

# Machine Learning

Daniel Drake

May 18, 2020

## 1 Metric Space, Limit, Derivative, and Partial Derivative

### Definition 1.1. *Metric*

Let  $X$  be a non-empty set..

Let  $d : X \times X \rightarrow \mathbb{R}_0^+$  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  $T : X \rightarrow Y$  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:

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

Reference

### Definition 1.3. *Derivative*

Let  $\hat{f} : \mathbb{R} \rightarrow \mathbb{R}$

Further let  $f = \hat{f}|_U$  where  $U \in \tau_{\mathbb{R}}$

Then  $f$  is said to be differentiable at  $a \in U$  if:

$$f'(a) = \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h}$$

Reference

### Definition 1.4. *Partial Derivative*

Reference

## 2 Gradients, Jacobian, Fréchet derivative, and Sub-Gradients

### Definition 2.1. Gradient

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be a Differentiable function.

Then  $\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n$  where:

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

is called the Gradient of  $f$ .

Reference

### Definition 2.2. Jacobian

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  where:

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix} \quad \forall x \in \mathbb{R}^n \text{ and } (\forall j \in \mathbb{N}_m)(f_j : \mathbb{R}^n \rightarrow \mathbb{R})$$

Then  $J_f : \mathbb{R}^n \rightarrow \mathbb{R}^{m \times n}$  where:

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

Reference

**Theorem 2.1.** Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}^k$  and  $g : \mathbb{R}^k \rightarrow \mathbb{R}^m$

Define:  $h(x) = f(g(x))$

where:

$$h(x) = \begin{bmatrix} f_1(g(x)) \\ \vdots \\ f_k(g(x)) \end{bmatrix}$$

$$J_h(x) := \begin{bmatrix} \frac{\partial}{\partial x_1} h_1(x) & \cdots & \frac{\partial}{\partial x_1} h_m(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_n} h_1(x) & \cdots & \frac{\partial}{\partial x_n} h_m(x) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1} f_1(g(x)) & \cdots & \frac{\partial}{\partial x_1} f_m(g(x)) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_n} f_1(g(x)) & \cdots & \frac{\partial}{\partial x_n} f_m(g(x)) \end{bmatrix}$$

### Definition 2.3. Derivative with respect to a vector and the subspace gradient.

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

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

Let  $x \in X_j$  for some  $j \in \{0, \dots, n-1\}$

Where

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

$$D_x T(z) = \begin{bmatrix} \frac{\partial}{\partial x_{1,1}} T_1(z) & \cdots & \frac{\partial}{\partial x_{1,n_j}} T_1(z) & \cdots & \frac{\partial}{\partial x_{1,1}} T_{n-1}(z) & \cdots & \frac{\partial}{\partial x_{1,n_j}} T_{n-1}(z) \\ \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_{m_j,1}} T_1(z) & \cdots & \frac{\partial}{\partial x_{m_j,n_j}} T_1(z) & \cdots & \frac{\partial}{\partial x_{m_j,1}} T_{n-1}(z) & \cdots & \frac{\partial}{\partial x_{m_j,n_j}} T_{n-1}(z) \end{bmatrix}$$

If  $k_y = 1$  then  $D_x T(z)$  is called the subspace gradient with respect to  $X_j$  at the point  $z$  and is written  $\nabla_x T(z)$

**Definition 2.4. 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  $T : \prod_{j=0}^{n-1} X_j \rightarrow Y$  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:  $(\forall j \in \{0, \dots, n-1\})(x_j \in X_j)$

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

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

**Example 2.1.**

Let  $X_0 = \mathbb{R}^n, X_1 = \mathbb{R}^m, X_2 = \mathbb{R}^m, X_3 = \mathbb{R}^m$

Let  $Y_0 = \mathbb{R}^{n \times m}, Y_1 = \mathbb{R}^m, Y_2 = \{0\}$  and  $Y = \prod_{j=0}^2 Y_j$

$$T^3 : X_0 \times \prod_{j=0}^2 Y_j \rightarrow X_3 \text{ where } T^3(x, A, a, 0) = \text{atan}(a + Ax)$$

$$T_0 : X_0 \times Y_0 \rightarrow X_1 \text{ where } T_0(x, A) = Ax$$

$$T_1(x, a) = a + x$$

$$T_2(x) = \text{atan}(x)$$

$$T^3(x, A, a, 0) = T_2(T_1(T_0(x, A), a))$$

Now fix  $x \in X_0$

Then we have a new operator,  $T_x : \prod_{j=0}^2 Y_j \rightarrow X_3$  where:

$$T_x(y) = T^3(x, y) \text{ and } y = (A, a, 0)$$

With all this we can now look at this:

Let  $z \in \prod_{j=0}^2 Y_j$

$$D_X ||T_x(z)|| \text{ where } ||\cdot|| \text{ is the one norm.}$$

First:

$$\nabla_{y_0} ||T_x(z)|| = \begin{bmatrix} \frac{\partial}{\partial a_{1,1}} ||T_x(z)|| & \cdots & \frac{\partial}{\partial a_{1,n}} ||T_x(z)|| \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial a_{m,1}} ||T_x(z)|| & \cdots & \frac{\partial}{\partial a_{m,n}} ||T_x(z)|| \end{bmatrix} \text{ Since } y_0 = A = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{bmatrix}$$

So then we will just look at:  $\frac{\partial}{\partial a_{i,j}} ||T_x(z)||$

$$\frac{\partial}{\partial a_{i,j}} ||T_x(z)|| = \frac{\partial}{\partial a_{i,j}} ||\text{atan}(a + Ax)||$$

$$\text{atan}(Ax + a) = \begin{bmatrix} \arctan(\sum_{l=1}^n a_{1,l}x_l + a_1) \\ \vdots \\ \arctan(\sum_{l=1}^n a_{m,l}x_l + a_m) \end{bmatrix}$$

and so  $||\text{atan}(Ax + a)|| = \sum_{i=1}^m |\arctan(\sum_{l=1}^n a_{i,l}x_l + a_i)|$

Thus the partial with respect to  $a_{i,j}$  zeros out all but:

$$\frac{\partial}{\partial a_{i,j}} |\arctan(\sum_{l=1}^n a_{i,l}x_l + a_i)| = \frac{\arctan(\sum_{l=1}^n a_{i,l}x_l + a_i)}{((\sum_{l=1}^n a_{i,l}x_l + a_i)^2 + 1) |\arctan(\sum_{l=1}^n a_{i,l}x_l + a_i)|} a_{i,j}x_j \text{ by the chain rule.}$$

**Definition 2.5. Fréchet derivative**

Let  $V, W$  be normed vector spaces and  $U \subset V$  be an open set.

An operator  $f : U \rightarrow W$  is said to be Fréchet differentiable if there exists a bounded linear operator  $A : V \rightarrow W$  such that:

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

Reference

**Theorem 2.2. Fréchet derivative of a bounded linear operator**

Let  $V, W$  be normed vector spaces and  $U \subset V$  be an open set.

Let  $\hat{f} : V \rightarrow W$  be a bounded linear operator.

Then let's look at  $f = \hat{f}|_U$

My guess is that  $A = \hat{f}$

Let  $x \in U$  and  $h \in U$  with  $\|h\| \neq 0$  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:

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

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

**2.1 Finite Composition Operator****Definition 2.6. 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  $(\forall j \in \mathbb{N}_{n-1})(T_j : X_j \rightarrow X_{j+1})$

Then  $T^n : X_0 \rightarrow X_n$  defined by:

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

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

**Theorem 2.3. Finite Composition Jacobian**

Let  $T^n$  be defined as above.

Then:

$$J_{T^n}(x) = J_{T_{n-1} \circ T^{n-1}}(x) = J_{T_{n-1}}(T^{n-1}(x)) J_{T^{n-1}}(x)$$

where:

$$J_{T^1}(x) = J_{T_0}(x)$$

**Definition 2.7. 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:  $(\forall j \in \mathbb{N}_{n-1})(T_j : X_j \times Y_j \rightarrow X_{j+1})$

Let  $T^n : X_0 \times \prod_{j=0}^{n-1} Y_j \rightarrow X_n$  where:

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

**Example 2.2. MVFCI**

Let  $X_0 = \mathbb{R}^n, X_1 = \mathbb{R}^m, X_2 = \mathbb{R}^m, X_3 = \mathbb{R}^m$

Let  $Y_0 = \mathbb{R}^{n \times m}, Y_1 = \mathbb{R}^m, Y_2 = \{0\}$

$$T^3 : X_0 \times \prod_{j=0}^2 Y_j \rightarrow X_n \text{ where } T^3(x, A, a, 0) = \text{atan}(a + Ax)$$

$$T_0 : X_0 \times Y_0 \rightarrow X_1 \text{ where } T_0(x, A) = Ax$$

$$T_1(x, a) = a + x$$

$$T_2(x) = \text{atan}(x)$$

$$T^3(x, A, a, 0) = T_2(T_1(T_0(x, A), a))$$

**Definition 2.8. Gradient Descent**

Let  $E : \mathbb{R}^n \rightarrow \mathbb{R}$  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 2.3. 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  $T$ .

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

### 3 Surjective Continuous Non-decreasing Bounded Functionals

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

**Theorem 3.1.  $B$  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  $\alpha$  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 3.2.  $B$  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  $\exists x \in \mathbb{R} \cap f(x) = \alpha$

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  $f(x) \leq f(y) \Rightarrow f(x + c) \leq f(y + c)$

$\Rightarrow g(x) \leq g(y)$  and so  $g$  is non-decreasing.

Thus  $g \in B$  and  $B$  is therefore translation invariant.

**Theorem 3.3.  $B$  is not complete.**

**Theorem 3.4. Every element in  $B$  can be decomposed as a finite non-trivial convex combination from  $B$**