CHAPTER 2

Differentiation

1. Directional and Partial Derivatives

DEFINITION 1.1. Let $U \subset \mathbb{R}^d$ be a domain, $f : U \to \mathbb{R}$ be a function, and $v \in \mathbb{R}^d - \{0\}$ be a vector. We define the *directional derivative* of f in the direction v at the point a by

$$D_v f(a) \stackrel{\text{def}}{=} \frac{d}{dt} f(a+tv)\Big|_{t=0}$$

EXAMPLE 1.2. If $f(x) = |x|^2$, then $D_v f(x) = 2x \cdot v$.

REMARK 1.3. Be aware that some authors define $D_v f$ by additionally dividing by the length of v. We will never do that!

DEFINITION 1.4. We define the i^{th} partial derivative of f (denoted by $\partial_i f$) to be the directional derivative of f in direction e_i (where e_i is the i^{th} elementary basis vector).

Practically, to compute the i^{th} partial derivative of f differentiate it with respect to x_i treating all the other coordinates as constant.

EXAMPLE 1.5. For $x \neq 0$ we have $\partial_i |x| = x_i/|x|^2$.

2. Derivatives

DEFINITION 2.1. Let $U \subseteq \mathbb{R}^d$ be a domain, $f : \mathbb{R}^d \to \mathbb{R}$ be a function, and $a \in U$. We say f is *differentiable at* a if there exists a linear transformation $T : \mathbb{R}^d \to \mathbb{R}$ and a function e such that

(1) $f(a+h) = f(a) + Th + e(h)$
(2) *and* $\lim_{h\to 0} |e(h)|/|h| = 0$.

In this case, the linear transformation T is called *the derivative of* f *at* a, and denoted by Df_a.

PROPOSITION 2.2. *If f is differentiable at a, then all the directional derivatives $D_v f(a)$ exist. Further,*

$$Df_a = \begin{pmatrix} \partial_1 f(a) & \partial_2 f(a) & \cdots & \partial_d f(a) \end{pmatrix}$$

and

$$D_v f(a) = Df_a v = \sum_{i=1}^{d} v_i \partial_i f(a).$$

REMARK 2.3. This shows that the linear transformation appearing in the definition of f is unique!

The converse of Proposition 2.2 is (surprisingly?) false. All directional derivatives can exist, however, the function need not be differentiable (or even continuous!)

EXAMPLE 2.4. Let $f(x,y) = x^2 y/(x^4+y^2)$. Then for every $v \in \mathbb{R}^2 - \{0\}$, $D_v f(0)$ exists, but f is not differentiable (or even continuous) at 0.

The converse of Proposition 2.2 is true under the additional assumption that the partial derivatives are continuous.

THEOREM 2.5. *If all partial derivatives of f exist in a neighbourhood of a,* and are continuous at a, *then f is differentiable at a.*

PROOF. For simplicity we assume $d = 2$. By the mean value theorem

$$\begin{aligned} f(a+h) - f(a) &= f(a_1+h_1, a_2+h_2) - f(a_1+h_1, a_2) + f(a_1+h_1, a_2) - f(a_1, a_2) \\ &= h_2 \partial_2 f(a_1+h_1, a_2+\xi_2) + h_1 \partial_1 f(a_1+\xi_1, a_2) \end{aligned}$$

for some ξ_1, ξ_2 such that ξ_i lies between 0 and h_i. Now let T be the matrix $(\partial_1 f(a)\ \partial_2 f(a))$ and observe

$$f(a+h) = f(a) + Th + e(h),$$

where

$$e(h) = h_2(\partial_2 f(a_1+h_1, a_2+\xi_2) - \partial_2 f(a)) + h_1(\partial_1 f(a_1+\xi_1, a_2) - \partial_1 f(a)).$$

Clearly

$$\frac{|e(h)|}{|h|} \leqslant |\partial_2 f(a_1+h_1, a_2+\xi_2) - \partial_2 f(a)| + |\partial_1 f(a_1+\xi_1, a_2) - \partial_1 f(a)|,$$

which converges to 0 as $h \to 0$. □

Note, however, it is possible for a function to be differentiable, and for the partial derivatives to exist and be discontinuous (e.g. $f(x) = |x|^2 \sin(1/|x|)$).

DEFINITION 2.6. Let $U \subset \mathbb{R}^m$ be a domain, and $a \in U$. We say a function $U \to \mathbb{R}^n$ is differentiable if if there exists a linear transformation $T : \mathbb{R}^m \to \mathbb{R}^n$ and a function e such that

(1) $f(a+h) = f(a) + Th + e(h)$
(2) *and* $\lim_{h\to 0} |e(h)|/|h| = 0$.

Note this is *exactly the same* as Definition 2.1. In this case Df is a $n \times m$ matrix given by

$$Df_a = \begin{pmatrix} \partial_1 f_1(a) & \partial_2 f_1(a) & \cdots & \partial_m f_1(a) \\ \partial_1 f_2(a) & \partial_2 f_2(a) & \cdots & \partial_m f_2(a) \\ \vdots & \vdots & & \vdots \\ \partial_1 f_n(a) & \partial_2 f_n(a) & \cdots & \partial_m f_n(a) \end{pmatrix}$$

and is called the *Jacobian Matrix*.

3. Tangent planes and Level Sets

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable.

DEFINITION 3.1. The *graph* of f is the set $\Gamma \subset \mathbb{R}^{d+1}$ defined by

$$\Gamma = \{(x, f(x)) \mid x \in \mathbb{R}^d\}.$$

Given a point $(a, f(a)) \in \Gamma$ we define the *tangent plane* of f at the point a by the equation

$$y = f(a) + Df_a(x - a)$$

The tangent plane is the *best linear approximation* to the graph Γ at the point a. Projecting the tangent plane into 2 dimensions (by freezing other coordinates) gives you a tangent line.

DEFINITION 3.2. Given $c \in \mathbb{R}$ we define the *level set of* f to be the set $\{x \in \mathbb{R}^d \mid f(x) = c\}$.

If $d = 2$, then level sets are typically curves. If $d = 3$, then level sets are typically surfaces. In higher dimensions (for "nice functions") level sets of f are typically $d-1$-dimensional hyper-surfaces.

EXAMPLE 3.3. Let $d = 3$ and $f(x) = |x|^2$. Then $\{f(x) = c\}$ is the sphere of radius $\sqrt{c}$ for $c > 0$, a point for $c = 0$ and the empty set for $c < 0$.

Level sets are very useful in plotting, and are often used to produce *contour plots*. We will see later that if v is tangent to a level set of f, then $D_v f = 0$.

4. Chain rule

The one variable calculus rules for differentiation of sums, products and quotients (when they make sense) are still valid in higher dimensions.

PROPOSITION 4.1. *Let $f, g : \mathbb{R}^d \to \mathbb{R}$ be two differentiable functions.*

- *$f + g$ is differentiable and $D(f + g) = Df + Dg$.*
- *fg is differentiable and $D(fg) = fDg + gDf$.*
- *At points where $g \neq 0$, f/g is also differentiable and*

$$D\Big(\frac{f}{g}\Big) = \frac{gDf - fDg}{g^2}$$

These follow in a manner very similar to the one variable analogues, and are left for you to verify. The one rule that is a little different in this context is the differentiation of composites.

THEOREM 4.2 (Chain Rule). *Let $U \subseteq \mathbb{R}^m$, $V \subseteq \mathbb{R}^n$ be domains, $g : U \to V$, $f : V \to \mathbb{R}^d$ be two differentiable functions. Then $f \circ g : U \to \mathbb{R}^d$ is also differentiable and*

$$D(f \circ g)_a = (Df_{g(a)})(Dg_a)$$

Note Df_g and Dg are both matrices, and the product above is the *matrix product* of Df and Dg.

PROOF. The basic intuition is as follows:

$$\begin{aligned} f(g(a+h)) &= f(g(a) + Dg_a h + e(h)) \approx f(g(a) + Dg_a h) \\ &= f(g(a)) + Df_{g(a)}(Dg_a h) + e_2(Dg_a h) \approx f(g(a)) + Df_{g(a)}(Dg_a h), \end{aligned}$$

since the composition of linear transformations is again linear. A more detailed version was done in class, and the complete ε-δ version is on your homework. □

Note if $d = 1$, then

$$\partial_i(f \circ g) = (Df_g)(Dg)e_i = \sum_{j=1}^{n} \partial_j f\Big|_g \partial_i g_j.$$

This is extremely useful, so I recommend remembering it (and not just the fancy matrix product version).

As a consequence, here is a "proof" that directional derivatives in directions tangent to level sets vanish.

PROPOSITION 4.3. *Let $\Gamma = \{x \mid f(x) = c\}$ be a level set of a differentiable function f. Let $\gamma : [-1, 1] \to \Gamma$ be a differentiable function, $v = D\gamma(0)$, and $a = \gamma(0)$. Then $D_v f(a) = 0$.*

Think of $\gamma(t)$ as the position of a particle at time t. If for all t, $\gamma(t)$ belongs to the curve Γ, then the velocity $D\gamma$ should be tangent to the curve γ, and thus thus the vector v above should be tangent to Γ. (When we can define this rigorously, we will revisit it and prove it.)

PROOF. Note $f \circ \gamma = c$ (since $\gamma(t) \in \Gamma$ for all t). By the chain rule $D(f \circ \gamma) = Df_\gamma D\gamma$. At $t = 0$ this gives $Df_{\gamma(0)} v = 0 \implies D_v f(\gamma(0)) = 0$ as desired. □

DEFINITION 4.4. If $f : \mathbb{R}^d \to R$ is differentiable, define the *gradient* of f (denoted by ∇f) to be the transpose of the derivative of f.

We've seen above that if v is tangent to a level set of f at a, then $D_v f(a) = 0$. This is equivalent to saying $\nabla f(a) \cdot v = 0$, or that *the gradient of f is perpendicular to level sets of f*. Intuitively, in directions tangent to level sets, f is changing the *least*. In the perpendicular direction (given by ∇f), the function f is changing the most.

5. Higher order derivatives

Given a function f, treat $\partial_i f$ as a function. If $\partial_i f$ is itself a differentiable function, we can differentiate it again. The second derivative (denoted by $\partial_j \partial_i f$) is called a second order partial of f. These can further be differentiated to obtain third order partials.

THEOREM 5.1 (Clairaut). *If $\partial_i \partial_j f$ and $\partial_j \partial_i f$ both exist in a neighbourhood of a, and are continuous at a then they must be equal.*

If the mixed second order partials are not continuous, however, they need not be equal.

EXAMPLE 5.2. Let $f(x,y) = x^3y/(x^2+y^2)$ for $(x,y) \neq 0$ and $f(0,0) = 0$. Then $\partial_x\partial_y f(0,0) = 1$ but $\partial_y\partial_x f(0,0) = 0$.

PROOF OF CLAIRAUT'S THEOREM. Here's the idea in 2D (the same works in higher dimensions). For simplicity assume $a = 0$.

- Let R be the rectangle with corners $(0,0)$, $(h,0)$, $(0,k)$, (h,k).
- Using the mean value theorem, show $f(h,k) - f(h,0) - f(0,k) + f(0,0) = hk\partial_x\partial_y f(\alpha)$ for some point $\alpha \in R$.
- Observe $f(h,k) - f(h,0) - f(0,k) + f(0,0) = f(h,k) - f(0,k) - f(h,0) + f(0,0)$ and so using the mean value theorem show $f(h,k) - f(h,0) - f(0,k) + f(0,0) = hk\partial_y\partial_x f(\beta)$ for some point $\beta \in R$.
- Note that as $(h,k) \to 0$, we have $\alpha, \beta \to 0$. Consequently, if $\partial_x\partial_y f$ and $\partial_y\partial_x f$ are both continuous at 0 we must have

$$\partial_x\partial_y f(0,0) = \lim_{(h,k)\to 0} \frac{f(h,k) - f(h,0) - f(0,k) + f(0,0)}{hk} = \partial_y\partial_x f(0,0),$$

proving equality as desired. □

DEFINITION 5.3. A function is said to be of class C^k if all its k^{th}-order partial derivatives exist and are continuous.

By Clairaut's theorem, we know that mixed partials are equal for C^k functions.

6. Maxima and Minima

DEFINITION 6.1. A function f has a local maximum at a if $\exists \varepsilon > 0$ such that whenever $|x - a| < \varepsilon$ we have $f(x) \leqslant f(a)$.

Our aim is now to understand what having a local maximum / minimum translates to in terms of derivatives of f. For this we do a simple calculation: Observe that if f has a local maximum at a, then for all $v \in \mathbb{R}^d - \{0\}$ the function $f(a+tv)$ must have a local maximum at $t = 0$. Hence we must have $\partial_t f(a+tv)|_{t=0} = 0$ and $\partial_t^2 f(a+tv)|_{t=0} \leqslant 0$. Using the chain rule, we compute

$$\partial_t f(a+tv) = \sum_{i=1}^d \partial_i f(a+tv)v_i \quad \text{and} \quad \partial_t^2 f(a+tv) = \sum_{i,j=1}^d \partial_i\partial_j f(a+tv)v_iv_j$$

Thus at a local maximum we must have

$$\sum_{i=1}^d \partial_i f(a)v_i = 0 \quad \text{and} \quad \sum_{i,j=1}^d \partial_i\partial_j f(a)v_iv_j \leqslant 0$$

for every $v \in \mathbb{R}^d$. This translates to the following proposition.

PROPOSITION 6.2. *If f is a C^2 function which has a local maximum at a, then*

(1) The first derivative Df must vanish at a (i.e. $Df_a = 0$). $Df_a = 0$
(2) The Hessian *Hf is negative semi-definite at a.*

For a local maximum, we replace negative semi-definite *above with* positive semi-definite.

DEFINITION 6.3. The *Hessian* of a C^2 function (denoted by Hf) is defined to be the matrix

$$Hf = \begin{pmatrix} \partial_1\partial_1 f & \partial_2\partial_1 f & \cdots & \partial_d\partial_1 f \\ \partial_1\partial_2 f & \partial_2\partial_2 f & \cdots & \partial_d\partial_2 f \\ \vdots & \vdots & & \vdots \\ \partial_1\partial_d f & \partial_2\partial_d f & \cdots & \partial_d\partial_d f \end{pmatrix}$$

Note if $f \in C^2$, Hf is symmetric.

DEFINITION 6.4. Let A be a $d \times d$ *symmetric* matrix.

- If $(Av) \cdot v \leqslant 0$ for all $v \in \mathbb{R}^d$, then A is called *negative semi-definite*.
- If $(Av) \cdot v < 0$ for all $v \in \mathbb{R}^d$, then A is called *negative definite*.
- If $(Av) \cdot v \geqslant 0$ for all $v \in \mathbb{R}^d$, then A is called *positive semi-definite*.
- If $(Av) \cdot v > 0$ for all $v \in \mathbb{R}^d$, then A is called *positive definite*.

Recall a symmetric matrix is positive semi-definite if all the eigenvalues are non-negative. In 2D this simplifies to the following:

PROPOSITION 6.5. *Let A be the symmetric 2×2 matrix $\left(\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right)$.*

- *A is positive semi-definite if and only if $a \geqslant 0$ and $ac - b^2 \geqslant 0$.*
- *A is negative semi-definite if and only if $a \leqslant 0$ and $ac - b^2 \geqslant 0$.*

For positive/negative definite we only need to additionally insist $ac - b^2 > 0$.

Finally, we address the converse: Namely, we look for a condition on the derivatives of f that guarantees that f attains a local maximum or minimum at a.

THEOREM 6.6. *Let f be a C^2 function.*

- *If $Df_a = 0$ and further Hf_a is positive definite, then f attains a local minimum at a.*
- *If $Df_a = 0$ and further Hf_a is negative definite, then f attains a local minimum at a.*

The proof uses Taylor's theorem, and we will revisit it later.

DEFINITION 6.7. We say a is a *saddle point* of f if $Df_a = 0$ and Hf_a has at least one strictly positive eigenvalue, and at least one strictly negative eigenvalue.

This corresponds to points where f has a local maximum in one direction and a local minimum in the other.

EXAMPLE 6.8. The function $|x|^2$ has a local minimum at 0. The function $-|x|^2$ has a local maximum at 0 The function $x_1^2 - x_2^2$ has a saddle at 0.

EXAMPLE 6.9. Let Γ be the hyper-surface $y = f(x)$, and $(z,t) \in \mathbb{R}^{d+1}$. Let $(a, f(a))$ be the point on Γ which is closest to (z,t). Then $z - a$ is parallel to ∇f and $(z - a, t - f(a))$ is normal to the tangent plane at $(a, f(a))$.

PROOF. Let $d(x) = |x - z|^2 + (f(x) - t)^2$. At a max $\nabla d = 0$, and hence $2(x - z) + 2(f(x) - t)\nabla f(x) = 0$ at $x = a$. This shows $a - z$ is parallel to $\nabla f(a)$ and $(Df_a, -1)^T$ is parallel to $(z - a, t - f(a))$. □

7. Taylors theorem

THEOREM 7.1. *If $f \in C^2$, then*

$$f(a+h) = f(a) + Df_a h + \frac{1}{2} h \cdot Hf_a h + R_2(h), \tag{7.1}$$

where $R_2(h)$ is some function such that

$$\lim_{h\to 0} \frac{R_2(h)}{|h|^2} \to 0.$$

In coordinates equation (7.1) is

$$f(a+h) = f(a) + \sum_i \partial_i f(a) h_i + \frac{1}{2} \sum_{i,j} \partial_i \partial_j f(a) h_i h_j + R_2(h).$$

PROOF. Let $g(t) = f(a+th)$. Using the 1D Taylors theorem we have

$$g(1) = g(0) + g'(0) + \frac{1}{2} g''(\xi)$$

for some $\xi \in (0,1)$. Writing this in terms of f finishes the proof. □

The same technique can show the following mean value theorem:

THEOREM 7.2 (Mean value theorem). *If f is differentiable on the entire line joining a and b,*

$$f(b) = f(a) + (b-a) \cdot \nabla f(\xi)$$

for some point ξ on the line segment joining a and b.

Taylor's theorem allows us to prove Theorem 6.6.

PROOF OF THEOREM 6.6. Suppose $Df_a = 0$ and Hf_a is positive definite. Let λ_0 be the smallest eigenvalue of Hf_a. Expanding in terms of an orthonormal basis of eigenfunctions of Hf_a we see $Hh \cdot h \geqslant \lambda_0 |h|^2$.

Now choose $\delta > 0$ so that $|R_2(h)| < \lambda_0 |h|^2/2$ for $h < \delta$, and note $f(a+h) \geqslant f(a) + \frac{|h|^2}{2} \geqslant f(a)$, showing f has a local min at a. □

A higher order version of Taylor's theorem is also true. It is usually stated using the multi-index notation, collecting all mixed partials that are equal.

DEFINITION 7.3. Let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$, with $\alpha_i \in \mathbb{N} \cup \{0\}$. If $h \in \mathbb{R}^d$ define

$$h^\alpha = h_1^{\alpha_1} h_2^{\alpha_2} \cdots h_d^{\alpha_d}, \quad |\alpha| = \alpha_1 + \cdots + \alpha_d, \quad \text{and} \quad \alpha! = \alpha_1! \, \alpha_2! \cdots \alpha_d!.$$

Given a $C^{|\alpha|}$ function f, define

$$D^\alpha f = \partial_1^{\alpha_1} \partial_2^{\alpha_2} \cdots \partial_d^{\alpha_d} f,$$

with the convention that $\partial_i^0 f = f$.

THEOREM 7.4. *If f is a C^n function on $\mathbb{R}^d$ and $a \in \mathbb{R}^d$ we have*

$$f(a+h) = \sum_{|\alpha|<n} \frac{1}{\alpha!} D^\alpha f(a) + R_n(h),$$

for some function R_n such that

$$\lim_{h\to 0} \frac{R_n(h)}{|h|^n} = 0.$$

The proof follows from the one variable Taylor's theorem in exactly the same as our second order version does, and collecting all mixed partials that are equal puts it in the above form.