Multivariate Real Analysis

Muchang Bahng

Spring 2025

Contents

In here, we extend the results of univariate real analysis to multivariate and/or vector-valued functions. In practice, multivariate calculus is used, and there are many new results that arise in the multivariate case. The case for continuity and convergence of multivariate functions is very straightforward, since these are topological properties. However, the definition of the derivative and the integral will need to be generalized.

In continuity, to prove that a limit is something, we just use ϵ-δ. However, to actually *compute* what the limit is, we have multiple ways to do this in practice.

1. Just compute the function assuming it is continuous.
2. Take some sort of path p and take the univariate limit.

We can also show that the multivariate limit doesn't exist by taking two sequences or two path functions where the limits do not equal each other.

1 Banach Spaces

Note that both the domain and codomain of a function $f : \mathbb{R}^n \to \mathbb{R}^m$ are vector spaces. However, Euclidean spaces have a lot of structure on them, and it is nice to identify the essential properties we need from these sets. It is essential that we work in vector spaces because when we define a derivative, we usually see some form that looks like

$$\frac{f(x+h) - f(x)}{h} \tag{1}$$

Note the operations used here. First, we want a notion of addition in the domain $(x + h)$ and the codomain $f(x + h) - f(x)$, along with some scalar multiplication when we multiply by $1/h$. A vector space precisely supports these operations and therefore is a natural choice. It is immediate that to define convergence, we definitely need a topology. We will see later that we want to define multivariate derivatives by adding a norm to this term, requiring the use of a normed vector space. Completion is clearly essential as we have seen in single-variable analysis.

Definition 1.1 (Banach Space)

A **Banach space** is a normed completed vector space.

Note that by extending the dimension, we have essentially lost the ordering $\leq$ on these spaces, along with the field properties. Therefore, we will need to adapt our definitions accordingly.

1.1 Coordinate Systems

Frenet frame?

2 Differentiation

Definition 2.1 (Frechet Derivative)

A function $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is **differentiable** at $x \in D$ if there exists a unique linear map $Df_x : \mathbb{R}^n \longrightarrow \mathbb{R}^m$—called the **total derivative** or **Frechet derivative**—such that

$$\lim_{h \to 0} \frac{\|f(x+h) - f(x) - Ah\|}{\|h\|} = 0 \tag{2}$$

If such a map Df_x exists, then it is unique.

Proof. The only claim to prove is uniqueness if defined.

The reason we want D to be open is that we require that $x + h$ to also be in D for sufficiently small h, and we can guarantee this since an ϵ-ball around x is guaranteed to be in D. Just as with the univariate case, the fundamental increment lemma also holds.

Lemma 2.1 (Fundamental Increment Lemma)

Suppose the derivative of $f : [a, b] \subset \mathbb{R}^n \to \mathbb{R}$ at x exists. Then, there exists a function $\Phi : \mathbb{R}^n \to \mathbb{R}$ such that

$$f(x+h) = f(x) + Df_x h + \Phi(h)\|h\| \tag{3}$$

for sufficiently small nonzero h, and

$$\lim_{h \to 0} \Phi(h) = 0 \tag{4}$$

Note that when $m = 1$, then Df_x is a linear functional, i.e. a dual vector. A simple way to extract derivatives is to fix all of the input components except for one, and treat f as a single-variable function. This results in—for now—a completely separate notion of a derivative.

Definition 2.2 (Directional, Partial Derivative)

The **directional derivative** of a multivariate function $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ at a point $\mathbf{a} \in D$ in direction $\mathbf{v} \in \mathbb{R}^n$ is the instantaneous rate of change of f when moving along direction $\mathbf{v}$ at $\mathbf{a}$. Formally,

$$\nabla_{\mathbf{v}} f(\mathbf{a}) \coloneqq \lim_{h \to 0} \frac{f(\mathbf{a} + h\mathbf{v})}{h}$$

When computing directional derivatives, it is convenient to normalize the directional vector $\mathbf{v}$ to be unit length so that it coincides with the partial derivatives. We don't technically need to set $||\mathbf{v}|| = 1$, but if we have two vectors $\mathbf{v}$ and a scaled $c\mathbf{v}$, then the directional derivatives will also be scaled ($\nabla_{c\mathbf{v}} f(\mathbf{a}) = c\nabla_{\mathbf{v}} f(\mathbf{a})$), so we will only work with unit directional vectors. Some say that this restriction is undesirable, since it loses the linearity of the function $\mathbf{v} \mapsto \partial_{\mathbf{v}} f(\mathbf{a})$.
If $\mathbf{v}$ is a unit basis vector $\mathbf{e}_i$, then we define this specific instance to be the **partial derivative** of f with respect to argument $\mathbf{x_i}$.

$$\partial_{x_i} f(\mathbf{a}) = \frac{\partial}{\partial x_i}\bigg|_{\mathbf{a}} f \coloneqq \lim_{h \to 0} \frac{f(\mathbf{a} + h\mathbf{e}_i)}{h}$$

which can be calculated by differentiating the function w.r.t. x_i and fixing all other variables. The partial derivative looks at the function as it is approaching $\mathbf{a}$ along an axis, while a directional derivative looks at the function as it is approaching from any direction in the domain.

Therefore, we have sort of three (or two) separate notions of derivatives. A natural question to ask is whether the existence of one derivative implies the existence of another.

Theorem 2.1 (Existence of Derivatives)

Let us have a function $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ and a point $\mathbf{a} \in D$.

1. If f is differentiable at $\mathbf{a} \in D$, then all of its directional derivatives exist. Furthermore, the total derivative $Df_{\mathbf{a}}$ applied to the directional unit vector $\mathbf{v}$ is equal to the directional derivative at $\mathbf{a}$ in direction $\mathbf{v}$.
$$Df_{\mathbf{a}}\mathbf{v} = \nabla_{\mathbf{v}} f(\mathbf{a}) \tag{5}$$
2. If all directional derivatives exist, then the partials exist (since we can just set the directional vectors to be the unit vectors).

Therefore, we know that differentiability (i.e. existence of the total derivative) is the strongest. If this is the case, then the partial and directional derivatives exist. It turns out that if we know the partial derivative

Furthermore, since our vector spaces come with a basis, we can realize Df_x in a matrix form. Furthermore, we can compute it quite easily!

Definition 2.3 (Jacobian)

Given $f : E \subset \mathbb{R}^n \to \mathbb{R}^m$ differentiable at x, the **Jacobian** of f at x is the matrix realization of the total derivative, denoted Jf_x.

Theorem 2.2 (Partial Derivatives as Image of Total Derivative)

The partial derivative equals the image of the basis vector
$$\frac{\partial f_j}{\partial x_i} = (Df_x e_i)_j \tag{6}$$

Corollary 2.1 (Directional Derivatives as Image of Total Derivative)

The directional derivative equals the image of the direction vector.
$$\nabla_v f(x) = (Df_x v)_j \tag{7}$$

Proof. By linearity.

Corollary 2.2 (Entries of Jacobian are Partials)

The entries of the Jacobian matrix are precisely the partial derivatives.

Proof. Immediate result of the previous corollary.

Analogous to the univariate case, a nice way to visualize the derivative is by looking at the tangent *plane*. In general, we have an *affine hyperplane*. Define this here. Unfortunately, the only types of functions that we can meaningful visualize are those mapping $f : \mathbb{R}^2 \to \mathbb{R}$.

Figure 1

So, to prove that a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is differentiable at $\mathbf{a}$, and if so, what its total derivative is, there are essentially two steps.

1. We find a candidate for M by evaluating the partials.

$$M = \begin{pmatrix} \partial_{x_1} f(\mathbf{a}) & \partial_{x_2} f(\mathbf{a}) & \dots & \partial_{x_n} f(\mathbf{a}) \end{pmatrix}$$

2. We check to see if the limit is true.

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) - M\mathbf{h}}{||\mathbf{h}||} = 0$$

If it is, then $Df_{\mathbf{a}} = M$, and the "tangent plane" of f at $\mathbf{a}$ is defined by the equation

$$y = f(\mathbf{a}) + M\mathbf{h}$$

Example 2.1 (Computing Total Derivative)

The function $f(x_1, x_2) = x_1^2 + x_2^2$ is differentiable at $(1,1)$. We let $M = \begin{pmatrix} \partial_{x_1} f(1,1), \partial_{x_2} f(1,1) \end{pmatrix} = \begin{pmatrix} 2, 2 \end{pmatrix}$ and see that

$$\begin{aligned} \lim_{\mathbf{h} \to \mathbf{0}} \frac{f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) - M\mathbf{h}}{||\mathbf{h}||} &= \lim_{\mathbf{h} \to \mathbf{0}} \frac{f(1+h_1, 1+h_2) - f(1,1) - 2h_1 - 2h_2}{\sqrt{h_1^2 + h_2^2}} \\ &= \lim_{\mathbf{h} \to \mathbf{0}} \frac{(1+h_1)^2 + (1+h_2)^2 - 2 - 2h_1 - 2h_2}{\sqrt{h_1^2 + h_2^2}} \\ &= \lim_{\mathbf{h} \to \mathbf{0}} \sqrt{h_1^2 + h_2^2} = 0 \end{aligned}$$

So, $Df(1,1) = (2,2)$.

Example 2.2 (Computing Tangent Plane)

Let us find the equation of the tangent plane to $f(x,y) = \ln(2x+y)$ at $(-1,3)$. Our total derivative, if it exists, is the covector of partials

$$Df = \begin{pmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{pmatrix} = \begin{pmatrix} \frac{2}{2x+y} & \frac{1}{2x+y} \end{pmatrix}$$

which is indeed continuous at a neighborhood of $(-1,3)$ (in fact, in every neighborhood not containing $\mathbf{0}$). By continuity of partials, f is differentiable at $(-1,3)$, with $Df_{(-1,3)} = (2,1)$. The equation of the plane is then

$$z = f(-1,3) + Df_{(-1,3)} \begin{pmatrix} x+1 \\ y-3 \end{pmatrix} = 2(x+1) + 1(y-3) \implies z = 2x + y - 1$$

However, the converse is not true for either statements. You cannot just evaluate all the partial derivatives and assume that the total derivative exists!

Example 2.3 (Existence of Directional Derivatives $\not\Rightarrow$ Differentiability)

The function $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$, defined

$$f(x_1, x_2) \coloneqq \begin{cases} 0 & \text{for } (x_1, x_2) = (0,0) \\ \frac{x_1^3}{x_1^2 + x_2^2} & \text{for } (x_1, x_2) \neq (0,0) \end{cases}$$

is not differentiable at $(0,0)$, but all directional derivatives exist. That is, for any (conventionally) unit $\mathbf{v} = (v_1, v_2)$, its directional derivative is always well defined to be

$$\nabla_{\mathbf{v}} f(0,0) = \lim_{h \to 0} \frac{\frac{h^3 v_1^3}{h^2(v_1^2+v_2)^2}}{h} = \frac{v_1^3}{v_1^2 + v_2^2} = v_1^3$$

Now assuming that there is such a linear M, we can find the partials by setting $\mathbf{v} = (1,0)$ and $\mathbf{v} = (0,1)$, giving

$$M = \begin{pmatrix} \partial_{x_1} f(0,0) & \partial_{x_2} f(0,0) \end{pmatrix} = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

But

$$\begin{aligned} \lim_{\mathbf{h} \to \mathbf{0}} \frac{f(a+h) - f(a) - Mh}{||\mathbf{h}||} &= \lim_{\mathbf{h} \to \mathbf{0}} \frac{f(h_1, h_2) - f(0,0) - 1 \cdot h_1 - 0 \cdot h_2}{\sqrt{h_1^2 + h_2^2}} \\ &= \lim_{\mathbf{h} \to \mathbf{0}} \frac{\frac{h_1^3}{h_1^2 + h_2^2} - h_1}{\sqrt{h_1^2 + h_2^2}} \\ &= \lim_{\mathbf{h} \to \mathbf{0}} -\frac{h_1 h_2^2}{(h_1^2 + h_2^2)^{3/2}} \end{aligned}$$

and taking along the path $\mathbf{h} = (k, k)$ gives

$$\lim_{(k,k) \to \mathbf{0}} -\frac{k^3}{(2k^2)^{3/2}} = -\frac{1}{2^{3/2}} \neq 0$$

Example 2.4 (Existence of Partials $\not\Rightarrow$ Existence of Directional Derivatives)

Consider the function

$$f(x_1, x_2) = \begin{cases} \frac{x_1 x_2}{x_1^2 + x_2^2} & \text{if } (x_1, x_2) \neq (0,0) \\ 0 & \text{if } (x_1, x_2) = (0,0) \end{cases}$$

The partial derivatives exist everywhere. Away from the origin we can simply compute

$$\begin{aligned} \frac{\partial f}{\partial x_1} &= \frac{x_2(x_1^2 + x_2^2) - x_1 x_2 \cdot 2x_1}{(x_1^2 + x_2^2)^2} = \frac{-x_1^2 x_2 + x_2^3}{(x_1^2 + x_2^2)^2} \\ \frac{\partial f}{\partial x_2} &= \frac{x_1(x_1^2 + x_2^2) - x_1 x_2 \cdot 2x_2}{(x_1^2 + x_2^2)^2} = \frac{x_1^3 - x_1 x_2^2}{(x_1^2 + x_2^2)^2} \end{aligned}$$

As for the partials at the origin, we must compute using the limit rule.

$$\begin{aligned} \partial_{x_1} f(\mathbf{0}) &= \lim_{h \to 0} \frac{f(\mathbf{0} + h\mathbf{e}_1) - f(\mathbf{0})}{h} = \lim_{h \to 0} \frac{f(h,0)}{h} = 0 \\ \partial_{x_2} f(\mathbf{0}) &= \lim_{h \to 0} \frac{f(\mathbf{0} + h\mathbf{e}_2) - f(\mathbf{0})}{h} = \lim_{h \to 0} \frac{f(0,h)}{h} = 0 \end{aligned}$$

However, the directional derivative taken in direction $\mathbf{v} = (1, 1)$ gives

$$\begin{aligned}\nabla_{(1,1)} f(0,0) &= \lim_{h \to 0} \frac{f(\mathbf{0} + h(1,1)) - f(\mathbf{0})}{h} \\ &= \lim_{h \to 0} \frac{f(h,h)}{h} \\ &= \lim_{h \to 0} \frac{1}{2h}\end{aligned}$$

which does not have a limit as $h \to 0$. To visualize this, let's look at the values of f along various lines in $\mathbb{R}^2$.

1. $f = 0$ at the line $x_1 = 0$ and $x_2 = 0$, which is why the partials are 0.
2. $f = \frac{1}{2}$ at the line where $x_1 = x_2$, except for the point $(0, 0)$, where $f = 0$, which is why the limit in the direction doesn't exist.

2.1 Rules of Differentiation

Just like single variable calculus, the total derivative behaves in predictable ways: it is linear, product/quotient rules, and the chain rule.

Theorem 2.3 (Linearity of Total Derivatives)

Let $f, g : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be differentiable at $\mathbf{a} \in D$. Then, the total derivative at $\mathbf{a}$ is linear w.r.t. the function arguments.

1. $D(f + g)_{\mathbf{a}} = Df_{\mathbf{a}} + Dg_{\mathbf{a}}$
2. $D(cf)_{\mathbf{a}} = cDf_{\mathbf{a}}$

Furthermore, if f and g are differentiable over D, then

1. $D(f + g) = Df + Dg$
2. $D(cf) = cDf$

Note that for the product and quotient rules, our scope is only for scalar valued functions.

Theorem 2.4 (Product Rule)

Let $f, g : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ be differentiable at $\mathbf{a}$. Then,

$$D(fg)_{\mathbf{a}} = Df_{\mathbf{a}} g(\mathbf{a}) + f(\mathbf{a}) Dg_{\mathbf{a}} \tag{8}$$

If f, g are differentiable over D, then

$$D(fg) = Df \cdot g + f \cdot Dg$$

Theorem 2.5 (Quotient Rule)

Let $f, g : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ be differentiable at $\mathbf{a}$ with $g(\mathbf{a}) \neq 0$. Then,

$$D\Big(\frac{f}{g}\Big)_{\mathbf{a}} = \frac{Df_{\mathbf{a}} g(\mathbf{a}) - f(\mathbf{a}) Dg_{\mathbf{a}}}{g(\mathbf{a})^2}$$

If f, g are differentiable over D and g never vanishes on D, then

$$D\Big(\frac{f}{g}\Big) = \frac{Df \cdot g - f \cdot Dg}{g^2}$$

Theorem 2.6 (Chain Rule)

Let $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ and $g : E \subset \mathbb{R}^p \longrightarrow \mathbb{R}^n$ be two functions such that $f \circ g : E \subset \mathbb{R}^p \longrightarrow \mathbb{R}^m$ is defined on E. Suppose g is differentiable at $\mathbf{a} \in E$ and f is differentiable at $g(\mathbf{a}) \in D$. Then, $f \circ g$ is differentiable at $\mathbf{a}$, and

$$D(f \circ g)_{\mathbf{a}} = Df_{g(\mathbf{a})} \circ Dg_{\mathbf{a}}$$

If g is differentiable over E and f over $g(E) \subset D$, then $f \circ g$ is differentiable over E, and

$$D(f \circ g)(\cdot) = Df_{g(\cdot)} \circ Dg_{(\cdot)}$$

Therefore, given the composition of function $f \circ g$, we have two methods of finding the derivative matrix of $f \circ g$ at point x_0. First is to explicitly compute $f \circ g$ and find its $m \times p$ derivative matrix $D(f \circ g)$, and plug in $\mathbf{a}$ to get $D(f \circ g)_{\mathbf{a}}$. The second way is to use the chain rule to find the individual total derivatives $Df_{g(\mathbf{a})}$ and $Dg_{\mathbf{a}}$, and multiply them together.

2.2 Continuously Differentiable Functions

An even stronger condition beyond differentiability is continuous partials, and we often prove continuity of partials to prove differentiability.

Theorem 2.7 (Continuous Partials $\implies$ Differentiability)

Given a function $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ and a point $\mathbf{a} \in D$, if all the partials $\partial_{x_i} f$ exist and are continuous at $\mathbf{a}$, then f is differentiable at $\mathbf{a}$.

Example 2.5 (Differentiability $\not\implies$ Continuous Partials)

The function

$$g(x) \equiv \begin{cases} x^2 \sin\left(\frac{1}{x}\right) & x \neq 0 \\ 0 & x = 0 \end{cases}$$

is differentiable, with derivative at $x = 0$ to be $g'(0) = 0$, since $g(h)$ is bounded by h^2.

$$\lim_{h \to 0} \frac{h^2 \sin\left(\frac{1}{h}\right) - 0}{h} \leq \lim_{h \to 0} \frac{h^2}{h} = 0$$

which makes

$$g'(x) \equiv \begin{cases} -\cos\left(\frac{1}{x}\right) + 2x \sin\left(\frac{1}{x}\right) & x \neq 0 \\ 0 & x = 0 \end{cases}$$

But because $\cos(\frac{1}{x})$ oscillates at $x \to 0$, $g'(x)$ is not continuous at $x = 0$. Therefore $g(x)$ is differentiable but not in $C^1(\mathbb{R})$.

Definition 2.4 (C^1 Space)

The vector space of all functions $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ with continuous partials is denoted $C^1(D; \mathbb{R})$ or $C^1(D)$. They are called **continuously differentiable**.

We can also visualize this theorem. Since the partials are continuous, then the tangent subspace, which is determined by the span of the tangent vectors determined by the partials, also changes continuously, and therefore the total derivative within a neighborhood of $\mathbf{a}$ exists. Note that from now, whenever we talk

about differentiating a function f, we will assume that it is C^1. This is overkill, since the set of all k-times differentiable functions is a subset of C^k, but it is conventional to work with C^k functions.

2.3 Extrema and Concavity

Definition 2.5 (Local Extrema)

Given a function $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$, a point $\mathbf{x}_0 \in D$ is a *local minimum* if there exists a neighborhood U of $\mathbf{x}_0$ such that

$$f(\mathbf{x}) \geq f(\mathbf{x}_0) \text{ for every } \mathbf{x} \in U$$

Similarly, $\mathbf{x}_0$ is a *local maximum* if there exists a neighborhood U of $\mathbf{x}_0$ such that

$$f(\mathbf{x}) \leq f(\mathbf{x}_0) \text{ for every } \mathbf{x} \in U$$

Theorem 2.8 (1st Derivative Test)

If $\mathbf{x}_0$ is a local extremum of a differentiable function f, then $Df_{\mathbf{x}_0} = \mathbf{0}$. That is, $\mathbf{x}_0$ is a critical point of f, i.e. every directional derivative through $\mathbf{x}_0$ is 0.

Note that even though the converse of this theorem is not true, we can use the contrapositive to determine that every point that has a nonzero derivative cannot be a extremum. A function may also have an infinite amount of critical points (e.g. if they lie in a circle). In order to determine whether a critical point $\mathbf{x}_0$ is a relative maximum, minimum, or neither, we use the second derivative test.

Theorem 2.9 (2nd Derivative Test)

Let $\mathbf{x}_0$ be a critical point of C^2 function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$. That is, $Df_{\mathbf{x}_0} = \mathbf{0}$. Then,

1. $\mathbf{x}_0$ is a local minimum if $Hf_{\mathbf{x}_0}$ is positive definite.
2. $\mathbf{x}_0$ is a local maximum if $Hf_{\mathbf{x}_0}$ is negative definite.
3. $\mathbf{x}_0$ is a *saddle point* if $Hf_{\mathbf{x}_0}$ is not positive definite nor negative definite.

Visually, this makes sense since given a critical point $\mathbf{x}_0$, the derivative matrix would be 0, meaning that the 2nd degree Taylor expansion of f near $\mathbf{x}_0$ would be in form

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T Hf_{\mathbf{x}_0}(\mathbf{x} - \mathbf{x}_0)$$

If $Hf_{\mathbf{x}_0}$ is positive definite, then by definition $\frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T Hf_{\mathbf{x}_0}(\mathbf{x}-\mathbf{x}_0) > 0$ for all x near $\mathbf{x}_0$, and so f would increase in every direction within the neighborhood of $\mathbf{x}_0$. The logic follows similarly for negative definite matrix $Hf_{\mathbf{x}_0}$. If $Hf_{\mathbf{x}_0}$ is not positive nor negative definite, then $\frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T Hf_{\mathbf{x}_0}(\mathbf{x}-\mathbf{x}_0)$ could be positive or negative, depending on which direction vector $\mathbf{h} = \mathbf{x}-\mathbf{x}_0$ we choose for computing the directional derivative. Therefore, f will increase for certain $\mathbf{h}$ and decrease for other $\mathbf{h}$.

Definition 2.6 (Global Extrema)

Given $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$, a point $\mathbf{x}_0 \in A$ is said to be an *absolute, or global, maximum* if

$$f(\mathbf{x}_0) \geq f(x) \text{ for all } \mathbf{x} \in D$$

and a *global minimum* if

$$f(\mathbf{x}_0) \leq f(x) \text{ for all } \mathbf{x} \in D$$

Unfortunately, determining whether a point $\mathbf{x}_0$ is a local extremum requires us to define an open neighborhood

around $\mathbf{x}_0$. This means that we can only determine local extrema within open sets in $\mathbb{R}^n$. Therefore, we must modify our procedure when looking for extrema on functions defined over closed bounded sets. We now describe a method of computing the global extrema. Let $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ be a multivariable function defined on a closed and bounded set $D \equiv U \cup \partial U$, where U is open and ∂U is the boundary of D. To find the global extrema on D, we find all local extrema of

1. f defined over the interior of U, which is open, by locating all points where $Df = \mathbf{0}$
2. f defined over ∂U, preferably defined as a composition of the path function $p : I \subset \mathbb{R} \longrightarrow \partial U$ and $f : \partial U \longrightarrow \mathbb{R}$. That is, find the values of t such that $D(f \circ p)(t) = 0$ and identify $p(t)$.

We take all these critical points and choose the largest to be the global maximum and the smallest to be the global minimum.

Definition 2.7 (Convex Set)

A subset $D \subset \mathbb{R}^n$ is a **convex set** if for any two points $\mathbf{x}, \mathbf{y} \in D$, the line segment joining them is also in D. That is,

$$\{\theta\mathbf{x} + (1-\theta)\mathbf{y} \mid 0 \leq \theta \leq 1\} \subset D$$

Definition 2.8 (Convex Function)

Let D be a convex set. A function $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ is a **convex function** if

$$f\big(\theta\mathbf{x} + (1-\theta)\mathbf{y}\big) \leq \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y})$$

We can visualize this as the graph of the function in $D \oplus \mathbb{R}$ always being "under" every line segment connecting $\big(\mathbf{x}, f(\mathbf{x})\big)$ and $\big(\mathbf{y}, f(\mathbf{y})\big)$.

If we assume that f is C^1 or C^2, we can use additional tools to prove convexity. The theorem for C^1 functions is quite intuitive, since for a convex function, the tangent plane on its graph must never be "above" the graph. In other words, the first order approximation must be a global underestimate of f.

Theorem 2.10 (Convexity of C^1 Functions)

Let D be a convex set and $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ be C^1. Then, f is convex over D if and only if

$$f(\mathbf{x}_0) + Df_{\mathbf{x}_0}(\mathbf{x} - \mathbf{x}_0) \leq f(\mathbf{x})$$

for all $\mathbf{x}_0, \mathbf{x} \in D$.

Theorem 2.11 (Convexity of C^2 Functions)

Let D be a convex set and $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ be C^2. Then, f is convex over D if and only if Hf is positive semidefinite over all interior points of D (i.e. all eigenvalues of $Hf_{\mathbf{a}}$ are nonnegative for all $\mathbf{a} \in D$).

The computation of the Hessian now gives us much more information about the graph of the function of interest.

Theorem 2.12 ()

A function $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ defined on a convex set D is convex if and only if its Hessian matrix Hf is positive semidefinite for all $\mathbf{x} \in D$.

Once we have computed the Hessian, let's take the eigendecomposition of it. Since $Hf_{\mathbf{a}}$ is a real symmetric matrix, by the spectral theorem, it will have n real eigenvalues $\lambda_1, \dots, \lambda_n$ (in descending values) and corresponding orthonormal eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_n$. Now given the gradient $\nabla f(\mathbf{a})$ at $\mathbf{a}$, we can approximate $\nabla f(\mathbf{a}+\mathbf{h})$ at $\mathbf{a}+\mathbf{h}$ using its total derivative as

$$\nabla f(\mathbf{a}+\mathbf{h}) \approx \nabla f(\mathbf{a}) + D\nabla f_{\mathbf{a}}\mathbf{h} = \nabla f(\mathbf{a}) + Hf(\mathbf{a})\mathbf{h}$$

The eigenvalues of $Hf(\mathbf{a})$ will tell us how "fast" the gradient changes at $\mathbf{a}$. That is, given a small displacement vector $\mathbf{h}$, we can take an orthonormal decomposition of it in the form

$$\mathbf{h} = \sum_i h_i \mathbf{v}_i$$

and now the approximate gradient can be written as

$$\nabla f(\mathbf{a}+\mathbf{h}) = \nabla f(\mathbf{a}) + \sum_i h_i \lambda_i \mathbf{v}_i$$

Therefore, bigger λ_i's will contribute to a greater change in $f(\mathbf{a})$, and smaller ones will contribute less. We can use this information to speed up convergence by scaling along different axes of $\mathbf{h}$ when sampling.

2.4 Optimization with Lagrange Multipliers

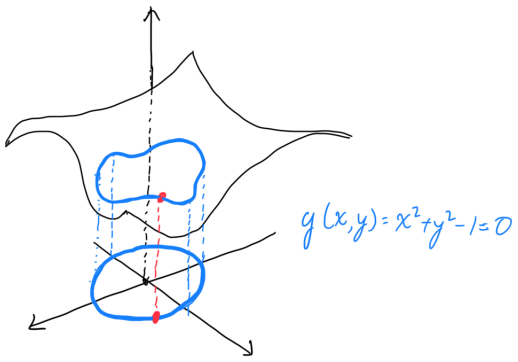
In many cases we are required to find the local extrema of a function $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ subject to a system of equality constraints (i.e. subject to the condition that one of more equations have to be satisfied exactly by the chosen values of the variables) of the form:

$$g_1(\mathbf{x}) = 0, g_2(\mathbf{x}) = 0, \dots, g_c(\mathbf{x}) = 0$$

which can be summarized into the constraint $\mathbf{g} : \mathbb{R}^n \longrightarrow \mathbb{R}^c$

$$\mathbf{g}(x) = \begin{pmatrix} g_1(x) \\ \vdots \\ g_c(x) \end{pmatrix} = \mathbf{0}$$

which really just represents a level set of $\mathbf{g}$ at $\mathbf{0}$, i.e. a set described by an implicit representation. In physics, these types of "well-behaved" constraints are known as *holonomic constraints*. Here is an example of a function $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ constrained to the unit circle, where $g(x, y) = x^2 + y^2 - 1 = 0$.



To solve this constraint problem, we use the method of Lagrange multipliers. The basic idea is to convert a constrained problem into a form such that the derivative test of an unconstrained problem can still be applied. The relationship between the gradient of the function and gradients of the constraints rather naturally leads to a reformulation of the original problem, known as the *Lagrangian function*. That is, in order to find the maximum/minimum of f subjected to the equality constraint $\mathbf{g}(\mathbf{x}) = \mathbf{0}$, we form the Lagrangian function

$$\mathcal{L}(\mathbf{x}, \lambda) \equiv f(\mathbf{x}) - \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})$$

and find the stationary points of $\mathcal{L}$ considered as a function of $\mathbf{x} \in D$ and the Lagrange multiplier $\lambda \in \mathbb{R}$. The main advantage to this method is that it allows the optimization to be solved without explicit parameterization in terms of the constraints.

Theorem 2.13 (Lagrange Multipliers Theorem)

Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a C^1 function and let $g(\mathbf{x}) = 0$, where $g : \mathbb{R}^n \longrightarrow \mathbb{R}^c$, be a system of C^1 constraint equations: $g \coloneqq (g_1, g_2, \dots, g_c)$. Let $\mathbf{x}^*$ be an optimal solution to the optimization problem of maximizing $f(\mathbf{x})$ subject to the constraint $g(\mathbf{x}) = 0$ such that $\operatorname{rank} Dg_{\mathbf{x}^*} = c < n$. Then, their exists a unique vector $\boldsymbol{\lambda}^*$ of Lagrange multipliers $\lambda_1^*, \lambda_2^*, \dots, \lambda_c^*$ s.t.

$$Df_{\mathbf{x}^*} = \lambda^{*T} Dg_{\mathbf{x}^*}$$

where $Df_{\mathbf{x}^*}$ can be interpreted as the $1 \times n$ Jacobian matrix of f and $Dg_{\mathbf{x}^*}$ as the $c \times n$ Jacobian of g. Since both $Df_{\mathbf{x}^*}$ and $\lambda^{*T} Dg_{\mathbf{x}^*}$ are maps from $\mathbb{R}^n$ to $\mathbb{R}$, we can invoke Riesz representation theorem to turn this into gradients:

$$\nabla f(\mathbf{x}^*) = \nabla \mathbf{g}(\mathbf{x}^*)(\boldsymbol{\lambda}^*)$$

which has a matrix realization of

$$\begin{aligned}
\begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}^*) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}^*) \end{pmatrix} &= \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(\mathbf{x}^*) & \dots & \frac{\partial g_c}{\partial x_1}(\mathbf{x}^*) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_1}{\partial x_n}(\mathbf{x}^*) & \dots & \frac{\partial g_c}{\partial x_n}(\mathbf{x}^*) \end{pmatrix} \begin{pmatrix} \lambda_1^* \\ \vdots \\ \lambda_c^* \end{pmatrix} \\
&= \lambda_1^* \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(\mathbf{x}^*) \\ \vdots \\ \frac{\partial g_1}{\partial x_n}(\mathbf{x}^*) \end{pmatrix} + \lambda_2^* \begin{pmatrix} \frac{\partial g_2}{\partial x_1}(\mathbf{x}^*) \\ \vdots \\ \frac{\partial g_2}{\partial x_n}(\mathbf{x}^*) \end{pmatrix} + \dots + \lambda_c^* \begin{pmatrix} \frac{\partial g_c}{\partial x_1}(\mathbf{x}^*) \\ \vdots \\ \frac{\partial g_c}{\partial x_n}(\mathbf{x}^*) \end{pmatrix}
\end{aligned}$$

This equation tells us that at any critical points $\mathbf{x}^*$ of f evaluated under the equality constraints, the gradient of f at $\mathbf{x}^*$ can be expressed as a unique linear combination of the gradients of the constraints $\nabla g_i(\mathbf{x}^*)$ (at $\mathbf{x}^*$), with the Lagrange multipliers acting as coefficients. Therefore, finding the critical points $\mathbf{x}^*$ of f constrained with g is equivalent to solving the system of $c + n$ equations for the n unknowns in $\mathbf{x}$ and c unknowns in $\boldsymbol{\lambda}$:

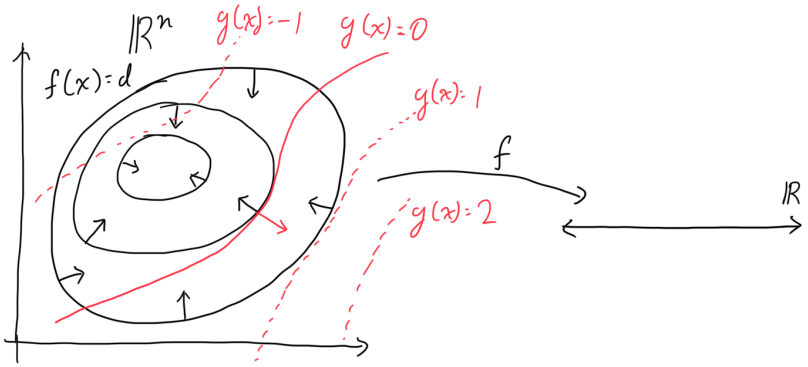
$$\begin{aligned}
\mathbf{g}(\mathbf{x}) &= 0 \\
\nabla f(\mathbf{x}) &= \nabla \mathbf{g}(\mathbf{x})(\boldsymbol{\lambda})
\end{aligned}$$

which can be rewritten as

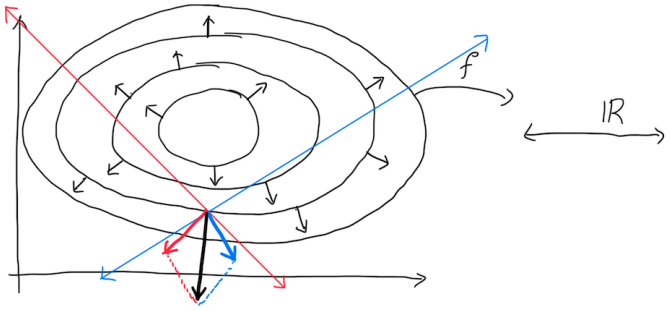
$$\begin{aligned}
c \text{ constraint equations} &\begin{cases} g_1(x) & = 0 \\ \dots & = 0 \\ g_c(x) & = 0 \end{cases} \\
n \text{ Lagranaian equations} &\begin{cases} \frac{\partial f}{\partial x_1}(\mathbf{x}^*) & = \lambda_1^* \frac{\partial g_1}{\partial x_1}(\mathbf{x}^*) + \lambda_2^* \frac{\partial g_2}{\partial x_1}(\mathbf{x}^*) + \dots + \lambda_c^* \frac{\partial g_c}{\partial x_1}(\mathbf{x}^*) \\ \dots & = \dots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}^*) & = \lambda_1^* \frac{\partial g_1}{\partial x_n}(\mathbf{x}^*) + \lambda_2^* \frac{\partial g_2}{\partial x_n}(\mathbf{x}^*) + \dots + \lambda_c^* \frac{\partial g_c}{\partial x_n}(\mathbf{x}^*) \end{cases}
\end{aligned}$$

More abstractly, $Df_{\mathbf{x}^*}$ is the linear functional in $(\mathbb{R}^n)^*$, and $Dg_{\mathbf{x}^*}$, which is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^c$, can be interpreted as a map from $(\mathbb{R}^c)^*$ to $(\mathbb{R}^n)^*$ Since $\boldsymbol{\lambda}^*$ "lives" in $(\mathbb{R}^c)^*$, $Dg_{\mathbf{x}^*}(\boldsymbol{\lambda}^*) \in (\mathbb{R}^n)^*$, which is the same space that $f_{\mathbf{x}^*}$ lives in.

Let us introduce a visualization for when where is a single constraint $g : \mathbb{R}^n \longrightarrow \mathbb{R}$. From the properties of the gradient, $\nabla f(x_0)$ is orthogonal to the level set of points satisfying $f(x) = f(x_0)$ at point x_0. Note that the constraint function g also maps $\mathbb{R}^n \longrightarrow \mathbb{R}$, and so it has its own level surfaces. We can see that the point where the contour line of $g(x) = 0$ tangentially touches the contours of f is the maximum. Since it intersects it tangentially, the gradient vector at that point $\nabla g(x_0)$ is parallel to $\nabla f(x_0)$.



We can visualize this for multiple constraints as well, where $\nabla f(x_0)$ (the gradient vector of f at x^*) can be expressed as a linear combination of $\nabla g_1(x_0)$ and $\nabla g_2(x_0)$ (gradient vectors of the constraint functions at x^*).



From the properties of the gradient introduced before, $\triangledown f(x_0)$ is orthogonal to the level set of points satisfying $f(x) = c$ at the point x_0. But this level set $f(x) = c$ actually intersects the level set determined by $g(x) = c$ at the point x_0 and is indistinguishable from each other at x_0. This means that $\triangledown g(x_0)$ is normal the level set of $g(x) = c$ at $x_0 \iff$ it is normal to the level set of $f(x) = c$ at x_0. But $\triangledown f(x_0)$ is also normal at that point, so $\triangledown f(x_0)$ must be parallel to $\triangledown g(x_0)$.

2.5 Inverse and Implicit Function Theorems

Theorem 2.14 (Inverse Function Theorem for Multivariable Functions and its Matrix Realization)

Let $\mathbf{f} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be a C^1 function defined on an open neighborhood of $\mathbf{x}_0$ in the domain. If the total derivative/Jacobian $D\mathbf{f}_{\mathbf{x}_0}$ at $\mathbf{x}_0$ is invertible, an inverse function of $\mathbf{f}$ is defined on some neighborhood

of $\mathbf{y}_0 = \mathbf{f}(\mathbf{x}_0)$. Given that we are working with a fixed basis, $\mathbf{f}$ can be modeled by the set of n equations

$$\begin{aligned} f_1(x_1, x_2, \dots, x_n) &= y_1 \\ \dots &= \dots \\ f_2(x_1, x_2, \dots, x_n) &= y_2 \end{aligned}$$

This theorem says that this system of n equations has a unique solution for $x_1, x_2, \dots, x_n$ in terms of $y_1, \dots, y_n$, provided that we restrict $\mathbf{x}$ and $\mathbf{y}$ to small enough neighborhoods of $\mathbf{x}_0$ and $\mathbf{y}_0$. This inverse function $\mathbf{f}^{-1} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is also C^1, and its total derivative/Jacobian $D\mathbf{f}_{\mathbf{y}_0}^{-1}$ at $\mathbf{y}_0 = \mathbf{f}(\mathbf{x}_0)$ is the inverse linear map of $D\mathbf{f}_{\mathbf{y}_0}$.

$$D\mathbf{f}_{\mathbf{y}_0}^{-1} = \left(D\mathbf{f}_{\mathbf{x}_0}\right)^{-1}$$

Example 2.6 ()

Consider the vector-valued function $\mathbf{f} : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ defined by

$$\mathbf{f}(x_1, x_2) = \begin{pmatrix} e^{x_1} \cos(x_2) \\ e^{x_1} \sin(x_2) \end{pmatrix}$$

The total derivative/Jacobian matrix is

$$J\mathbf{f}(x_1, x_2) = \begin{pmatrix} e^{x_1} \cos(x_2) & -e^{x_1} \sin(x_2) \\ e^{x_1} \sin(x_2) & e^{x_1} \cos(x_2) \end{pmatrix} \implies \det J\mathbf{f}(x_1, x_2) = e^{2x_1} \cos^2(x_2) + e^{2x_1} \sin^2(x_2) = e^{2x_1}$$

Since the determinant e^{2x_1} is nonzero everywhere, $D\mathbf{f}_{\mathbf{x}}$ is nonsingular. Thus, the theorem guarantees that for every point $\mathbf{x}_0 \in \mathbb{R}^2$, there exists a neighborhood about $\mathbf{x}_0$ over which $\mathbf{f}$ is invertible. However, this does not mean $\mathbf{f}$ is invertible over its entire domain: in this case $\mathbf{f}$ isn't even injective since it is periodic: e.g. the preimage of $(e, 0)$ contains $(1, 0)$ and $(1, 2\pi)$.

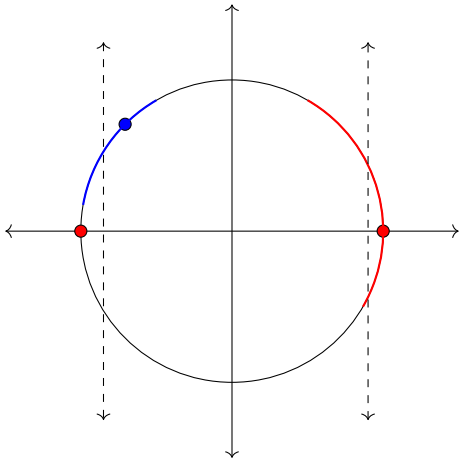
Remember that given an explicit representation of a set $\mathbf{y} = \mathbf{f}(\mathbf{x})$, we can easily find the implicit form as $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{y} - \mathbf{f}(\mathbf{x}) = \mathbf{0}$. What about the other way around? That is, given an implicit representation of some surface, what conditions must be met so that it can be represented as the graph of a function? The implicit function theorem is a tool that allows relations between points in $\mathbb{R}^n$ to be converted to functions of several real variables. That is, it states that for sufficiently "nice" points on a n-dimensional surface defined as $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ (where $\mathbf{F} : \mathbb{R}^{n+m} \longrightarrow \mathbb{R}^m$), we can locally pretend that this surface is a graph of a function $\mathbf{g} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ whose graph $\big(\mathbf{x}, \mathbf{g}(\mathbf{x})\big)$ is precisely the set of all $(\mathbf{x}, \mathbf{y})$ s.t. $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$. When $m = 1$, it basically states that if an implicit surface suffices the vertical line test in a neighborhood, then it can be written as a function.

Example 2.7 (Circle)

Let $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ be defined by $f(x, y) = x^2 + y^2 - 1$. The level set at $z = 0$ would be the set of points satisfying

$$x^2 + y^2 - 1 = 0$$

the unit circle. The derivative of f with respect to y is 0 at the points $(-1, 0)$ and $(1, 0)$, meaning that in any neighborhood of these points, we cannot define a function of y with respect to x. This is true, indeed, since any such function would fail the vertical line test, which can be seen in the red neighborhood around $(1, 0)$. However, the blue neighborhood of the point $(-\sqrt{2}/2, \sqrt{2}/2)$ does indeed define a function of y with respect to x satisfying the vertical line test.



Theorem 2.15 (Special Implicit Function Theorem)

Let $f : \mathbb{R}^{n+1} \longrightarrow \mathbb{R}$ be a C^1 function with a point $(\mathbf{a}, b) \in \mathbb{R}^{n+1}$ on the level set $f(\mathbf{x}, y) = 0$. If $\partial_y f(\mathbf{a}, b)$, which can also be thought of as the 1×1 truncated Jacobian matrix $D_{\mathbf{y}} f_{(\mathbf{a},b)} = \partial_y f(\mathbf{a}, b)$ w.r.t. y, of

$$Df_{(\mathbf{a},b)} = \begin{pmatrix} D_{\mathbf{x}} f_{(\mathbf{a},b)} & D_{\mathbf{y}} f_{(\mathbf{a},b)} \end{pmatrix} = \left(\begin{array}{ccc|c} \partial_{x_1} f(\mathbf{a}, b) & \dots & \partial_{x_n} f(\mathbf{a}, b) & \partial_y f(\mathbf{a}, b) \end{array} \right)$$

is invertible (in this case nonzero), then there exists an open neighborhood $U_{\mathbf{a}} \subset \mathbb{R}^n$ of $\mathbf{a}$ and a unique C^1 function $y : U \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ s.t. $f\big(\mathbf{x}, y(\mathbf{x})\big) = 0$ for all $\mathbf{x} \in U$. That is, we can find a $y : U \longrightarrow \mathbb{R}$ s.t. the graph $y(\mathbf{x})$ within U coincides with the graph of $f(\mathbf{x}, y) = 0$. Moreover, the total derivative/Jacobian of $y : \mathbb{R}^n \longrightarrow \mathbb{R}$ in U is the $1 \times n$ matrix given by the matrix product

$$Dg_{\mathbf{a}} = -(D_y f_{\mathbf{a}})^{-1} D_{\mathbf{x}} f_{\mathbf{a}}$$

Example 2.8 (Circle Example)

Let $n = m = 1$ and $f(x_1, x_2) = x_1^2 + x_2^2 - 1$. We would like to find out at which points $\mathbf{a}$ can this surface be explicitly represented by a function $g : U_{\mathbf{a}} \subset \mathbb{R} \longrightarrow \mathbb{R}$ defining x_2 from x_1. Its Jacobian is

$$Df = \begin{pmatrix} \partial_{x_1} f & \partial_{x_2} f \end{pmatrix} = \begin{pmatrix} 2x_1 & 2x_2 \end{pmatrix}$$

The truncated Jacobian w.r.t. x_2 is $2x_2$, which is invertible iff $x_2 \neq 0$. By the implicit function theorem, we can locally write the circle in the form $x_2 = g(x_1)$ for all points where $x_2 \neq 0$. This is easy to see. For example, we can choose the point $(0.8, 0.6)$ on the level set, and the appropriate explicit function is

$$x_2 = g(x_1) = \sqrt{1 - x_1^2}$$

within the neighborhood of $x_1 = 0.8$. For $(\pm 1, 0)$, we cannot since every function defined within a neighborhood of $x_1 = \pm 1$ fails the vertical line test. The derivative of g, by the theorem, can be defined implicitly as

$$Dg = -(\partial_{x_2} f)^{-1} D_{x_1} f = -(2x_2)^{-1}(2x_1) = -\frac{x_1}{x_2}$$

which leads to the differential equation

$$g'(x_1) = -\frac{x_1}{g(x_1)} \text{ where we solve for } g$$

If we would have liked to find a function $h : U_{a_{\mathbf{x}_2}} \subset \mathbb{R} \longrightarrow \mathbb{R}$ defining x_1 from x_2, then we can redo everything to find that the truncated Jacobian w.r.t. x_1 is $2x_1$, which is invertible iff $x_1 \neq 0$, and the the derivative is

$$Dh = -(\partial_{x_1} f)^{-1} D_{x_2} f = -(2x_1)^{-1}(2x_2) = -\frac{x_2}{x_1}$$

which leads to the differential equation

$$h'(x_2) = -\frac{x_2}{h(x_2)} \text{ where we solve for } h$$

Theorem 2.16 (General Implicit Function Theorem)

Let $f : \mathbb{R}^{n+m} \longrightarrow \mathbb{R}^m$ be a C^1 function with a point $(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^{n+m}$ on the level set $f(\mathbf{x}, \mathbf{y}) = 0$. If the $m \times m$ truncated Jacobian matrix $D_{\mathbf{y}} f(\mathbf{a}, \mathbf{b})$ w.r.t. $\mathbf{y}$, of

$$Df_{(\mathbf{a},\mathbf{b})} = \begin{pmatrix} D_{\mathbf{x}} f_{(\mathbf{a},\mathbf{b})} & D_{\mathbf{y}} f_{(\mathbf{a},\mathbf{b})} \end{pmatrix}$$

is invertible, then there exists an open neighborhood $U_{\mathbf{a}} \subset \mathbb{R}^n$ of $\mathbf{a}$ and a unique C^1 function $\mathbf{y} : U \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ s.t. $f\big(\mathbf{x}, \mathbf{y}(\mathbf{x})\big) = 0$ for all $\mathbf{x} \in U$. That is, we can find a $\mathbf{y} : U \longrightarrow \mathbb{R}^m$ s.t. the graph $\mathbf{y}(\mathbf{x})$ within U coincides with the graph of $f(\mathbf{x}, \mathbf{y}) = 0$. Moreover, the total derivative/Jacobian of $\mathbf{y}$ is the $m \times n$ matrix given by the matrix product

$$Dg_{\mathbf{a}} = -(D_{\mathbf{y}} f_{\mathbf{a}})^{-1} D_{\mathbf{x}} f_{\mathbf{a}}$$

3 Higher Order Derivatives

Redefine derivatives by talking about tensors.

Definition 3.1 ()

Definition of derivatives again with tensors.

Since $\nabla_{\mathbf{v}} f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$, we can take the directional derivative (assuming it exists) of it again in direction $\mathbf{u}$ to get a second derivative $\nabla_{\mathbf{u}} \nabla_{\mathbf{v}} f$. We usually work with iterated partial derivatives, and we can compute derivatives as many times as we want, given that they exist. Therefore, the second-order iterated partial derivatives of f are

$$\partial_{x_i x_j} \coloneqq \partial_{x_j} \partial_{x_i} f \text{ for } i, j = 1, \dots n$$

Definition 3.2 (C^k Functions)

A function $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ is said to be a C^k function if all k-times iterated partial derivatives

$$\partial_{x_{i_1} x_{i_2} \dots x_{i_k}} f$$

exist and are continuous. The vector space of all C^k functions is denoted $C^k(D; \mathbb{R})$, or $C^k(D)$.

Whenever we want to get the kth iterated partial derivative of f, we will assume that $f \in C^k$. Again, this is overkill, but it is conventional since we don't really work with the set of functions with existing partial derivatives.

Theorem 3.1 (Nested C^k and $\mathcal{D}^k$ Function Spaces)

Let the space of all k-times differentiable functions over $\mathbb{R}^n$ be denoted $\mathcal{D}(\mathbb{R}^n)$. Then,

$$C^0(\mathbb{R}^n) \supset \mathcal{D}^1(\mathbb{R}^n) \supset C^1(\mathbb{R}^n) \supset \mathcal{D}^2(\mathbb{R}^n) \supset C^2(\mathbb{R}^n) \dots \mathcal{D}^k(\mathbb{R}^n) \supset C^k(\mathbb{R}^n) \dots C^\infty(\mathbb{R}^n)$$

Note that mathematicians throw around the word "smooth" a lot. Usually, it means one of three things

1. it is of class C^1
2. it is of class C^∞
3. it is of class C^k, where k is however high it needs to be to satisfy our assumptions. For example, if I say let us differentiate smooth f two times, then I am assuming that $f \in C^2(\mathbb{R}^n)$.

Visualizing C^k-functions is easy for low orders. A C^0 function produces a graph that isn't "ripped" or "punctured," since this is exactly what a discontinuity would look like. A C^1 function requires the surface to be smooth in such a way that there is a well defined affine tangent subspace at every point. This means that there cannot be any sharp "points" or "edges" on the graph since a tangent subspace cannot be well defined.

Theorem 3.2 (Clairut's Theorem)

Given $f \in C^2$ at point $\mathbf{a}$, its second iterated partials are equal.

$$\partial_{x_i x_j} f(\mathbf{a}) = \partial_{x_j x_i} f(\mathbf{a}) \text{ for } i, j = 1, 2, \dots, n$$

Proof. For clarity, denote x_i, x_j as x, y and ignore the rest of the variables. Then, the partial derivatives $\partial_{xy} f$ and $\partial_{yx} f$ at a point (x_0, y_0) can be expressed as double limits:

$$\partial_{xy} f(x_0, y_0) = \lim_{y \to y_0} \frac{\partial_x f(x_0, y) - \partial_x f(x_0, y_0)}{y - y_0}$$

where $\partial_x f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$. We can use the two limit definitions of partial derivatives

$$\partial_x f(x_0, y) = \lim_{x \to x_0} \frac{f(x, y) - f(x_0, y)}{x - x_0} \text{ and } \partial_x f(x_0, y_0) = \lim_{x \to x_0} \frac{f(x, y_0) - f(x_0, y_0)}{x - x_0}$$

and substitute them to get the two partials

$$\begin{aligned}
\partial_{xy} f(x_0, y_0) &= \lim_{y \to y_0} \frac{\lim_{x \to x_0} \frac{f(x,y) - f(x_0,y)}{x - x_0} - \lim_{y \to y_0} \frac{f(x,y_0) - f(x_0,y_0)}{x - x_0}}{y - y_0} \\
&= \lim_{y \to y_0} \lim_{x \to x_0} \left(\frac{f(x, y) - f(x_0, y) - f(x, y_0) + f(x_0, y_0)}{(x - x_0)(y - y_0)} \right) \\
\partial_{yx} f(x_0, y_0) &= \lim_{x \to x_0} \frac{\lim_{y \to y_0} \frac{f(x,y) - f(x,y_0)}{y - y_0} - \lim_{y \to y_0} \frac{f(x_0,y) - f(x_0,y_0)}{y - y_0}}{x - x_0} \\
&= \lim_{x \to x_0} \lim_{y \to y_0} \left(\frac{f(x, y) - f(x, y_0) - f(x_0, y) + f(x_0, y_0)}{(y - y_0)(x - x_0)} \right)
\end{aligned}$$

Now invoking our assumption that f is C^2, the two limits, which approach (x_0, y_0) along different paths, both exist and are equal to

$$\lim_{(x,y) \to (x_0, y_0)} \frac{f(x, y) - f(x_0, y) - f(x, y_0) + f(x_0, y_0)}{(x - x_0)(y - y_0)}$$

and therefore $\partial_{xy} f = \partial_{yx} f$.

Corollary 3.1 ()

Given $f \in C^k$, its kth iterated partials are equal. That is, given any permutation σ,

$$\partial_{x_{i_1} x_{i_2} \dots x_{i_k}} f = \partial_{x_{\sigma(i_1)} x_{\sigma(i_2)} \dots x_{\sigma(i_k)}} f \text{ for } i_1, \dots, i_k = 1, \dots, n$$

Corollary 3.2 ()

Higher order derivatives live in the exterior algebra.

3.1 Second Order Derivatives

Definition 3.3 (Hessian Matrix)

The $n \times n$ matrix of second iterated partials of $f \in C^2$ at $\mathbf{a}$ is called the **Hessian matrix**.

$$Hf_{\mathbf{a}} \coloneqq \begin{pmatrix} \partial_{x_1x_1}(\mathbf{a}) & \dots & \partial_{x_1x_n}(\mathbf{a}) \\ \vdots & \ddots & \vdots \\ \partial_{x_nx_1}(\mathbf{a}) & \dots & \partial_{x_nx_n}(\mathbf{a}) \end{pmatrix} \text{ and } Hf \coloneqq \begin{pmatrix} \partial_{x_1x_1} & \dots & \partial_{x_1x_n} \\ \vdots & \ddots & \vdots \\ \partial_{x_nx_1} & \dots & \partial_{x_nx_n} \end{pmatrix}$$

By equality of mixed partials, it is symmetric.

Theorem 3.3 ()

The Hessian matrix of a function $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ is the Jacobian matrix of the gradient of f. That is, interpreting $\nabla f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$, we have

$$Hf_{\mathbf{a}} = D\nabla f_{\mathbf{a}} \text{ for all } \mathbf{a} \in D$$

which we can also write as $Hf = D\nabla f$. This theorem is very useful, especially for optimization and sampling methods, since we can now interpret the Hessian as the rate of change of the gradient of f.

3.2 Taylor Series

To talk about convergence, the big-O notation is very useful.

Definition 3.4 (Classes of Infinitesimal Functions)

A function $\alpha : \mathbb{R}^n \longrightarrow \mathbb{R}$ is **infinitesimal** if $\alpha \to 0$ as $\mathbf{x} \to \mathbf{x}_0$. There are multiple "levels" of infinitesimal functions, i.e. how fast they converge to 0. We can classify them by comparing their limits to polynomials.

1. α is of class $O(1)$ if
$$\lim_{\mathbf{x} \to \mathbf{x}_0} \frac{\alpha(\mathbf{x})}{1} = 0$$
This means that $\alpha(\mathbf{x})$ tends to 0 infinitely faster than 1 (which just means that it tends to 0).
2. α is of class $O(h)$ if
$$\lim_{\mathbf{x} \to \mathbf{x}_0} \frac{\alpha(\mathbf{x})}{||\mathbf{x} - \mathbf{x}_0||} = 0$$
This means that $\alpha(\mathbf{x})$ tends to 0 infinitely faster than the linear $||\mathbf{h}||$, where $\mathbf{h} = \mathbf{x} - \mathbf{x}_0$.
3. α is of class $O(h^2)$ if
$$\lim_{\mathbf{x} \to \mathbf{x}_0} \frac{\alpha(\mathbf{x})}{||\mathbf{x} - \mathbf{x}_0||^2} = 0$$
This means that $\alpha(\mathbf{x})$ tends to 0 infinitely faster than the quadratic $||\mathbf{h}||^2$, where $\mathbf{h} = \mathbf{x} - \mathbf{x}_0$.
4. α is of class $O(h^k)$ if
$$\lim_{\mathbf{x} \to \mathbf{x}_0} \frac{\alpha(\mathbf{x})}{||\mathbf{x} - \mathbf{x}_0||^k} = 0$$
This means that $\alpha(\mathbf{x})$ tends to 0 infinitely faster than the kth-order $||\mathbf{h}||^k$, where $\mathbf{h} = \mathbf{x} - \mathbf{x}_0$.

Clearly, $O(h^k) \supset O(h^{k+1})$.

Now given a $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$, we present some polynomial approximations of f at $\mathbf{x}_0 \in D$:

1. If $f \in C^0$, the zeroth (constant) approximation is just
$$P_0(\mathbf{x}) = f(\mathbf{x}_0)$$

This is not interesting at all, since it is just constant. Furthermore, the error term $\epsilon_0(\mathbf{x}) = f(\mathbf{x}) - P_0(\mathbf{x})$ is an infinitesimal function as $\mathbf{x} \to \mathbf{x}_0$ and is of class $O(1)$, since

$$\lim_{\mathbf{x} \to \mathbf{x_0}} \frac{f(\mathbf{x}) - P_0(\mathbf{x})}{1} = 0$$

2. If $f \in C^1$, the first (linear) approximation requires us to use our total derivative:

$$P_1(\mathbf{x}) = f(\mathbf{x}_0) + Df_{\mathbf{x}_0}(\mathbf{x} - \mathbf{x}_0)$$

and we know that the error $\epsilon_1(\mathbf{x}) = f(\mathbf{x}) - P_1(\mathbf{x})$ is infinitesimal as $\mathbf{x} \to \mathbf{x}_0$ and is of class $O(h)$, since

$$\lim_{\mathbf{x} \to \mathbf{x}_0} \frac{f(\mathbf{x}) - P_1(\mathbf{x})}{||\mathbf{x} - \mathbf{x}_0||} = 0$$

3. If $f \in C^2$, the second (quadratic) approximation requires us to use a quadratic term (i.e. a bilinear form of $\mathbf{h} = \mathbf{x} - \mathbf{x}_0$) centered at $\mathbf{x}_0$. Call it $H_{\mathbf{x}_0} : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$, and our estimation is

$$P_2(\mathbf{x}) = f(\mathbf{x}_0) + Df_{\mathbf{x}_0}(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2} H(\mathbf{x} - \mathbf{x}_0, \mathbf{x} - \mathbf{x}_0)$$

which we would like the error term $\epsilon_2(\mathbf{x}) = f(\mathbf{x}) - P_2(\mathbf{x})$ to be $O(h^2)$, or in limit terms, P_2 must satisfy

$$\lim_{\mathbf{x} \to \mathbf{x}_0} \frac{f(\mathbf{x}) - P_2(\mathbf{x})}{||\mathbf{x} - \mathbf{x}_0||^2} = 0$$

We show that this form H is precisely the Hessian matrix.

Theorem 3.4 (Hessian)

The second order approximation of a C^2-differentiable function f about a point $\mathbf{x}_0$ is

$$f(\mathbf{x}) = f(\mathbf{x}_0) + Df_{\mathbf{x}_0}(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T Hf_{\mathbf{x}_0}(\mathbf{x} - \mathbf{x}_0) + O(h^2)$$

where $Df_{\mathbf{x}_0}$ is the total derivative at $\mathbf{x}_0$ and $Hf_{\mathbf{x}_0}$ is the Hessian matrix at $\mathbf{x}_0$.

4 Matrix Calculus

Now we will take a look at functions that have either an input or output as matrices. Essentially, matrices are also vectors, so there is nothing new here to learn, but having a concrete set of notation is useful. First, note that when we talk about a total derivative $D\mathbf{f_a}$, we can interpret this as a linear map that takes in some small perturbation $\mathbf{h}$ and gives us the result $D\mathbf{f_a}(\mathbf{h})$. In our column-vector setting, this just corresponded to left matrix multiplication:

$$D\mathbf{f_a}(\mathbf{h}) = D\mathbf{f_a}\mathbf{h}$$

This is not the case in the matrix setting. Let us compare the following:

1. The derivative of $f : \mathbb{R} \to \mathbb{R}$ at a is a linear function $Df_a : \mathbb{R} \longrightarrow \mathbb{R}$ satisfying $f(a+h) \approx f(a) + Df_a(h) + O(h^2)$. But linearity reduces Df_a to simply a scalar, and so our condition reduces to

$$f(a+h) \approx f(a) + Df_a h + O(h^2)$$

2. The derivative of a path function $\mathbf{f} : \mathbb{R} \to \mathbb{R}^m$ is a linear function $D\mathbf{f}_a : \mathbb{R} \longrightarrow \mathbb{R}^m$ satisfying $\mathbf{f}(a+h) = \mathbf{f}(a) + D\mathbf{f}_a(h) + O(h^2)$. Linearity implies that $D\mathbf{f}_a$ is a rank-1 linear map, which reduces to it being a column vector, and our condition reduces to

$$\underbrace{\mathbf{f}(a+h)}_{m \times 1} \approx \underbrace{\mathbf{f}(a)}_{m \times 1} + \underbrace{D\mathbf{f}_a}_{m \times 1} \underbrace{h}_{1 \times 1}$$

3. The derivative of a matrix function $\mathbf{F} : \mathbb{R} \to \mathbb{R}^{m \times n}$ is a linear map $D\mathbf{F}_a : \mathbb{R} \longrightarrow \mathbb{R}^{m \times n}$ satisfying $\mathbf{F}(a+h) \approx \mathbf{F}(a) + D\mathbf{F}_a(h)$. This time, linearity does not reduce it to simple left-hand matrix multiplication. We could just say that this is a left-hand scalar multiplication, but this doesn't generalize well, so we are stuck with just saying that $D\mathbf{F}_a$ is a linear map.

$$\underbrace{\mathbf{F}(a+h)}_{m \times n} \approx \underbrace{\mathbf{F}(a)}_{m \times n} + \underbrace{D\mathbf{F}_a(h)}_{m \times n}$$

Now let us take a look at when we have matrix inputs.

1. The derivative of $f : \mathbb{R}^{m \times n} \longrightarrow \mathbb{R}$ is a linear function $Df_{\mathbf{A}} : \mathbb{R}^{m \times n} \longrightarrow \mathbb{R}$ satisfying $f(\mathbf{A}+\mathbf{H}) \approx f(\mathbf{A}) + Df_{\mathbf{A}}(\mathbf{H})$. We could let $Df_{\mathbf{A}}$ be some linear map like $\mathbf{M} \mapsto \mathbf{v}^T\mathbf{M}\mathbf{u}$, where $\mathbf{v}, \mathbf{u}$ is fixed. But in generality, we just have the condition

$$\underbrace{f(\mathbf{A}+\mathbf{H})}_{1 \times 1} \approx \underbrace{f(\mathbf{A})}_{1 \times 1} + \underbrace{Df_{\mathbf{A}}(\mathbf{H})}_{1 \times 1}$$

2. The derivative of $\mathbf{f} : \mathbf{R}^{m \times n} \longrightarrow \mathbb{R}^d$ is some linear map $D\mathbf{f}_{\mathbf{A}} : \mathbf{R}^{m \times n} \longrightarrow \mathbb{R}^d$ satisfying $\mathbf{f}(\mathbf{A}+\mathbf{H}) \approx \mathbf{f}(\mathbf{A}) + D\mathbf{f}_{\mathbf{A}}(\mathbf{H})$. Again, we could construct some form that would give us a linear map in terms of some matrix multiplication, but in generality, we have the condition

$$\underbrace{\mathbf{f}(\mathbf{A}+\mathbf{H})}_{d \times 1} \approx \underbrace{\mathbf{f}(\mathbf{A})}_{d \times 1} + \underbrace{D\mathbf{f}_{\mathbf{A}}(\mathbf{H})}_{d \times 1}$$

4.1 Simple Differentiation Rules

Now we present some theorems on basic differentiation. Proving these just requires us to expand the function and compute the derivatives component-wise.

Theorem 4.1 (Derivative of Affine Map)

Given $f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ defined $f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ (where $\mathbf{A}, \mathbf{b}$ is not dependent on $\mathbf{x}$), its derivative is

$$D\mathbf{f} = \mathbf{A}$$

Theorem 4.2 ()

Given the scalar α defined by

$$\alpha = \mathbf{y}^T\mathbf{A}\mathbf{x}$$

where $\mathbf{y} \in \mathbb{R}^{m \times 1}, \mathbf{A} \in \mathbb{R}^{m \times n}$, and $\mathbf{x} \in \mathbb{R}^{n \times 1}$, then

$$\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{y}^T\mathbf{A} : \mathbb{R}^n \longrightarrow \mathbb{R}$$

and

$$\frac{\partial \alpha}{\partial \mathbf{y}} = \mathbf{x}^T\mathbf{A}^T : \mathbb{R}^m \longrightarrow \mathbb{R}$$

Rewritten in the total derivative notation, we can interpret α as a function of both $\mathbf{x}$ and $\mathbf{y}$ and write

$$D\alpha_{(\mathbf{x},\mathbf{y})} = \begin{pmatrix} \mathbf{y}^T\mathbf{A} \\ \mathbf{x}^T\mathbf{A}^T \end{pmatrix} : \mathbb{R}^{n+m} \longrightarrow \mathbb{R}$$

Theorem 4.3 ()

Given the scalar α defined by the quadratic form

$$\alpha = \mathbf{x}^T \mathbf{A} \mathbf{x}$$

where $\mathbf{x} \in \mathbb{R}^{n \times 1}$ and $\mathbf{A} \in \mathbb{R}^{n \times n}$, then

$$\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{x}^T \big(\mathbf{A} + \mathbf{A}^T \big)$$

or in the total derivative notation,

$$D\alpha_{\mathbf{a}} = \mathbf{a}^T \big(\mathbf{A} + \mathbf{A}^T \big) : \mathbb{R}^n \longrightarrow \mathbb{R}$$

5 Vector Fields

5.1 Gradients

Definition 5.1 (Gradient)

The gradient of a C^1 scalar-valued function f is the vector field $\nabla f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ defined

$$\nabla f(\mathbf{a}) \coloneqq \begin{pmatrix} \partial_{\mathbf{x}_1}(\mathbf{a}) \\ \vdots \\ \partial_{\mathbf{x}_n}(\mathbf{a}) \end{pmatrix}$$

The gradient at a point is a **tangent vector**.

Note that the gradient is a vector field (a bundle of vectors), while the total derivative is a covector field (a bundle of covectors). Since $\mathbb{R}^n$ is an inner product space, we can invoke Riesz Representation theorem and see that they are related in the way that

$$Df_{\mathbf{a}}\mathbf{v} = \nabla f(\mathbf{a}) \cdot \mathbf{v}$$

where $\cdot$ represents the dot product. At this point, it's a bit hard to see the difference between these two, but in more abstract spaces the total derivative generalizes much better than the gradient, which exists for inner product spaces. From this, we can write a coordinate independent definition of the gradient.

Definition 5.2 (Gradient)

The gradient of a scalar valued function $f \in C^1$ is the unique vector field whose dot product with any vector $\mathbf{v}$ at each point is the directional derivative of f along $\mathbf{v}$. That is,

$$\nabla f_{\mathbf{a}} \cdot \mathbf{v} = Df_{\mathbf{a}}\mathbf{v} \text{ for all } \mathbf{a} \in D$$

Theorem 5.1 (Gradient as Direction of Fastest Increase)

Let f be a real-valued function such that $\nabla f(x) \neq 0$. Then, at the point x, $\nabla f(x)$ points in the direction along which f is increasing the fastest. Equivalently, $-\nabla f(x)$ points in the direction along which f is decreasing the fastest.

Proof. Note that this is a coordinate-independent proof. Given a directional vector $\mathbf{v}$, we can normalize it since we are only interested in direction. Evaluating it with the total derivative at x gives us $D_{\mathbf{a}}f\mathbf{v}$. But by definition,

$$\nabla f_{\mathbf{a}} \cdot \mathbf{v} = Df_{\mathbf{a}}\mathbf{v}$$

which means that

$$\begin{aligned} \sup_{||\mathbf{v}||=1} \{Df_{\mathbf{a}}\mathbf{v}\} &= \sup_{||\mathbf{v}||=1} \{\nabla f_{\mathbf{a}} \cdot \mathbf{v}\} \\ &= \sup_{||\mathbf{v}||=1} \{||\nabla f_{\mathbf{a}}||\,||\mathbf{v}||\cos(\theta)\} \\ &= \sup\{||\nabla f_{\mathbf{a}}||\cos(\theta)\} \\ &= ||\nabla f_{\mathbf{a}}|| \text{ when } \theta = 0 \end{aligned}$$

Therefore, v must point in the direction of $\nabla f_{\mathbf{a}}$.

Therefore, we can interpret the gradient evaluated at a point as the tangent vector that points in the direction

of fastest increase. We can also interpret the gradient ∇f itself as the vector field that determines some sort of "flow" in the domain $\mathbb{R}^n$. Therefore, if we drop a point in this field, the point will flow through $\mathbb{R}^n$ through a current determined by ∇f and will eventually end up at a local maximum.

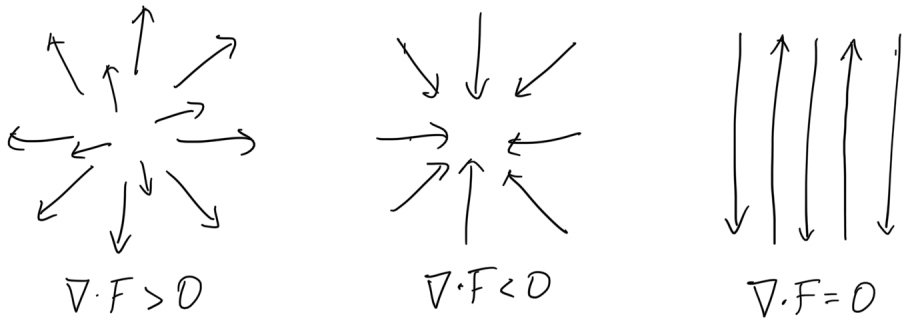
Definition 5.3 (Del Operator)

For convenience, we use the del operator to denote the gradient. The **del operator** $\nabla : f \mapsto \nabla f$ takes in a differentiable function and outputs the gradient of it.

5.2 Divergence

Colloquially, the divergence is an operator div that operates on a vector field and produces a scalar field which provides the quantity of the vector field's source at each point. Technically, the divergence represents the volume density of the outward flux of a vector field from an infinitesimal volume around a given point.

There is a very nice geometric interpretation for divergence. Imagine that the vector field F represents fluid flow in $\mathbb{R}^n$. Divergence is then the "measure" of the net amount of fluid flowing in and out of an infinitesimally small region, labeled at each point. If the net fluid flow is positive (i.e. more fluid is flowing in than out) at point x_0, then $\operatorname{div} F(x_0) > 0$. If the net fluid flow is negative (i.e. more fluid is flowing out than in) at point x_0, then $\operatorname{div} F(x_0) < 0$. This measure assigns a number to every point in the space (creating a scalar field). Therefore, each point either acts as a "source" of fluid emanating from it or as a "sink" that sucks in more fluid than it puts out.



Definition 5.4 (Divergence)

The **divergence** of a vector field $\mathbf{F} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is a scalar field defined

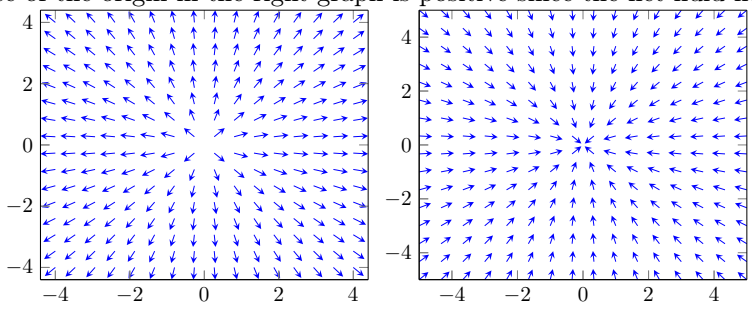
$$\operatorname{div} \mathbf{F} \coloneqq \nabla \cdot \mathbf{F} = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix} \cdot \begin{pmatrix} F_1 \\ \vdots \\ F_n \end{pmatrix} = \sum_i \frac{\partial \mathbf{F}_i}{\partial x_i}$$

When $n = 1$, $\mathbf{F}$ reduces to a regular function and $\operatorname{div} \mathbf{F}$ reduces to the ordinary derivative. Some further properties:

1. By linearity of partials, div is also a linear operator. That is, given two vector fields $\mathbf{F}, \mathbf{G}$ and two scalars α, β,
$$\operatorname{div}(\alpha \mathbf{F} + \beta \mathbf{G}) = \alpha \operatorname{div} \mathbf{F} + \beta \operatorname{div} \mathbf{G}$$
2. Divergence satisfies the product rule: Given a vector field $\mathbf{F} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ and a scalar function $\varphi : \mathbb{R}^n \longrightarrow \mathbb{R}$.
$$\nabla \cdot (\varphi \mathbf{F}) = \nabla \varphi \cdot \mathbf{F} + \varphi (\nabla \cdot \mathbf{F})$$

Example 5.1 ()

The divergence of the origin in the left graph is clearly negative since the net flow is out of the point, while the divergence of the origin in the right graph is positive since the net fluid flow is in.



Lemma 5.1 (Divergence in Cylindrical Coordinates)

For vector field $F : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ expressed in cylindrical coordinates as

$$F = \begin{pmatrix} F_r \\ F_\theta \\ F_z \end{pmatrix}$$

the divergence is

$$\operatorname{div} F = \nabla \cdot F = \frac{1}{r}\frac{\partial}{\partial r}(rF_r) + \frac{1}{r}\frac{\partial F_\theta}{\partial \theta} + \frac{\partial F_z}{\partial z}$$

Note that the condition of locality is important, since in general a global cylindrical coordinate system would be inconsistent.

Lemma 5.2 (Divergence in Spherical Coordinates)

For vector field $F : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ expressed in spherical coordinates (r, θ, ϕ), the divergence is

$$\operatorname{div} F = \nabla \cdot F = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2 F_r) + \frac{1}{r\,\sin\theta}\frac{\partial}{\partial \theta}(\sin\theta F_\theta) + \frac{1}{r\,\sin\theta}\frac{\partial F_\phi}{\partial \phi}$$

5.3 Curl

Colloquially, the curl is a vector operator that describes the infinitesimal circulation of a vector field in 3-dimensional Euclidean space, where the curl at each point is represented by a vector whose length and direction denote the magnitude and axis as the maximum circulation. That is, if one drops a twig or a ball with its center of mass at a certain point, the curl measures how much it will spin. In physics, the rotation of a rigid body in 3-dimensions can be described by a vector ω along the axis of rotation. ω is called the *angular velocity vector*, with $||\omega||$ denoting the angular speed of the body. The curl of this vector field measured at the center of mass of the body is measured as 2ω. That is, the curl outputs *twice* the angular velocity vector of any rigid body. Note that unlike the gradient and divergence operators, curl does not generalize as simply to other dimensions.

Definition 5.5 (Curl)

The *curl* of a 3-dimensional C^k vector field $F : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ is an operator

$$\operatorname{curl} : C^k(\mathbb{R}^3; \mathbb{R}^3) \longrightarrow C^{k-1}(\mathbb{R}^3; \mathbb{R}^3)$$

defined

$$\operatorname{curl} F \equiv \nabla \times F \equiv \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} \equiv \begin{pmatrix} \frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z} \\ \frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x} \\ \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \end{pmatrix}$$

Definition 5.6 (Irrotational Vector Fields)

A vector field F is *irrotational* if

$$\operatorname{curl} F = \mathbf{0}$$

Visually, this indicates that there are no "whirlpools" everywhere, meaning that any rigid body placed anywhere, while it may travel along a path, will not rotate around its own axis.

It has been shown that fluid draining from a tub is usually irrotational except for right at the center, which is surprising since the fluid itself is "rotating" around the drain.

Theorem 5.2 ()

For any C^2 vector field F,

$$\operatorname{div} \operatorname{curl} F = \nabla \cdot (\nabla \times F) = 0$$

That is, the divergence of any curl is 0.

Proof. Proved by equality of mixed partials.

Definition 5.7 ()

The *Laplace operator*, or *Laplacian*, of a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is the divergence of the gradient.

$$\nabla^2 f \equiv \nabla \cdot (\nabla f) \equiv \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}$$

5.4 Conservative, Solenoidal Vector Fields

Definition 5.8 (Conservative Vector Fields)

A vector field $F : U \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is a *conservative vector field* if and only if there exists a scalar field $f : U \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ such that

$$F = \nabla f$$

on U.

Conservative vector fields appear naturally in mechanics: they are vector fields representing forces of physical systems in which energy is conserved.

Theorem 5.3 ()

Given a C^2-function $f : \mathbb{R}^3 \longrightarrow \mathbb{R}$,

$$\nabla \times (\nabla f) = 0$$

That is, the curl of any gradient vector field is the zero vector.

Proof. $\nabla \times \nabla f$ can be expanded to

$$\left(\frac{\partial^2 f}{\partial y \partial z} - \frac{\partial^2 f}{\partial z \partial y}, \frac{\partial^2 f}{\partial z \partial x} - \frac{\partial^2 f}{\partial x \partial z}, \frac{\partial^2 f}{\partial x \partial y} - \frac{\partial^2 f}{\partial y \partial x} \right) = (0, 0, 0)$$

by equality of mixed partials.

Definition 5.9 (Solenoidal Vector Fields)

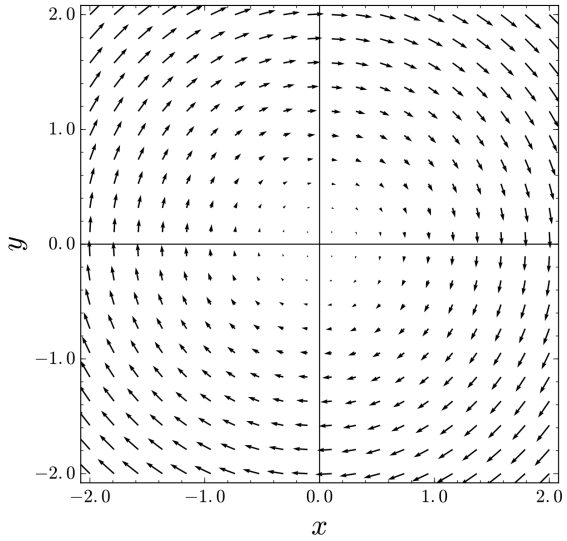
A *solenoidal, or incompressible, vector field* is a vector field $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ such that

$$\operatorname{div} F = \nabla \cdot F = 0$$

at all point in the field. That is, the field has no sources or sinks.

Example 5.2 ()

The vector field $F : (x, y) \mapsto (y, -x)$ is solenoidal.



6 Riemann and Darboux Integration

7 Surfaces

We can represent a K-dimensional subset $S \subset \mathbb{R}^N$ in multiple ways, where $K < N$. There are three conventional ways to do this.

1. We can parameterize it with a function $f : D \subset \mathbb{R}^k \longrightarrow \mathbb{R}^n$ to create a **parameterized set** defined as the image of an *injective* f under D. Letting $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{u} \in \mathbb{R}^k$, the parameterization is defined
$$\mathbf{u} \mapsto f(\mathbf{u}) = \big(f_1(\mathbf{u}), f_2(\mathbf{u}), \dots, f_n(\mathbf{u})\big)$$
2. A function $\mathbf{f} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ of the form $\mathbf{y} = \mathbf{f}(\mathbf{x})$ creates an **explicit representation** by defining all $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{n+m}$ satisfying
$$\mathbf{y} = \mathbf{f}(\mathbf{x})$$
3. A **level set** of the form $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ creates an **implicit representation** by defining all $\mathbf{x} \in \mathbb{R}^n$ satisfying
$$\mathbf{F}(\mathbf{x}) = \mathbf{0}$$
Now if $\mathbf{F}$ was scalar valued, then the equation $F(\mathbf{x}) = 0$ defines a hypersurface in $\mathbb{R}^n$ with codimension 1. If $\mathbf{F}$ is a k-vector valued function, then the implicit surface generally has codimension k, since we can interpret $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ as a system of k constraint equations.

Generally, the change of representations is simple only when the explicit representation $\mathbf{y} = \mathbf{f}(\mathbf{x})$ is given. The implicit form is $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{y} - \mathbf{f}(\mathbf{x}) = \mathbf{0}$, and the parameterized form is the map $\mathbf{x} \mapsto (\mathbf{x}, \mathbf{f}(\mathbf{x}))$. However, the explicit representation is very limited in usefulness, because it can only describe sets that are graphs of functions that pass the vertical line test. The implicit function theorem, stated later, states conditions under which an equation $\mathbf{F}(\mathbf{x}) = 0$ can be solved explicitly for any of the x_i's. The other two representations are much more versatile, with the implicit representation being slightly more general, but the parametric form being more useful, since we can directly compute points on the S. Some examples are:

1. a 1-dimensional path/curve in $\mathbb{R}^n$
2. a 2-dimensional surface in $\mathbb{R}^3$
3. a k-dimensional set in $\mathbb{R}^n$

If these surfaces are smooth enough, then there must exist geometric tangent vectors, geometric tangent planes, and geometric orthogonal vectors on them. We say "geometric" to distinguish them from the vectors in the tangent space $T_{\mathbf{x}_0}\mathbb{R}^n$. It is important to know how to derive them.

Theorem 7.1 (Explicit Representation)

Let us have the surface $S \subset \mathbb{R}^{n+1}$ defined by $y = f(\mathbf{x})$ and a point on the surface $(\mathbf{x}_0, f(\mathbf{x}_0))$.

1. To get the equation of the set of affine points forming the geometric tangent plane, we look at all points $(\mathbf{x}, y)$ satisfying
$$y = f(\mathbf{x}_0) + Df_{\mathbf{x}_0}(\mathbf{x} - \mathbf{x}_0)$$
and to get an arbitrary tangent vector protruding from $\mathbf{x}_0$, we look at all vectors $(\mathbf{v}, w)$ of form
$$w = Df_{\mathbf{x}_0}\mathbf{v}$$
i.e. all vectors of form $(\mathbf{v}, Df_{\mathbf{x}_0}\mathbf{v})$.
2. To get the equation of the orthogonal vector, convert this to the implicit representation $g(\mathbf{x}, y) = y - f(\mathbf{x}) = 0$, and see that the gradient is orthogonal to the tangent plane. So, the orthogonal vector at $\mathbf{x}_0$ is
$$\nabla g(\mathbf{x}_0, f(\mathbf{x}_0)) = \begin{pmatrix} -\nabla f(\mathbf{x}_0) \\ 1 \end{pmatrix}$$

Note that indeed, dotting this with an arbitrary tangent vector of the form above gives

$$\begin{pmatrix} -\nabla f(\mathbf{x}_0) \\ 1 \end{pmatrix} \cdot \begin{pmatrix} \mathbf{v} \\ Df_{\mathbf{x}_0}\mathbf{v} \end{pmatrix} = -\nabla f(\mathbf{x}_0) \cdot \mathbf{v} + Df_{\mathbf{x}_0}\mathbf{v} = 0$$

Given a level set $S = \{\mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) = c\}$, a vector $\mathbf{v}$ is a **tangent vector** of S at $\mathbf{a}$ if the directional derivative (if it exists) satisfies

$$\nabla_{\mathbf{v}} f(\mathbf{a}) = 0$$

If f is differentiable at $\mathbf{a}$, then this condition is equivalent to

$$Df_{\mathbf{a}}\mathbf{v} = 0$$

Intuitively, $Df_{\mathbf{a}}\mathbf{v}$ answers the question: "If I move infinitesimally in the direction $\mathbf{v}$, what happens to f?" We would want this direction to preserve the value of $f = c$, and so the derivative should be 0. Therefore, we look for the vectors $\mathbf{v}$ satisfying $Df_{\mathbf{a}}\mathbf{v} = 0$, i.e. the annihilator $(Df_{\mathbf{a}})^0 \subset \mathbb{R}^n$. This result is precisely the well-known theorem that states that "gradients are orthogonal to level sets." It is intuitive to claim that if we have some sort of directional vector $\mathbf{v}$, then this $\mathbf{v}$ must be "tangent" if the directional derivative towards $\mathbf{v}$ must be 0, essentially staying within the level set of value c.

Theorem 7.2 (Implicit Representation)

Let us have the surface $S \subset \mathbb{R}^n$ defined by $F(\mathbf{x}) = 0$ and a point $\mathbf{a} \in S$.

1. The gradient $\nabla F(\mathbf{x}_0)$ is simply the orthogonal vector at $\mathbf{x}_0$.
2. The set of all directional tangent vectors protruding from $\mathbf{x}_0$ is defined by the set of directional vectors $\mathbf{v}$ satisfying
$$\nabla F(\mathbf{x}_0) \cdot \mathbf{v} = 0$$
and the set of all affine points forming the geometric tangent plane are all $\mathbf{x} \in \mathbb{R}^n$ satisfying
$$\nabla F(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) = 0$$

Proof. This is trivial since we can invoke Reisz representation theorem and see that

$$Df_{\mathbf{a}}\mathbf{v} = 0 \implies \nabla_{\mathbf{a}} f \cdot \mathbf{v} = 0$$

This theorem now simplifies our derivation of tangent planes of a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$. To find the equation of a tangent plane of $y = f(\mathbf{x})$ at $\mathbf{x} = \mathbf{a}$, we can simply write the one-line equation as

$$y = f(\mathbf{a}) + Df_{\mathbf{a}}(\mathbf{x} - \mathbf{x}_0)$$

However, if we had an implicit function of the form $g(\mathbf{x}, y) = c$, then separating this into an explicit function of y is hard. Therefore, we can simply treat g itself as a function of the $n + 1$ variables $(\mathbf{x}, y)$, and treat $g(\mathbf{x}, y) = c$ as a level set.

Theorem 7.3 (Parametric Representation)

Let us have $f : D \subset \mathbb{R}^k \longrightarrow \mathbb{R}^n$, with injective f defining a surface $f(D) \subset \mathbb{R}^n$. Let us have $\mathbf{u}_0 \in D$ with $f(\mathbf{u}_0) = \mathbf{x}_0 \in f(D)$. Our idea is this: we compute k directional derivatives of f in k linearly independent direction vectors at $\mathbf{u}_0$, which will give us k (linearly independent, due to injectiveness of f, but not necessarily orthogonal) geometric tangent vectors protruding from $\mathbf{x}_0$ that span the tangent space $T_{\mathbf{x}_0}$. If $k = n - 1$, then the orthogonal vector is uniquely defined to be the vector spanning $T_{\mathbf{x}_0}^{\perp}$, and is $k < n - 1$, there is no unique orthogonal vector defined.

1. The set of all directional tangent vectors $\mathbf{v}$ protruding from $\mathbf{x}_0$ is represented by the set of all

linear combinations of the partials, aka the image of the Jacobian of f

$$\{c_1\partial_{u_1}f(\mathbf{u}_0)+\ldots+c_k\partial_{u_k}f(\mathbf{u}_0)\mid \mathbf{c}\in\mathbb{R}^n\}=\text{Im}\begin{pmatrix}| & \ldots & | \\ \partial_{u_1}f(\mathbf{u}_0) & \ldots & \partial_k f(\mathbf{u}_0) \\ | & \ldots & |\end{pmatrix}=\text{Im}Df_{\mathbf{u}_0}$$

The tangent space is the space of all $\mathbf{x}\in\mathbb{R}^n$ of the form

$$f(\mathbf{x}_0)+Df_{\mathbf{x}_0}\mathbf{u} \text{ for all } \mathbf{u}\in\mathbb{R}^k$$

2. If $k=n-1$, the orthogonal vector is the unique vector that is orthogonal to all $\partial_{u_i}f(\mathbf{u}_0)$, which can be computed using linear algebra techniques (e.g. kernel of $Df_{\mathbf{x}_0}$). If $n=3,k=2$, then this can simply be computed using the cross product.

8 Integration of Forms

9 Sequences of Functions