11

The differential

11.1 Multivariable functions

In this chapter we will study multivariable functions $f : \mathbb{R}^n \to \mathbb{R}$, i.e. functions taking inputs in $\mathbb{R}^n$ and returning outputs in $\mathbb{R}$. We may think of the inputs in $\mathbb{R}^n$ as points or vectors of $\mathbb{R}^n$.

Many concepts seen for single variable functions translate with just very minor adjustments into the multivariable context. Let's first review the concepts of limit and continuity in this new context.

We say that a function $f : \mathbb{R}^n \to \mathbb{R}$ has limit L when x tends to a if for any $\epsilon > 0$ there is some $\delta > 0$ such that $0 < \|\mathbf{x} - \mathbf{a}\| < \delta$ implies that $|f(x) - f(a)| < \epsilon$, denoted

$$\lim_{x \to a} f(x) = L.$$

A function f will be *continuous* at a if

$$\lim_{x \to a} f(x) = f(a).$$

As with the single variable functions, the *domain* of f, denoted $\mathrm{Dom}(f)$, is the set of input values for which the function can return an output.

Unlike with single variable functions we are little more constrained to represent functions graphically, although the abstract concept of graph remains the same. Formally, the *graph* of a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ will be given by the following set of points:

$$\Gamma(f) = \{(x_1, \ldots, x_n, y) \in \mathbb{R}^{n+1} \mid f(x_1, \ldots, x_n) = y\}.$$

In the case of functions $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ we can think of $\Gamma(f)$ as a surface and in general, for functions of the form $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, we can think of $\Gamma(f)$ as a hypersurface embedded in $\mathbb{R}^{n+1}$.

We define the *level set* of the function f corresponding to level $c \in R$, denoted $f^{-1}(c)$ as the following set of input points:

$$f^{-1}(c) = \{x \in \mathbb{R}^n \mid f(x) = c\}.$$

In the case of functions $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ we can think of the level sets as curves as a surface, although in some cases the set will not be a proper curve. For example, for the function $f(x, y) = x^2 + y^2$ we can describe three types of level sets:

1. If $c < 0$ there is no point $(x, y) \in \mathbb{R}^2$ in the domain such that $f(x, y) = c$, consequently $f^{-1}(c)$ is the empty set.

2. If $c = 0$ then $f^{-1}(c)$ is just one point, namely $\{(0,0)\}$.

3. If $c > 0$ then $f^{-1}(c)$ is the circumference of center $(0,0)$ and radius $\sqrt{c}$.

11.2 The differential

When trying to translate the concept of local trend, a.k.a. derivative, to the multivariable setting, we encounter an important difference with the single variable setting: a function can exhibit many different local trends depending on the direction by which we approach our point of interest.

11.2.1 Slicing multivariable functions

As a first instructive example, let's analyse of the local behaviour of the function $f(x,y) = xy$ at the point $a = (1,2)$.

First imagine that we approach a following a trajectory that is parallel to the x-axis, in other words, we take inputs of the form $(t,2)$ setting the y-component of a fixed. What is the trend of f restricted to this line? It must be the same trend given by the single variable function $\phi(t) = f(t,2) = 2t$, namely, $\phi'(1) = 2$.

Now imagine that we approach a following a trajectory that is parallel to the y-axis, in other words, we take inputs of the form $(1,s)$ setting the x-component of a fixed. What is the trend of f restricted to this new line? It must be the same trend given by the single variable function $\psi(s) = f(1,s) = s$, namely, $\psi'(2) = 1$. The local trend seems to be a little bit more complex than it was for the single variable case.

11.2.2 Partial derivatives

We can define a flexible way of computing the derivative of f that uses as additional input information the direction we follow to approach a. The *directional derivate* of f at a following the direction of a unitary vector v is defined as the following limit expression:

$$\frac{\partial f}{\partial v}(a) = \lim_{h \to 0} \frac{f(a + hv) - f(a)}{h}.$$

When $v = e_i$, the i-th vector of the canonical basis, the directional derivative of f at a following the direction of e_i is known as the *i-th partial derivative* of f at a and it is denoted $\frac{\partial f}{\partial x_i}(a)$.

Using the limit expression we can observe the following relationship,

$$f(a + he_i) = f(a) + \frac{\partial f}{\partial x_i}(a)h + \mathfrak{o}(h),$$

and $\frac{\partial f}{\partial x_i}(a)$ is the only value that fulfills this equation.

11.2.3 Differentiable functions

We have seen that in the multivariable case we can compute many local trends of f at a, each attached to a possible direction from which we can approach a. When working with multivariable functions, we will look for classes of functions that are easy to study and understand but they are not so restrictive. The next definition is motivated by the requirement that the local trend values are not completely arbitrary, but they keep instead some coherence in a sense that will be made precise soon.

We say that the function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is *differentiable* at a if there is a linear transformation $L : \mathbb{R}^n \longrightarrow \mathbb{R}$, known as the *differential*, such that

$$\lim_{h \to \vec{0}} \frac{f(a+h) - f(a) - L(h)}{\|h\|} = 0,$$

or equivalently,

$$f(x) = f(a) + L(x-a) + \mathfrak{o}(\|x-a\|).$$

If f is differentiable the linear transformation L is uniquely determined (why?). What is the matrix of the differential L? Observe that the limit above must hold no matter which path the vector h takes towards the zero vector $\vec{0}$. If we now take $x = a + he_i$ we can see that

$$f(x) = f(a) + L(he_i) + \mathfrak{o}(h) = f(a) + L_i h + \mathfrak{o}(h),$$

where L_i denotes the i-th element of the row matrix L. But in view of the discussion in Section 11.2.2 we can conclude that $L_i = \frac{\partial f}{\partial x_i}(a)$.

The matrix of the differential

$$Jf(a) = [\frac{\partial f}{\partial x_1}(a) \ldots \frac{\partial f}{\partial x_n}(a)] \in \mathbb{R}^{1 \times n}$$

is known as the *Jacobian matrix* of f at a and defines a linear transformation $\mathbb{R}^n \longrightarrow \mathbb{R}$ that attaches to each possible direction a value for the local trend of f at a.

Proposition 36. *If f is differentiable at a, we can compute the directional derivative of f at a following the direction of a unitary vector v as follows*

$$\frac{\partial f}{\partial v}(a) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(a)v_i = Jf(a)v$$

11.2.4 Local affine approximation

The discussion above already implies that for any differentiable function f and point a there is a multivariable local affine approximation that takes the following form:

$$G(x) = f(a) + Jf(a)(x - a)$$

We can readily verify that the function G satisfies the following conditions:

1. $G(a) = f(a)$
2. $\frac{\partial G}{\partial x_i}(a) = \frac{\partial f}{\partial x_i}(a)$ for all $1 \leq i \leq n$

Also following the discussion in the previous section, the affine approximation satisfies a somewhat modified approximation condition compared to the single variable case, i.e. the approximation error is $\mathfrak{o}(\|x-a\|)$. We can state the result in the following proposition:

Proposition 37. *If f is differentiable at a, then*

$$f(x) = f(a) + Jf(a)(x - a) + \mathfrak{o}(\|x - a\|).$$

11.3 The gradient vector

Among all the possible directions v that we can use to compute the directional derivative $\frac{\partial f}{\partial v}(a)$, which unitary vector v would render the highest possible value? In other words, in which direction from the point a the growth of f is steepest?

As we already know that

$$\frac{\partial f}{\partial v}(a) = Jf(a)v,$$

the problem admits an easy answer. Since $Jf(a)$ is a row matrix, we can think of $Jf(a)v$ as the following dot product:

$$\nabla f(a) \cdot v,$$

where

$$\nabla f(a) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(a) \\ \vdots \\ \frac{\partial f}{\partial x_n}(a) \end{bmatrix}.$$

Assuming that $Jf(a)$ is not zero (what would happen if $Jf(a)$ is a matrix with zeros?), since $\nabla f(a) \cdot v = \|\nabla f(a)\| \|v\| \cos(\theta)$, with θ being the angle formed by the two vectors, the maximum value would be attained by

$$\hat{v} = \frac{\nabla f(a)}{\|\nabla f(a)\|},$$

a unitary vector that points in the same direction as $\nabla f(a)$. The vector $\nabla f(a)$ is known as the *gradient* of f at a.

Another quiz: among all the possible directions v that we can use to compute the directional derivative $\frac{\partial f}{\partial v}(a)$, which unitary vectors v would render a null local trend? Following the same rationale, the solution is the set of unitary vectors that are orthogonal to $\nabla f(a)$, i.e. the unitary vectors of the vector subspace $\nabla f(a)^{\perp}$. This remark corroborates the intuition that the direction of steepest ascent, given by the gradient vector, must be perpendicular to the level curves in the case of functions $f : \mathbb{R}^2 \rightarrow \mathbb{R}$.

11.4 TL;DR

The derivative and the more general concept of differential are just convenient linear transformations to describe the local behaviour of functions about points of interest: those functions that admit local affine approximation are known as differentiable functions. The differential is the linear transformation that provides the local trend for any possible direction to approach the point of interest, the matrix of the differential is known as the Jacobian matrix and its entries are the partial derivatives. The gradient vector has also the partial derivatives as its entries, although it has a different interpretation: it signals the direction of local steepest ascent.