

Machine Learning

Daniel Drake

May 14, 2020

1 Gradients, Jacobian, Ferchet Drivative, and Sub-Gradients

Definition 1.1. *Gradient*

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

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

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

is called the Gradient of f .

Reference

Definition 1.2. *Jacobian*

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

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

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

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

Reference

Theorem 1.1. *When the Jacobian is the Gradient*

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

Then $(\nabla f(x) = (J_f(x))^T)(\forall x \in \mathbb{R}^n)$

Definition 1.3. *Derivative with respect to a vector.*

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

$$D_x T(z) = \begin{bmatrix} \frac{\partial}{\partial x_1} T_1(z) & \cdots & \frac{\partial}{\partial x_{k_j}} T_1(z) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} T_{k_y}(z) & \cdots & \frac{\partial}{\partial x_{k_j}} T_{k_y}(z) \end{bmatrix}$$

Definition 1.4. 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 1.2. 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 $h \in U \cap \{h \mid \|h\| \neq 0 \text{ and } x+h \in U\}$, 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 .

Example 1.1. Let $\text{atan} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that:

$$\text{atan}(x) := \begin{bmatrix} \arctan(\pi_1(x)) \\ \vdots \\ \arctan(\pi_n(x)) \end{bmatrix} = \begin{bmatrix} \arctan(x_1) \\ \vdots \\ \arctan(x_n) \end{bmatrix}$$

Then

$$\begin{aligned} J_{\text{atan}}(x) &= \begin{bmatrix} \frac{\partial}{\partial x_1} \arctan(x_1) & \cdots & \frac{\partial}{\partial x_n} \arctan(x_1) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} \arctan(x_n) & \cdots & \frac{\partial}{\partial x_n} \arctan(x_n) \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial}{\partial x_1} \arctan(x_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\partial}{\partial x_n} \arctan(x_n) \end{bmatrix} = \begin{bmatrix} \frac{1}{x_1^2+1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{x_n^2+1} \end{bmatrix} = I \begin{bmatrix} \frac{1}{x_1^2+1} \\ \vdots \\ \frac{1}{x_n^2+1} \end{bmatrix} \end{aligned}$$

Example 1.2. Let $T_x : \mathbb{R}^n \rightarrow \mathbb{R}^n$ for each $x \in \mathbb{R}^n$ such that:

$$T_x(b) := \begin{bmatrix} x_1 + \pi_1(b) \\ \vdots \\ x_n + \pi_n(b) \end{bmatrix} = \begin{bmatrix} x_1 + b_1 \\ \vdots \\ x_n + b_n \end{bmatrix}$$

Then

$$J_{T_x}(b) = \begin{bmatrix} \frac{\partial}{\partial b_1} [x_1 + b_1] & \cdots & \frac{\partial}{\partial b_n} [x_1 + b_1] \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial b_1} [x_n + b_n] & \cdots & \frac{\partial}{\partial b_n} [x_n + b_n] \end{bmatrix} = \begin{bmatrix} b_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & b_n \end{bmatrix} = Ib$$

Example 1.3. Let $Z_x : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^m$ such that:

$$Z_x(A) = Ax$$

Let $A, B \in \mathbb{R}^{m \times n}$ and $\alpha, \beta \in \mathbb{R}$

$$Z_x(\alpha A + \beta B) = (\alpha A + \beta B)x = \alpha Ax + \beta Bx = \alpha Z_x(A) + \beta Z_x(B) \text{ and so } Z_x \text{ is linear.}$$

Since Z_x is continuous, it is also bounded.

Thus by Theorem 1.2, it is Frechet differentiable everywhere.

1.1 Finite Composition Operator

Definition 1.5. 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 1.3. 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 1.6. 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 1.4. MVFCI

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

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

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

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

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

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

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

Definition 1.7. 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 1.5. 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 $B = \{f : \mathbb{R} \rightarrow [0, 1] \mid f \text{ is surjective, continuous, and non-decreasing.}\}$

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 $\exists x \in \mathbb{R} \cap f(x) = \alpha$

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 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