_0) \leq f(x_1)$ and $g(x_0) \leq g(x_1)$
$\Rightarrow \lambda f(x_0) \leq \lambda f(x_1)$ *and* $(1-\lambda)g(x_0) \leq (1-\lambda)g(x_1)$
$\Rightarrow \lambda f(x_0) + (1-\lambda)g(x_0) \leq \lambda f(x_1) + (1-\lambda)g(x_1)$
$\Rightarrow h(x_0) \leq h(x_1)$
Thus h is non-decreasing.
Since h is surjective, continuous, and non-decreasing, then $h \in B$
Thus B is convex.

Theorem 2.2. ***B is translation invariant.***
Let $f \in B$ and $g(x) := f(x + c)$ where $c \in \mathbb{R}$
f is continuous and so is the addition operator so g is continuous.
Let $\alpha \in [0,1]$ since f is surjective then $\exists x \in \mathbb{R} \pitchfork f(x) = \alpha$
Then $g(x - c) = f(x + c - c) = f(x) = \alpha$ and so g is surjective.
Let $x < y$ be elements in $\mathbb{R}$
Then $f(x) \leq f(y) \Rightarrow f(x + c) \leq f(y + c)$
$\Rightarrow g(x) \leq f(y)$ *and so g is non-decreasing.*
Thus $g \in B$ and B is therefore translation invariant.

Theorem 2.3. ***B is not complete.***

Theorem 2.4. ***Every element in B can be decomposed as a finite non-trivial convex combination from B***