Back to Fundamental things for DL

Differentiation and Convolution

Konstantin Burlachenko (burlachenkok@gmail.com)

Last update: 18 APR 2020

Contents

1 Introduction

This document is suggested to be evolved in time and contains some general math things in which DL aspects is leveraging on or heuristically or non-heuristically.
Anybody welcome to append into it **pure mathematical** type things more or less relative to: Differentiation and Convolution.

2 Differentiation

2.1 Introduction about Differentiation

This days people talk about chain rule when prediction schema $F(x;\theta) : X \to Y$ constructed as composition of several differentiable functions.

Leveraging in chain rule for differentiable function allows computing gradient of function in systematic way, such that the computation will be correct from math point of view, even numerically it can be various problems.

2.2 Differentiation Definition in Mathematics

Function $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable if for arbitrarily $\varDelta x \in \mathbb{R}^n$ which element-wise consist of some components (without any interdependency in them) we can have:

$$f(x + \varDelta x) - f(x) = A\varDelta x + \alpha(\varDelta x)|\varDelta x| \tag{1}$$

Where:

1. $\alpha(\Delta x)$ is infinite small function i.e. $lim_{\Delta x \to 0} \alpha(\Delta x) = 0$
2. A is a rectangular matrix and its items should not depend on specific direction or values of Δx
3. $\alpha(\Delta x)$ is infinite dimensional function relative to Δx

Differentiation is a local property of the function. If function is differentiable in all points in some set χ then people say that function is differentiable in the whole set χ.

2.3 Some Theorems formulations in context of Differentiation

Theorem 2.1 (Partial Derivatives). *If function f is differentiable in point then all partial derivatives of the function exist and equal to corresponded elements of the matrix A.*

To derive this theorem you should vary $\Delta x = [0, 0, ..., 0, dx_i, 0, ..., 0]^T$. This simple variation in input will allow eliminating every from A, except i-th column.

Theorem 2.2 (Continuity Property). *If function f is differentiable in point then function is continuous in the point.*

Theorem 2.3 (Differentiability from existing partial derivatives). *If function f has all partial derivatives in point, $\boldsymbol{a}$ and this partial derivative are continuous functions in a point $\boldsymbol{a}$ then function is differentiable in point $\boldsymbol{a}$.*

Theorem 2.4 (Differentiability of function composition). *Also known as the chain rule.*

1. *If function f is differentiable in point of it's domain $\boldsymbol{a}$.*
2. *If function g is differentiable in point of it's domain $\boldsymbol{b}$.*
3. *If $\boldsymbol{b} = f(\boldsymbol{a})$*

Then function $z = g(f(x))$ is differentiable in points a and $z' = g' \cdot f'$

2.4 Taylor Series for $\mathbb{R} \to \mathbb{R}$

Even this days more things are based on convexity, but nevertheless it's a very fundamental thing for numerical methods and mathematical analysis. Named after the English mathematician Brook Taylor (1685–1731). The form that I have studied in

Bauman Moscow State Technical University from lectures Prof. V.V.Feoktistov is a form of Taylor Series with remainder in form of variation of Schlömilch Remainder.

$$f(x) = T_n(x) + R_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + R_n(x) \tag{2}$$

$$R_n(x) = \frac{(x - x_0)^{n+1}(1 - \theta)^{n-p+1}}{pn!} f^{(n+1)}(x_0 + \theta(x - x_0)) \tag{3}$$

In this form we can choose: x_0,p,n, but not $\theta \in (0{,}1)$

Steps to prove this brilliant theorem:

1. $f(x) = T_n(x) + R_n(x)$, for some undefined $R_n(x)$. Nothing more than notation trick

2. But if take x_0 then in fact $f(x_0) - T_n(x_0) = 0 = R_n(x_0)$, so it seems that x_0 is a zero of function (aka as a root) $R_n(x_0) = 0$

3. Without loss of generality we can represent $R_n(x_0) = (x - x_0)^p M(x)$. Now via pushing all complications into new term $M(x)$

4. Now we fix x to some constant value. Now we will select some function $M(x)$ such that equality $f(x) = T_n(x) + (x - x_0)^p M(x)$ holds in point x

5. We consider the difference between the function f and its representation via Taylor Series $F(t) = f(t) - T_n(t) - (t - x_0)^p M(t)$

6. What is special about function $F(t)$, that $F(x_0) = F(x) = 0$. By Rolle's theorem we're in a situation that there exists at least one $c = x_0 + \theta(x -_0), \theta \in (0{,}1)$ such that. $F'(c) = 0$

7. After formal differentiation there are a lot of cancellation which finally will bring us that we can obtain form of $M(x)$

Various residuals

1. If we will select $p = n + 1$ then name of residual is Lagrange Remainder

2. If we will select $p = 1$ then name of residual is Cauchy Remainder

3. Peano's Form of Remainder $R_n(x) = o((x - x_0)^n)$. Its remainder is infinite small w.r.t. to $(x - x_0)^n$

2.5 Taylor Series for $\mathbb{R}^n \to \mathbb{R}$

Steps to prove this theorem:

1. We fix a point in linear space $\mathbb{R}^n$
2. We consider some displacement dx.
3. Now instead original function we consider $g(t) = f(a + dx \cdot t)$ and apply one dimensional Taylor series for it.
4. We look what is a value for $g(1)$ and write it down to write formally value of $f(x + dx)$

Formulation

$$f(x + dx) = \sum_{k=0}^{m} \frac{d^k f(x)}{k!} + \frac{d^{m+1} f(a + \nu dx)}{(m+1)!} \tag{4}$$

$$d^k f(x) = \left(\frac{\partial}{\partial x_1} dx_1 + ... + \frac{\partial}{\partial x_n} dx_n\right)^k f(x) \tag{5}$$

To be absolutely honest, the Taylor Theorem with a remainder in this *Lagrange-Form* stop to be true for the case of a vector-valued function, instead of a scalar argument.

3 What is convolution in mathematics

3.1 Convolution definition

I think you know that convolution is well-defined operation:

$$f * g(x) = \int_{-\infty}^{+\infty} f(y) g(x - y) dy \tag{6}$$

Sometimes limits of integration can be reduced to some subinterval of $[-\infty, +\infty]$.

3.2 Convolution commutativity

$$\begin{multline} f * g(x) = \int_{-\infty}^{+\infty} f(y) g(x - y) dy = \\ |x - y = z, dy = -dz| = \\ \int_{+\infty}^{-\infty} f(x - z) g(z)(-dz) = \\ \int_{-\infty}^{+\infty} g(z) f(x - z) dz = g * f(x) \end{multline} \tag{7}$$

So we have proved that $f * g = g * f$

3.3 Time invariant property of convolution

$$\begin{aligned}\tau(f * g,b) = f * g(x-b) = \int_{-\infty}^{+\infty} f(y)g((x-b)-y)dy = \\ |g(x-b) = \tau(g,b) \\ \int_{-\infty}^{+\infty} f(y)\tau(g,b)(x-y)dy = \\ f * \tau(g,b)\end{aligned} \tag{8}$$

So we have proved that $f * g = g * f$

3.4 Some Convolution Algebraic Properties

$$f * g = g * f \tag{9}$$

$$(f * g) * h) = f * (g * h) \tag{10}$$

$$f * (g + h) = f * g + f * h \tag{11}$$

$$\frac{d^k(f * g)}{dx^k} = f * \frac{d^k g}{dx^k} \tag{12}$$

$$\tag{13}$$

3.5 Convolution Connection with Sampling

$$f * \delta(a) = f(a) \tag{14}$$

$$\delta(x-a) * \delta(x-b) = \delta_a * \delta_b = \delta_{a+b} \tag{15}$$

$$\underset{p}{\text{Ш}} * f = \left(\sum_{n=-\infty}^{+\infty} \delta(x - np) \right) * f(x) = \sum_{k=-\infty}^{+\infty} f(x - kp) \tag{16}$$

4 What is cross-correlation in math

4.1 Cross-Correlation definition

If consider functions $\mathbb{R} \to \mathbb{R}$ then value of cross-correlation for them is **defined pointwise for specific value x** as the following

$$f \star g(x) = \int_{-\infty}^{+\infty} f(y)g(x+y)dy \tag{17}$$

If compared with convolution there is no "flipping and dragging" of convolution kernel in algebraic definition.

What Neural Network community mean by Convolution is some form of Cross-Correlations

4.2 Cross-Correlation connection with convolution

Let's define function $f^- := f(-x)$. This operation takes a function and reverse it w.r.t. to function value axis.

$$\begin{aligned} f \star g(x) = \int_{-\infty}^{+\infty} f(y)g(x+y)dy = \\ |y = -z, dy = -dz| = \\ \int_{+\infty}^{-\infty} f(-z)g(x-z)(-dz) = \\ \int_{-\infty}^{+\infty} f(-z)g(x-z)dz = \\ \int_{-\infty}^{+\infty} f^- g(x-z)dz = f^- * g \end{aligned} \tag{18}$$

So we derived $f \star g = f^- * g$

4.3 Some Cross-Correlation Properties not necessary to hold

Convolution behaves like multiplication, but as you see $f \star g = f^- * g$ So due to that different algebraic properties is not necessary to hold. List of some properties which in general does not hold

$$f \star g = f^{-} * g = g * f^{-} \neq g^{-} * f = g \star f \tag{19}$$

$$(f \star g) \star h \neq f \star (g \star h) \tag{20}$$