COS 324, Precept #3:
Calculus Review

February 24, 2020

1 Introduction

This precept covers some basic concepts of calculus that will be useful throughout this course. We assume familiarity with (almost) all the material introduced below hence the precept 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.

2 Derivative

The whole reason behind the field of calculus is providing all the necessary tools that enable simple analysis of functions: we would like to investigate certain desirable properties of a given function. The concept of a derivative is the first stepping stone. 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:

$$\forall x \in \mathbb{R}, \quad f'(x) := \lim_{t \to 0} \frac{f(x+t) - f(x)}{t} \ . \tag{1}$$

It is important to note that the expression (1) is not always well-defined. In this class we mainly encounter differentiable functions. The existence of infinite differentiability is assumed throughout this document, namely $f', (f')', (f'')', \ldots, ..$ all exist and are well-defined. Now let us see what properties one can harness from the derivative of a function.

- **Monotonicity:** It is not difficult to see from (1) 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:** For a function $f : \mathbb{R} \to \mathbb{R}$, x^* is called an extremum point when f attains its maximum or minimum at x^*. Then, the derivative at that point must be zero, $f'(x^*) = 0$. This can be seen the following way: consider x^* to be a minimum point, the other case is analogous; if $f'(x^*) > 0$ then f is increasing around x^* so one can find a point slightly to the left of $x^* > \tilde{x}$ s.t. $\tilde{x}$ attains a smaller value of f, a contradiction; similarly, if $f'(x^*) < 0$ then the point $\tilde{x} > x^*$ would be slightly to the right and will still attain a smaller value of f, a contradiction. This whole argument can be made formal but the intuition behind is that an extremum point of a function $f : \mathbb{R} \to \mathbb{R}$ over the real line (if it exists) can be found by looking at all the critical points of the equation $f'(x) = 0$. The converse however is not true, i.e. $f'(x) = 0$ does not imply that x is an extremum point. The same rationale applies to open intervals (a, b) whereas for closed intervals $[a, b]$ the boundary points a and b are also extremum point candidates along with the critical points in (a, b).

- **Convexity:** Recall that for a single variable scalar-valued function $f : \mathbb{R} \to \mathbb{R}$ its derivative $f' : \mathbb{R} \to \mathbb{R}$ is also a single variable scalar-valued function so one can consider the derivative of f' denoted $f'' : \mathbb{R} \to \mathbb{R}$, called the second derivative of f. The second derivative provides more information about f itself. In particular, the condition $f''(x) > 0$ over the real line (or an interval) is equivalent to f being a convex function. Similarly, $f''(x) < 0$ is equivalent to concavity of f. In general, a convex function f has a unique extremum point, which is a *minimum* point, since f'' being positive means f' is increasing and can't have multiple roots to $f'(x) = 0$. The same relation holds with concave functions and maximum points. Note that higher order derivatives are defined the same way: the k^{th} order derivative of f, $f^{(k)}$, is the derivative of the $(k-1)^{\text{th}}$ order derivative function $f^{(k-1)}$.

3 Taylor Series

A function $f : \mathbb{R} \to \mathbb{R}$ can sometimes have a complicated, messy expression or no analytic expression at all. It is often useful to express the function as a polynomial and that's what the Taylor series does. Formally, 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

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

There's no need to worry about the analytic part in this class for two reasons: most functions we encounter are actually analytic. More importantly, one can drop the analytic part when assuming x is sufficiently close to x_0 which is the usual use case of Taylor series. The identity (2) is a step in the right direction however it is not always convenient to deal with an infinite sum. Furthermore, since we usually consider x, x_0 being close to each other each term vanishes by an additional order of $|x - x_0|$. A combination of the Taylor's theorem (not the same as above) and the Mean Value theorem under mild conditions results in the following expression for $f(x)$ that is often found to be much easier to work with. For any $k \geq 1$, any $x, x_0 \in \mathbb{R}$

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \cdots + \frac{f^{(k-1)}(x_0)}{(k-1)!}(x - x_0)^{k-1} + \frac{f^{(k)}(z)}{k!}(x - x_0)^k \tag{3}$$

where $z \in [x_0, x]$ lies on the interval from x_0 to x (change order if $x < x_0$).

4 Chain Rule

Another useful tool necessary for the class is the chain rule. It is a fairly simple rule for differentiating a composition of functions. In particular, given $f, g : \mathbb{R} \to \mathbb{R}$ single variable scalar-valued functions, define their composition, more specifically g composed with f, to be $h : \mathbb{R} \to \mathbb{R}$ such that $h(x) = g(f(x))$. The composition is denoted as $h \equiv g \circ f$. Note that the order of composition is important, $g \circ f$ and $f \circ g$ are different functions. The chain rule states that

$$\forall x \in \mathbb{R}, \quad (g \circ f)'(x) = g'(f(x)) \cdot f'(x) \tag{4}$$

In other words, $(g \circ f)' = (g' \circ f) \cdot f'$. This rule enables us to differentiate composite functions and can be useful in differentiating an example function you are given in class, for instance, 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.

5 Multivariate Calculus

5.1 Gradient

All of the concepts and properties (well, almost) transfer to the case of multivariate scalar-valued functions. First, we define the analog of a derivative, the gradient, for a function $f : \mathbb{R}^n \to \mathbb{R}$ where n now is any positive integer. The goal is to compute the gradient of f at a point $\mathbf{x} \in \mathbb{R}^n$ where $\mathbf{x} = [x_1\, x_2\, \ldots\, x_n]^T$ is a multidimensional vector.

One way to do this is by using the already defined derivative in the single variable case: (i) for each $i \in [n] = \{1, \ldots, n\}$ define the contracted single variable function $g_i : \mathbb{R} \to \mathbb{R}$ s.t. $g_i(y) = f([x_1\, \ldots\, x_{i-1}\, y\, x_{i+1}\, \ldots\, x_n]^T)$ for all $y \in \mathbb{R}$, i.e. for each coordinate of $\mathbf{x}$ fix all the values at the other coordinates then look at f as simply a function of that single coordinate; (ii) for the single variable functions $g_i(\cdot), i \in [n]$ compute the derivative at x_i, i.e. get $g_i'(x_i)$, which are called the partial derivatives of f and are otherwise denoted as $\frac{\partial f}{\partial x_i} = g_i'(x_i)$ for all $i \in [n]$; (iii) combine these partial derivatives into an n-dimensional vector to obtain the gradient of f at $\mathbf{x}$ given by

$$\nabla f(\mathbf{x}) = \left[\frac{\partial f}{\partial x_1} \cdots \frac{\partial f}{\partial x_n} \right]^T \tag{5}$$

Another way to look at the gradient is by considering the amount of change function f occurs just around the point $\mathbf{x}$ – this is how we started with the derivative. This approach gives the more natural derivation of a gradient. However, notice that unlike in the single variable case one now needs to choose a direction in which we slightly move from $\mathbf{x}$ to observe the change in the function value. More precisely, the *directional derivative* of a function $f : \mathbb{R}^n \to \mathbb{R}$ at a point $\mathbf{x} \in \mathbb{R}^n$ along a vector $\mathbf{v} \in \mathbb{R}^n$ is defined as

$$Df(\mathbf{x})[\mathbf{v}] = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t} \tag{6}$$

Notice that this is completely analogous to the definition in (1). However, this idea of describing amount of scalar change in function value seems to depend on a picked direction $\mathbf{v}$, a vector. Hence, the gradient vector is defined as the unique (for our purposes) vector that for any direction produces the directional derivative when taking the inner product along that direction. Thus, the gradient is defined as

$$\forall \mathbf{v} \in \mathbb{R}^n, \quad \nabla f(\mathbf{x}) \cdot \mathbf{v} = Df(\mathbf{x})[\mathbf{v}] \tag{7}$$

We remark that the first definition given in (5) is mostly useful for computing the gradient of an explicitly given function while the second definition in (7) gives the intuition behind the gradient and is potentially a useful property to keep track of. Next we quickly state the properties from the previous sections for the multivariate case (except monotonicity due to irrelevancy).

5.2 Properties

- **Critical points:** Any extremum point $\mathbf{x}^* \in \mathbb{R}^n$ of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ must satisfy the identity $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

- **Convexity:** A convex function $f : \mathbb{R}^n \to \mathbb{R}$ has a unique extremum point, a *minimum* point, while the same holds for a concave function $f : \mathbb{R}^n \to \mathbb{R}$ with a *maximum* point. There is an analogous condition given by the second order 'derivative' of f for convexity/concavity that is, for now, out of the scope.

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

  $$f(\mathbf{x}) = f(\mathbf{x}_0) + \frac{Df(\mathbf{x}_0)[\mathbf{x} - \mathbf{x}_0]}{1!} + \cdots = \sum_{i=0}^{k-1} \frac{D^i f(\mathbf{x}_0)[\mathbf{x} - \mathbf{x}_0]^i}{i!} + \frac{D^k f(\mathbf{z})[\mathbf{x} - \mathbf{x}_0]^k}{k!} \tag{8}$$

  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 $Df(\mathbf{x}_0)[\mathbf{x} - \mathbf{x}_0] = \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)$.

- **Chain Rule:** This is basically identical. For $g : \mathbb{R} \to \mathbb{R}, f : \mathbb{R}^n \to \mathbb{R}$, $g \circ f$ is defined identically, while chain rule states that $\nabla(g \circ f)(\mathbf{x}) = g'(f(\mathbf{x}))\nabla f(\mathbf{x})$.