Calculus Refresher

Introduction

This refresher covers some basic concepts of calculus that will be useful throughout the machine learning workshop. We assume familiarity with (almost) all the material introduced below hence the refresher contains little to no proofs and is intended to serve as a refresher to most of the students. For simplicity, we define and explain all the necessary concepts in the single variable case, then demonstrate the transition to multivariate calculus since machine learning usually deals with high-dimensional objects.

Derivatives

The whole reason behind the field of calculus is providing all the necessary tools that enable simple analysis of functions.

Let $f : \mathbb{R} \to \mathbb{R}$ be a single variable, scalar-valued function, i.e. both the input and the output are reals. The derivative of f at an arbitrary input $x \in \mathbb{R}$, denoted as $f'(x)$ for $f' : \mathbb{R} \to \mathbb{R}$, describes the amount of change in f only around that point x:

$$\forall x \in \mathbb{R}, \quad f'(x) \underset{\text{def}}{=} \lim_{t \to 0} \frac{f(x+t) - f(x)}{t} \ .$$

It is important to note that this equation is not always well-defined. In this refresher we mainly cover infinitely differentiable functions, namely $f', (f')', (f'')', \ldots, ..$ all exist. Now let us see what properties one can harness from the derivative of a function.

Monotonicity. It is not difficult to see from the equation above that a function f is increasing if and only if its derivative is always positive, i.e. $f'(x) > 0$ for all $x \in \mathbb{R}$.

A function f is said to be increasing on an interval (a, b) (or $[a, b]$) if its derivative is positive on that interval. Similarly, when a function f is decreasing, its derivative f' is negative on the whole real line, or a given interval. A function is called *monotone* if it is either increasing or decreasing.

Critical points. Given a function $f : \mathbb{R} \to \mathbb{R}$ a point x^* is called an extremum when f attains its maximum or minimum at x^*. The derivative at such point must be zero $f'(x^*) = 0$.

This can be proved in the following way. Assume for instance that x^* is a minimum point. If however $f'(x^*) > 0$ then f is increasing around x^* then we can find a point slightly to the left of $x^* > \tilde{x}$ such that $f(\tilde{x}) < f(x^*)$ which contradicts optimality. An analogous arguments holds assuming $f'(x^*) < 0$.

Therefore extremum points of a function $f : R \to \mathbb{R}$ can be found by examining the zeros of the equation $f'(x) = 0$. The converse however is not true, i.e. $f'(x) = 0$ does not necessarily imply that x is an extremum point. For instance, the derivative of the function $f(x) = x^3$ at $x = 0$ is 0 but the point $x = 0$ is neither a minimum nor a maximum of f. Such points are called inflection or saddle points. Often a relatively small perturbation of a function yields dramatic changes to location and number of extrema. In the figure below we show two functions, $f(x) = 1/(1 + e^{-x^3})$, on the right hand side, and $h(x) = f(x) + (x/10 - 1)^2$. Try to find and identify the type of their critical points.

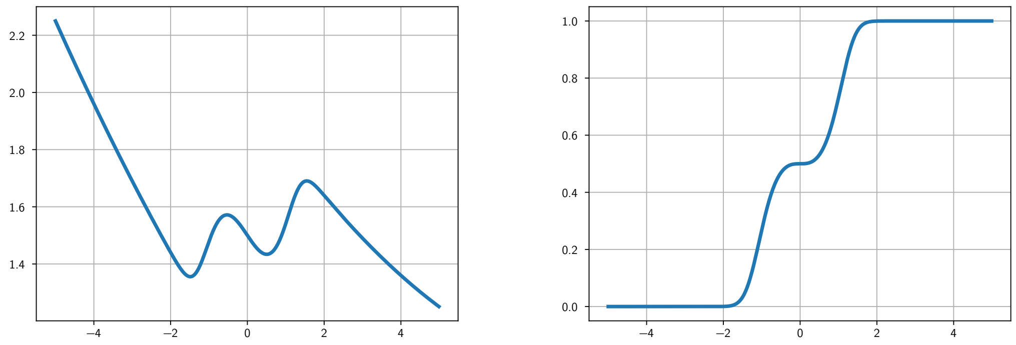

Figure 1. Plots of the function h (left) and f (see text).

Taylor Series

The Taylor series of an analytic function $f : \mathbb{R} \to \mathbb{R}$ for any $x \in \mathbb{R}$ at an arbitrary point $x_0 \in \mathbb{R}$ is given by,

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \ .$$

There is no need to be concerned by requirement analyticity as many useful functions are indeed analytic. The Taylor expansion results in an infinite series but it is often useful to build an approximation of f using a small number of summands. The Mean Value Theorem is a lesser known variant of Taylor's theorem that provides a finite representation of f for any $k \geq 1$ and $x, x_0 \in \mathbb{R}$,

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \cdots + \frac{f^{(k-1)}(x_0)}{(k-1)!}(x - x_0)^{k-1} + \frac{f^{(k)}(z)}{k!}(x - x_0)^k,$$

where $z \in [x_0, x]$ when $x > x_0$ and $z \in [x, x_0]$ otherwise.

Chain Rule

Another useful tool is the chain rule for simple rule for differentiating a composition of functions. Given $f, g : \mathbb{R} \to \mathbb{R}$ single variable functions, the composition of g with f is a function $h(x) = g(f(x))$. The composition is also denoted as $h \equiv g \circ f$. The order of composition is important – $g \circ f$ and $f \circ g$ are different functions. The chain rule states that,

$$(g \circ f)'(x) = g'(f(x))\, f'(x) \ \Leftrightarrow\ (g \circ f)' = (g' \circ f)\, f' \ .$$

To further understand the chain rule, let us define for a *given* input x a scalar $z = f(x)$. With this auxiliary variable we can write the chain rule as $(g(f(x)))' = g'(z)\, f'(x)$ which is the product of the derivatives of g evaluated at z and of f evaluated at x. The chain rule enables us to differentiate composite functions such as a logarithm of a polynomial or sum of exponentials, polynomial inside a logarithm, trigonometric or exponential functions, etc. However, the chain rule is also useful for functions not given explicitly so it is a good tool to have in the arsenal.

Multivariate Calculus

Gradients

Most of the concepts and properties available in single variable calculus transfer to calculus of multivariate functions. The analog of a derivative, the gradient, is defined for function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ where $n > 1$. The gradient of f at a point $\mathbf{x} = (x_1\, x_2\, \ldots\, x_n)$ is a vector of derivative. A standard way to define the gradient uses single-variable derivatives as follows.

1. For each $i \in= \{1, \ldots, n\}$ define the restricted single variable function
$$f_i(y) = f\big((x_1\, \ldots\, x_{i-1}\, y\, x_{i+1}\, \ldots\, x_n)\big) \; .$$
Meaning, for each coordinate of $\mathbf{x}$ the rest of the coordinates are fixed so as to define a scalar function $f_i : \mathbb{R} \rightarrow \mathbb{R}$.
2. Calculate the derivatives of each f_i with respect to x_i which are the partial derivatives of f,
$$\frac{\partial f}{\partial x_i} = f_i'(x_i) \; .$$
3. Concatenate the partial derivatives to an n-dimensional vector which is the gradient of f at $\mathbf{x}$,
$$\nabla f(\mathbf{x}) = \left[\frac{\partial f}{\partial x_1} \; \ldots \; \frac{\partial f}{\partial x_n} \right] \; .$$

Another useful concept is the directional derivative which is defined as the rate of change of f when moving away from a point $\mathbf{x}$ in along a direction defined by a vector $\mathbf{v}$,

$$Df(\mathbf{x})[\mathbf{v}] = \lim_{t \rightarrow 0} \frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t} \; .$$

Note that we can view $f(\mathbf{x} + t\mathbf{v})$ as a single variable function $h(t)$ and therefore,

$$Df(\mathbf{x})[\mathbf{v}] = \frac{dh}{dt} \; .$$

Moreover, we can use the chain rule and obtain that,

$$Df(\mathbf{x})[\mathbf{v}] = \mathbf{v}^\top \nabla f(\mathbf{x}) \; .$$

Critical Points

All extrema of f are points $\mathbf{x}^* \in \mathbb{R}^n$ which satisfy

$$\nabla f(\mathbf{x}^*) = \mathbf{0} = (0, 0, \ldots, 0) \; .$$

Multivariate Chain Rule

For a scalar function $h : \mathbb{R} \rightarrow \mathbb{R}$ and a multivariate function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ their composition $h \circ f$ is $h(f(\mathbf{x}))$. The chain rule implies that the gradient of $h \circ f$ is,

$$\nabla(h \circ f)(\mathbf{x}) = h'(f(\mathbf{x})) \nabla f(\mathbf{x}) \; .$$

Taylor Approximation

For this part, we simply state the more useful expression obtained via Taylor's and Mean Value theorems. For any $k \geq 1$, $\mathbf{x}, \mathbf{x}_0 \in \mathbb{R}^n$

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \frac{Df(\mathbf{x}_0)[\mathbf{x}-\mathbf{x}_0]}{1!} + \cdots = \sum_{i=0}^{k-1} \frac{D^i f(\mathbf{x}_0)[\mathbf{x}-\mathbf{x}_0]^i}{i!} + \frac{D^k f(\mathbf{z})[\mathbf{x}-\mathbf{x}_0]^k}{k!} \ ,$$

where $\mathbf{z}$ from the last term belongs to the line between $\mathbf{x}$ and $\mathbf{x}_0$ which is given by

$$\mathbf{z} \in [\mathbf{x}_0, \mathbf{x}] = \{t\mathbf{x}_0 + (1-t)\mathbf{x} \ : \ t \in [0,1]\} \ .$$

No need to worry about the higher order differentials in the given expression, there's not much novel about them, we just didn't cover them for simplicity, but note that $Df(\mathbf{x}_0)[\mathbf{x}-\mathbf{x}_0] = \nabla f(\mathbf{x}_0) \cdot (\mathbf{x}-\mathbf{x}_0)$.