Author: Imad Ali
Editor: Prof Ross Flek

This document is intended to be a brief overview of the salient topics in Vector Calculus. For a more comprehensive discussion of the field see Barr (2001) or Stewart (2007). Both texts were used in developing this document.

1 Linear Algebra

Outstanding Corrections

- vector addition: parallelogram method - a weighted average of the two vectors.
- p.3 show the direction of the vector that connects the two points.
- law of cosines
- p.8 generalize the cofactor expansion of the determinant
- p.8 generalize the inverse of a matrix
- p.10 explain how matrix **A** represents a linear transformation
- p.10 explain why the determinant of the matrix must not be equal to zero

1.1 Vectors

An *n-dimensional* vector is an *n-tuple* of numbers. The values of the tuple are the *components* of the vector. A vector is denoted as $\vec{v}$ or $\mathbf{v}$. A *zero vector* or a *null vector* is where all components of the vector are zero. A *unit vector* is a vector that has a length of one. Note that dividing a vector by its magnitude (i.e. $\frac{\vec{v}}{|\vec{v}|}$) will give you a unit vector.

Vectors can be represented as arrows and give a sense of direction. For instance, in two-dimensional space the vector $\vec{p} = (3, 5)$ is illustrated by drawing a line from the origin to the point represented by the coordinates $(3, 5)$.

There are two operations that can be performed on vectors.

- *Scalar multiplication*: An n-dimensional vector $\vec{p}$ can be multiplied by some scalar a. The resulting vector $a\vec{p}$ is where every component in $\vec{p}$ is multiplied by a.

$$a\vec{p} = (ap_1, ap_2, \ldots, ap_{n-1}, ap_n)$$

  Scalar multiplication stretches or contracts vector, and can change the direction of a vector.

- *Summation*: As long as vectors have the same dimension they can be added to one another. Similar to multiplication, addition is done by summing the n^{th} component in one vector with the n^{th} component in another vector.

$$\vec{p} + \vec{q} = (p_1 + q_1, p_2 + q_2, \ldots, p_{n-1} + q_{n-1}, p_n + q_n)$$

The figures above illustrate vector scalar multiplication (Figure 1) and vector summation (Figure 2). You can see how multiplying a vector by a negative scalar will cause the vector to point in the opposite direction. Multiplying a vector by a scalar less than one will reduce the magnitude of the vector and multiplying by a scalar greater than one will increase the magnitude.

Figure 1: Vector Multiplication

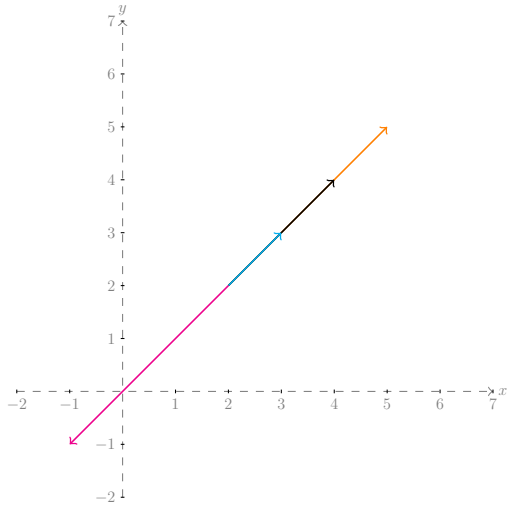
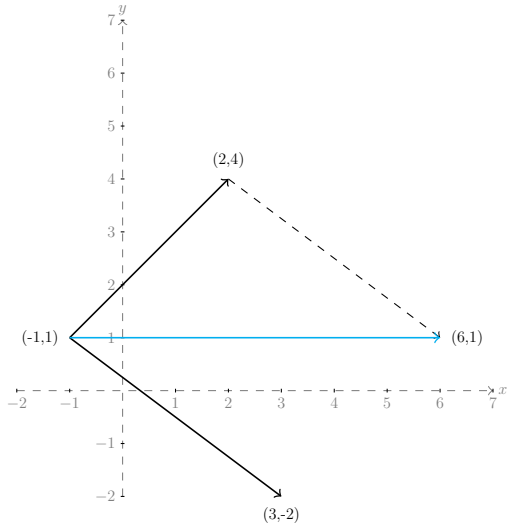


Figure 2: Vector Summation



The *dot product* (*scalar product* or *inner product*) for two n-dimensional vectors is obtained by multiplying the corresponding components and then summing over these products.

$$\vec{p} \cdot \vec{q} = p_1 \cdot q_1 + p_2 \cdot q_2 + \ldots + p_{n-1} \cdot q_{n-1} + p_n \cdot q_n$$

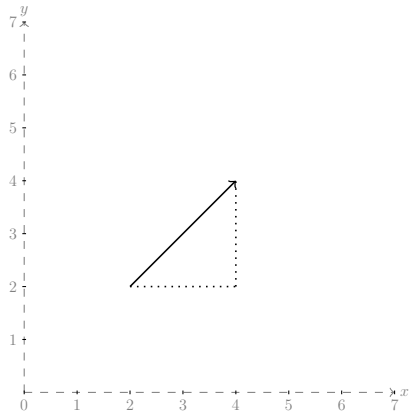
The dot product measures the extent to which two vectors point in the same direction. The dot product will be positive if they point in the same direction, negative if they point in opposite directions, and zero if they are *orthogonal* (i.e. perpendicular) to each another. This will be made clear when we determine the distance between two vectors.

Consider an arbitrary vector $\vec{p}$. The *length* of this vector, denoted $|\vec{p}|$, from the origin is calculated using the Pythagorean theorem.

$$|\vec{p}| = \sqrt{p_x^2 + p_y^2} = \sqrt{\vec{p} \cdot \vec{p}}$$

Figure 3 below illustrates the intuition behind finding the length of a vector in $\mathbb{R}^2$ space.

Figure 3: Vector Length



Finding the distance $d(\cdot)$ associated with the (x, y) coordinates of $\vec{p}$ and $\vec{q}$ is analogous to finding the length of the vector $\vec{r}$ that connects $\vec{p}$ and $\vec{q}$.

$$d(\vec{p}, \vec{q}) = |\vec{r}| = \sqrt{(q_x - p_x)^2 + (q_y - p_y)^2}$$

This can be written in terms of the dot product of $\vec{p}$ and $\vec{q}$.

$$\begin{aligned}
|\vec{r}|^2 &= (q_x - p_x)^2 + (q_y - p_y)^2 \\
&= q_x^2 - 2q_x p_x + p_x^2 + q_y^2 - 2q_y p_y + p_y^2 \\
&= p_x^2 + p_y^2 + q_x^2 + q_y^2 - 2(p_x q_x + p_y q_y) \\
&= |\vec{p}|^2 + |\vec{q}|^2 - 2(\vec{p} \cdot \vec{q}) \\
2(\vec{p} \cdot \vec{q}) &= |\vec{q}| + |\vec{p}| - |\vec{r}| \\
\vec{p} \cdot \vec{q} &= \frac{|\vec{q}|^2 + |\vec{p}|^2 - |\vec{r}|^2}{2}
\end{aligned}$$

Vectors that are orthogonal form a 90° angle. Using the vectors specified above, the Pythagorean theorem states that $|\vec{q}|^2 + |\vec{p}|^2 = |\vec{r}|^2 \implies |\vec{q}|^2 + |\vec{p}|^2 - |\vec{r}|^2 = 0$. Therefore, the dot product $\vec{p} \cdot \vec{q} = 0$ for orthogonal vectors. An extension of this is that the dot product for vectors that meet at an angle different from 90° will be different from zero. Figure 4 illustrates how we find the distance between two vectors in $\mathbb{R}^2$ space.

Two vectors whose dot product is zero are said to be *orthogonal* vectors. Graphically, orthogonal vectors are vectors that are perpendicular to one another as seen in Figure 5 below.

Figure 4: Vector Distance

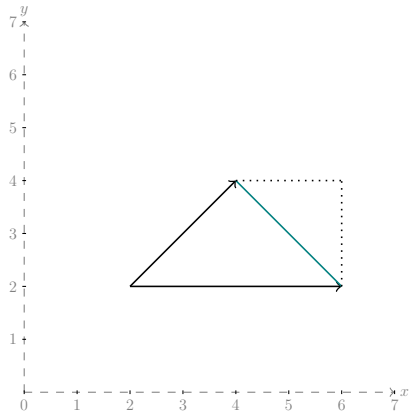
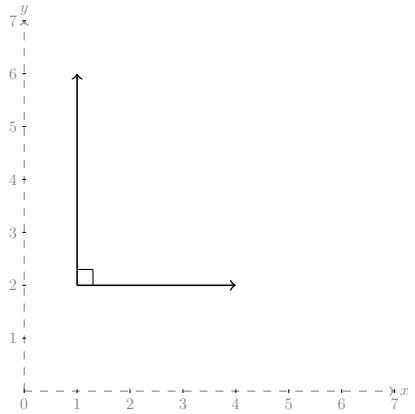


Figure 5: Orthogonal Vectors



The *Cauchy-Schwartz inequality* can be derived using the above. First, the triangle inequality states that the length of the third side of a triangle is greater than the absolute value of the difference of the other two sides and is less than the sum of the other two sides.

$$||\vec{q}| - |\vec{p}|| \leq |\vec{r}| \leq |\vec{q}| + |\vec{p}|$$

The three terms are non-negative so all parts of the compound inequality may be squared.

$$|\vec{q}|^2 - 2(|\vec{q}| \cdot |\vec{p}|) + |\vec{p}|^2 \leq |\vec{r}|^2 \leq |\vec{q}|^2 + 2(|\vec{q}| \cdot |\vec{p}|) + |\vec{p}|^2$$

Using the dot product formulation above, we can write this as follows.

$$\begin{aligned}
-|\vec{q}|^2 - 2(|\vec{q}| \cdot |\vec{p}|) - |\vec{p}|^2 \leq & -|\vec{r}|^2 \leq -|\vec{q}|^2 + 2(|\vec{q}| \cdot |\vec{p}|) - |\vec{p}|^2 \\
-|\vec{q}|^2 - 2(|\vec{q}| \cdot |\vec{p}|) - |\vec{p}|^2 \leq & 2(\vec{q} \cdot \vec{p}) - |\vec{q}|^2 - |\vec{p}|^2 \leq -|\vec{q}|^2 + 2(|\vec{q}| \cdot |\vec{p}|) - |\vec{p}|^2 \\
-2(|\vec{q}| \cdot |\vec{p}|) \leq & 2(\vec{q} \cdot \vec{p}) \leq 2(|\vec{q}| \cdot |\vec{p}|) \\
-|\vec{q}| \cdot |\vec{p}| \leq & (\vec{q} \cdot \vec{p}) \leq |\vec{q}| \cdot |\vec{p}| \\
-1 \leq & \frac{(\vec{q} \cdot \vec{p})}{|\vec{q}| \cdot |\vec{p}|} \leq 1
\end{aligned}$$

The above two formulations represent ways to measure the similarity between two vectors. He we have restricted ourself to the two-dimensional case, but the results hold in n-dimensions. $|\vec{r}| =$

$\sqrt{(\vec{q}-\vec{p})\cdot(\vec{q}-\vec{p})}$ says that the squared distance between two vectors is the dot product of the difference of those vectors. The Cauchy-Schwartz inequality is a scale invariant way to measure the similarity between two vectors.

The Cauchy-Schwartz inequality can be written in terms of the angle between $\vec{p}$ and $\vec{q}$. The law of cosines relates the length of a side of a triangle to the cosine of the angle opposite of that side. Therefore, the following holds regarding the distance between the coordinates of $\vec{p}$ and $\vec{q}$.

$$
\begin{aligned}
|\vec{r}|^2 &== |\vec{p}|^2 + |\vec{q}|^2 - 2(|\vec{p}|\cdot|\vec{q}|)\cos(\theta) \\
\cos(\theta) &= \frac{|\vec{p}|^2 + |\vec{q}|^2 - |\vec{r}|^2}{2(|\vec{p}|\cdot|\vec{q}|)} \\
\cos(\theta) &= \frac{\vec{p}\cdot\vec{q}}{|\vec{p}|\cdot|\vec{q}|} \\
\theta &= \cos^1\left(\frac{\vec{p}\cdot\vec{q}}{|\vec{p}|\cdot|\vec{q}|}\right)
\end{aligned}
$$

So the size of the angle between the two vectors is another measure of similarity.

An application of the dot product representing the similarity between two vectors can been seen in recommendation systems. Consider k consumers of various products. Each consumer has an n-dimensional vector $\vec{k}$ that represents the quantity of each of the n products that have been consumed. In order to make recommendations to a consumer we determine which consumers have similar vectors based on the methods described above. We then make recommendations to that consumer based on what the other consumers have purchased.

A function $f(\cdot)$ is a *linear transformation* if it satisfies the following properties.

- For any vector $\vec{v}$ and any scalar a, $f(a\cdot\vec{v}) = a\cdot f(\vec{v})$
- For any vectors $\vec{v}$ and $\vec{w}$, $f(\vec{v}+\vec{w}) = f(\vec{v}) + f(\vec{w})$

So if a function $f(\cdot)$ is a linear transformation from n-dimensional vectors to numbers, then there exists a unique vector $\vec{u}$ such that, for all $\vec{v}$, $f(\vec{v}) = \vec{u}\cdot\vec{v}$.

This is essentially saying that there is some vector $\vec{u}$ that represents the linear transformation that $f(\cdot)$ applies to $\vec{v}$.

If one vector can be written as a scalar multiple of another, then both vectors are said to be *linearly dependent*. If both vectors are not linearly dependent, then they are *linearly independent*.

Vectors have a *magnitude* and a *direction*. When drawing a vector the arrow indicates the direction. The magnitude is calculated using the Pythagorean theorem. Two vectors are identical if they have the same magnitude and direction. A zero vector has no direction and has a magnitude of zero. Both $(0,0)$ and $(0,0,0)$ are zero vectors in the two- and three-dimensional space, respectively A

unit vector will have some direction and will have a magnitude of one. Both $(0, 1)$ and $(0, 1, 0)$ are unit vectors in the two- and three- dimensional space, respectively. Applying the Pythagorean theorem will confirm the definitions of the the aforementioned vectors.

Consider two vectors $\vec{v}$ and $\vec{w}$. The *vector projection* of $\vec{w}$ onto $\vec{v}$ is obtained by dropping a perpendicular from the head of $\vec{w}$ to the line on which $\vec{v}$ lies, and drawing a vector from the tail of $\vec{v}$ to the point at which the perpendicular intersects the line on which $\vec{v}$ lies.

The *component* of $\vec{w}$ in the direction of $\vec{v}$ is the scalar that can transform $\vec{v}$ into the projection of $\vec{w}$ on $\vec{v}$.

$$comp_{\vec{v}}(\vec{w}) = \frac{\vec{v} \cdot \vec{w}}{|\vec{v}|}$$
$$\vec{proj}_{\vec{v}}(\vec{w}) = comp_{\vec{v}}(\vec{w}) \frac{\vec{v}}{|\vec{v}|}$$

(Note that the projection is a vector and the component is the magnitude of the vector).

The *cross product* of two nonzero, nonparallel vectors is the vector that is orthogonal to both of them. Given the nonzero, nonparallel vectors $\vec{a}$ and $\vec{b}$, vector $\vec{x} = (x, y, z)$ must be nonzero and satisfy

$$\vec{x} \cdot \vec{a} = a_1 x + a_2 y + a_3 z = 0$$
$$\vec{x} \cdot \vec{b} = b_1 x + b_2 y + b_3 z = 0$$

Solving this system yields

$$\vec{x} = (a_2 b_3 - a_3 b_2, a_3 b_1 - a_1 b_3, a_1 b_2 - a_2 b_1)$$

We can also find the cross product of two vectors using cofactor expansion. For instance, the cross product of the three dimensional vectors $\vec{a}$ and $\vec{b}$ can be written as the following determinant,

$$\vec{a} \times \vec{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} = \begin{vmatrix} a_2 & a_3 \\ b_2 & b_3 \end{vmatrix} \mathbf{i} - \begin{vmatrix} a_1 & a_3 \\ b_1 & b_3 \end{vmatrix} \mathbf{j} + \begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix} \mathbf{k}$$

Each determinant is a component in the resulting cross product vector.

1.2 Matrices

An $m \times n$ *matrix* is an array of elements (e.g. numbers) that contains m rows and n columns. It can also be defined as collection of m n-dimensional vectors or n m-dimensional vectors. Matrices are typically denoted as $\mathbf{M}$. The element corresponding to the ith row and the jth column is denoted

$\mathbf{M}[i, j]$. An $m \times m$ matrix is a square matrix.

A 3×3 matrix is given below.

$$\mathbf{M}_{3\times3} = \begin{bmatrix} m_{1,1} & m_{1,2} & m_{1,3} \\ m_{2,1} & m_{2,2} & m_{2,3} \\ m_{3,1} & m_{3,2} & m_{3,3} \end{bmatrix}$$

The transpose of a matrix, denoted $\mathbf{M}^\top$, consists of transforming the rows of $\mathbf{M}$ into the columns of $\mathbf{M}^\top$. If $\mathbf{M}$ is an $m \times n$ matrix then $\mathbf{M}^\top$ will be a $n \times m$ matrix. A square matrix is *symmetric* if it equals its transpose.

$$\mathbf{M}_{2\times3} = \begin{bmatrix} m_{1,1} & m_{1,2} & m_{1,3} \\ m_{2,1} & m_{2,2} & m_{2,3} \end{bmatrix}$$

$$\mathbf{M}^\top_{3\times2} = \begin{bmatrix} m_{1,1} & m_{2,1} \\ m_{1,2} & m_{2,2} \\ m_{1,3} & m_{2,3} \end{bmatrix}$$

Scalar multiplication on matrices and the summation of two matrices is performed component by component, as with vectors.

Let $\mathbf{M}$ be an $m \times n$ matrix and $\vec{v}$ be an n-dimensional vector. The product $\mathbf{M} \cdot \vec{v}$ is an m-dimensional vector. The product vector will consist of the dot product of each row of $\mathbf{M}$ and $\vec{v}$. This can also be thought of as a weighted sum of the columns of $\mathbf{M}$ where the weights for each column are the components of $\vec{v}$. The following properties hold for multiplying a matrix by a vector where $\mathbf{M}$ and $\mathbf{N}$ are $m \times n$ matrices, $\vec{u}$ and $\vec{v}$ are n-dimensional vectors, and a is a scalar.

$$\begin{aligned} (\mathbf{M} + \mathbf{N})\vec{v} &= \mathbf{M}\vec{v} + \mathbf{N}\vec{v} \\ \mathbf{M}(\vec{u} + \vec{v}) &= \mathbf{M}\vec{v} + \mathbf{M}\vec{u} \\ a(\mathbf{M})\vec{v} &= a(\mathbf{M}\vec{v}) = \mathbf{M}(a\vec{v}) \end{aligned}$$

Let $f(\cdot)$ be a function that represents a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then there exists a unique $m \times n$ matrix $\mathbf{F}$ such that, for all $\vec{v}$, $f(\vec{v}) = \mathbf{F}\vec{v}$. We say that matrix $\mathbf{F}$ corresponds to the transformation of $f(\cdot)$, and vice versa.

Matrices can represent systems of equations. An example is given below where $\vec{x}$ denotes a n-dimensional vector of unknown variables, $\mathbf{B}$ is a matrix of constants, and $\vec{a}$ is a m-dimensional vector of constants.

$$\mathbf{B} = \begin{bmatrix} b_{1,1} & \dots & b_{1,n} \\ \vdots & & \vdots \\ b_{m,1} & \dots & b_{m,n} \end{bmatrix} \vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \vec{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix}$$

$$\begin{bmatrix} b_{1,1} & \dots & b_{1,n} \\ \vdots & & \vdots \\ b_{m,1} & \dots & b_{m,n} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} b_{1,1}x_1 + b_{1,2}x_2 + \ldots + b_{1,n-1}x_{n-1} + b_{1,n}x_n = a_1 \\ b_{2,1}x_1 + b_{2,2}x_2 + \ldots + b_{2,n-1}x_{n-1} + b_{2,n}x_n = a_2 \\ \vdots \\ b_{m,1}x_1 + b_{m,2}x_2 + \ldots + b_{m,n-1}x_{n-1} + b_{m,n}x_n = a_m \end{bmatrix}$$

As mentioned, matrices can be thought of as a collection of vectors. Consider an n-dimensional vector $\vec{v}$ as matrix $\mathbf{V}$ and a $m \times n$ matrix $\mathbf{M}$. The product $\mathbf{M} \times \mathbf{V}$ is a m-dimensional vector in which each component is the dot product between each row of $\mathbf{M}$ and $\mathbf{V}$. Now consider a $n \times m$ matrix $\mathbf{U}$. The product $\mathbf{M} \times \mathbf{U}$ is a $m \times m$ matrix where each component represents the dot product between each row in $\mathbf{M}$ and each column in $\mathbf{M}$.

Given the above we can specify the following. Let $\mathbf{M}$ be a $m \times n$ matrix and $\mathbf{O}$ is a $n \times p$ matrix. The product $\mathbf{M} \times \mathbf{O}$ is a $m \times p$ matrix defined by

$$\mathbf{M} \cdot \mathbf{O} = \sum_{k=1}^{n} \mathbf{M}[i,k] \cdot \mathbf{O}[k,j]$$

A few special matrices include the *identity matrix* and *triangular matrices*. An identity matrix is a symmetric matrix (i.e. a square matrix that possesses symmetry along the main diagonal) where the components along the main diagonal are all equal to one. It has the ability to preserve the identity of the matrix that is being multiplied by it. Consider the $m \times n$ matrix $\mathbf{M}$. Then there exists two identity matrices, $\mathbf{I}_n$ and $\mathbf{I}_m$, that will preserve the $\mathbf{M}$ when multiplied on the right and left sides, respectively.

$$\begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix} \text{ and } \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix}$$

A *upper triangular matrix* is a square matrix where all the components below the main diagonal are zero. A *lower triangular matrix* is a square matrix where all the components above the main diagonal are zero. The transpose of an upper triangular matrix is a lower triangular matrix (and vice versa).

The *determinant* of a 2×2 matrix is the product of the components on the main diagonal minus the product of the components on the off diagonal.

$$det(\mathbf{A}) = \begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix} = a_1 \cdot b_2 - a_2 \cdot b_1$$

The determinant of a 3×3 matrix is computed as follows

$$\begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix} = a_1 \begin{vmatrix} b_2 & b_3 \\ c_2 & c_3 \end{vmatrix} - a_2 \begin{vmatrix} b_1 & b_3 \\ c_1 & c_3 \end{vmatrix} + a_3 \begin{vmatrix} b_1 & b_2 \\ c_1 & c_2 \end{vmatrix}$$

Higher order matrices can be solved by breaking them up into smaller parts, as done with the 3×3 matrix.

A $m \times m$ square matrix $\mathbf{A}$ is *invertible* if there exists an $m \times m$ square matrix $\mathbf{B}$ such that $\mathbf{BA} = \mathbf{I}_m$ and $\mathbf{AB} = \mathbf{I}_m$. In this context $\mathbf{B} = \mathbf{A}^{-1}$ is the inverse of $\mathbf{A}$, and vice versa. Note that this definition

implies that the inverse of an invertible matrix is unique.

The inverse for any square matrix $\mathbf{A}$ is the product of one over the determinant of $\mathbf{A}$ and the matrix that contains the negative of the components on the off diagonal and swaps the components on the main diagonal of the original matrix $\mathbf{A}$.

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$\mathbf{A}^{-1} = \frac{1}{det(\mathbf{A})} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Therefore, in order for a matrix to be invertible, we also require that the determinant be not equal to zero. The matrix that is being multiplied by the factor $1/det(\mathbf{A})$ is known as the *adjoint* of $\mathbf{A}$, denoted $adj(\mathbf{A})$.

1.3 Vector Space

Vector spaces are certain sets of n-dimensional vectors. Specifically, they are vectors that can undergo vector addition and scalar multiplication. A *subspace* of $\mathbb{R}^n$ is a particular type of set of n-dimensional vectors. Subspaces of $\mathbb{R}^n$ are a type of vector space.

Recall that a *linear combination* of a set of vectors $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n$ from $\mathbb{R}^n$ is the sum of these vectors scaled by some constants from $\mathbb{R}$: $c_1\vec{v}_1 + c_2\vec{v}_2 + \ldots + c_n\vec{v}_n$. These are *linear* combinations since we are simply adding the vectors and then scaling them up by some constant (we are not multiplying vectors by one another or anything like that). The set of all of the vectors that can be represented in $\mathbb{R}^n$ by all the possible linear combinations on $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n$ is denoted $span(\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n)$. Specifically, this represents the set of sum of all the possible scaled combinations of the vectors. Mathematically,

$$span(\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n) = \{c_1\vec{v}_1 + c_2\vec{v}_2 + \ldots + c_n\vec{v}_n \mid c_i \in \mathbb{R}^n \ \forall \ 1 \leq i \leq n\}$$

To determine the span of a set of vectors you have to solve the system of equations,

$$c_1\vec{v}_1 + c_2\vec{v}_2 + \ldots + c_n\vec{v}_n \ = x_n$$

for the constants.

A few important examples:

- $span(0) = 0$
- Let $\vec{v}_1$ and $\vec{v}_2$ be orthogonal 2-dimensional vectors. Then $span(\vec{v}_1, \vec{v}_2) = \mathbb{R}^2$.

The *plane through point* $x = (x_0, y_0, z_0)$ *perpendicular to nonzero vector* $\vec{n} = (n_1, n_2, n_3)$ consists of all points $x = (x, y, z)$, such that $(x - x_0) \cdot \vec{n} = 0$. The vector $\vec{n}$ is called a *normal* to the plane. The equation for the plane will take the following form.

$$n_1(x - x_0) + n_2(y - y_0) + n_3(z - z_0) = 0$$

Given a vector $\vec{m}$ and point $x_0 = (x_0, y_0, z_0)$ we define the *line through* x_0 *parallel to* $\vec{m}$ to be the set

$$\{x \in \mathbb{R}^3 | x = t\vec{m} + x_0, t \in \mathbb{R}\}$$

$\vec{m}$ is known as the *direction vector* of the line. The vector equation $x = t\vec{m} + x_0$ is known as the *[parametrization* of the line where the *parameter* is t. Since x is a set of three-dimensional coordinates we can obtain the following *parametric equations* for the line:

$$x = n_1 t + x_0$$
$$y = n_2 t + y_0$$
$$z = n_3 t + z_0$$

Consider a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$. If we apply T to the collection of points P in $\mathbb{R}^n$ then we will obtain a new collection of points $T(P) = \{T(x) | x \in P\}$. $T(P)$ is known as the *image of* P *under* T. A linear transformation of T on $\mathbb{R}^n$ is called a *dilation* if $T(x) = rx$ for some $r \in \mathbb{R}^n$. Dilation is a special case of *coordinate rescaling*. Coordinate rescaling involves a linear transformation of the form $T(x) = \mathbf{D}x$ where $\mathbf{D}$ is a diagonal matrix.

Consider $\mathbb{R}^n$ to be the space containing the n-dimensional vectors $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n$. More formally, The set of vectors $V = \{\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n\} \in \mathbb{R}^n$. The set of vectors V is a *subspace* of $\mathbb{R}^n$ if the following three conditions hold:

1. V contains the zero vector $\mathbf{0}$.
2. For any vector $\vec{x}$ in V and any scalar a, $a\vec{x}$ must also be in V.
3. For any vectors $\vec{x}$ and $\vec{y}$ in V, the sum $\vec{x} + \vec{y}$ must also be in V.

Point 2 refers to the subspace being closed under scalar multiplication and point 3 refers to the subspace being closed under addition.

1.4 Eigendecomposition

An *eigendecomposition* is the decomposition of a matrix into its constituent parts: *eigenvalues* and *eigenvectors*.

Consider the linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$. Given the linear transformation T there exists at most n vectors $\vec{v} \in \mathbb{R}^n$ such that $T(\vec{v}) = \lambda_n \vec{v}$, where λ is some scalar that transforms $\vec{v}$. Here, $\vec{v}$ is known as the eigenvector for the transformation T and λ is the eigenvalue associated with the eigenvector $\vec{v}$.

More formally, if you have a matrix that represents some linear transformation $\mathbf{A}$ then we have,

$$\lambda\vec{v} = \mathbf{A}\vec{v}$$

where $\vec{v}$ is the eigenvector of $\mathbf{A}$ and λ is the eigenvalue associated with the eigenvector $\vec{v}$.

We can solve for the eigenvectors and eigenvalues by using the identity $\vec{v} = \mathbf{I}_n\vec{v}$.

$$\begin{aligned}\lambda\vec{v} - \mathbf{A}\vec{v} &= \mathbf{0}\\ \lambda\mathbf{I}_n\vec{v} - \mathbf{A}\vec{v} &= \mathbf{0}\\ (\lambda\mathbf{I}_n - \mathbf{A})\vec{v} &= \mathbf{0}\end{aligned}$$

$\mathbf{A}\vec{v} = \lambda\vec{v}$ for nonzero $\vec{v}$ if and only if the determinant of $\lambda\mathbf{I}_n - \mathbf{A}$ is zero.

2 Equations of Lines and Planes

Here we generalize the equation of a line to three-dimensional space and introduce the equation of a plane in three-dimensional space.

An intuitive way to understand the equation of a line in $\mathbb{R}^3$ is by example. Suppose we want to graph the value $x = 6$ in $\mathbb{R}$, $\mathbb{R}^2$, or $\mathbb{R}^3$.

- In $\mathbb{R}$ it will just be a point where $x = 6$.
- In $\mathbb{R}^2$ it will be a vertical line that crosses the x-axis at $x = 6$. This line will consider all the potential y coordinates at $x = 6$. We can write this as $(6, y)$ for all $y \in \mathbb{R}$. Formally, we can say that the equation of this line is $a6 + by = c \implies y = c - 6$.
- In $\mathbb{R}^3$ it will be a plane. This follows naturally from the description in $\mathbb{R}^2$. We are considering all the potential z coordinates *and* all the potential y coordinates such that $x = 6$. More formally, we are considering $(6, y, z)$ for all $y, z \in \mathbb{R}$. Note that a plane is a special case of a surface where at least one of the coordinates is fixed (in this case x).

What about a familiar equation for a line $y = 3x + 9$ in $\mathbb{R}^2$ or $\mathbb{R}^3$.

- In $\mathbb{R}^2$ it will be a line. Passing any x value into the function will give you a corresponding y value and will allow you to trace out the function in two-dimensional space.
- In $\mathbb{R}^3$ this will be a plane. In $\mathbb{R}^2$ the equation traces out a line when a sequence of x values is passed into it. This exists for all $z \in \mathbb{R}$. So we are using the line to trace out a plane for every value of z.

include plots here

A *vector-valued function* or a *vector function* is a function that takes one or more variables (or vectors) as an input and outputs a set of vectors. In other words, the range of a vector-valued function is a set of vectors. For instance, consider the following vector-valued function,

$$\vec{v}(v_1, v_2) = (v_1, v_2, 3)$$

It takes two inputs (v_1 and v_2) and returns a vector (i.e. a point) in three-dimensional space.

2.1 The Vector and Parametric Form Equation of a Line

The mathematical form of a line in $\mathbb{R}^2$ is familiar. We also need a way to represent lines in $\mathbb{R}^n$ where $n > 2$. What follows is the mathematical representation of lines in $\mathbb{R}^3$. The extension to $\mathbb{R}^n$ should be intuitive.

The *vector form equation of a line* is written as $p = t\vec{m} + p_0$, where $p = (x, y, z)$ and $p_0 = (x_0, y_0, z_0)$, $\vec{m} = (a, b, c)$, and $t \in \mathbb{R}$. Expanding this gives,

$$(x, y, z) = t(a, b, c) + (x_0, y_0, z_0)$$

The *parametric form equation of the line* $p = t\vec{m} + p_0$ is,

$$x = ta + x_0$$
$$y = tb + y_0$$
$$z = tc + z_0$$

So, assuming that we have an initial point p_0, some vector $\vec{m}$, and scalar value t, we can find any point on the line in three-dimensional space. Typically one might be given two points in $\mathbb{R}^3$ and will be asked to find the equation of the line that connects them. In this case, one of the two points would give us the initial point p_0 and the difference between two of these points gives us $\vec{m}$, a line that is parallel to the line that connects the two given points.

The *symmetric equations of a line* are found by solving the parametric equations for t and equating the equations to one another.

$$\frac{x - x_0}{a} = \frac{y - y_0}{b} = \frac{z - z_0}{c}$$

This holds as long as $a, b, c \neq 0$.

As an example, say we want to find the equation of the line that passes through points $(2, -1, 3)$ and $(1, 4, -3)$ in $\mathbb{R}^3$ space. Recall that the difference between two vectors will give you a vector that is parallel to the line that passes through the points that describe those two vectors. Therefore, the vector that is parallel to the line that passes through the points above is,

$$\vec{m} = (2 - 1, -1 - 4, 3 + 3) = (1, -5, 6)$$

Using $\vec{m}$ and some point on the line, say $p_0 = (1, 4, -3)$, our parametric equations are,

$$x = t + 1$$
$$y = -5t + 4$$
$$z = 6t - 3$$

which gives us the parametric form equation of a line.

The vector form equation of this line is,

$$(x, y, z) = t(1, -5, 6) + (1, 4, -3)$$

2.2 The Scalar Equation of a Plane

The equation of a plane requires a bit more work. Let $p_0 = (x_0, y_0, z_0)$ be a point on the plane. There exists some vector $\vec{n} = (a, b, c)$ that is orthogonal to the plane. Let $p = (x, y, z)$ be another point on the plane. So we have now defined two points on the plane and a vector orthogonal to the plane. Now let $\vec{r} = p - p_0$. (Recall that $\vec{r}$ will give us a vector that is parallel to the line that connects points p_0 and p.) We defined $\vec{n}$ to be orthogonal to the plane and $\vec{r}$ as a vector on the

plane. Therefore, by definition, $\vec{n}$ is orthogonal to $\vec{r}$, which means the dot product between the two vectors is zero. More formally,

$$\begin{aligned}\vec{r} \cdot \vec{n} &= 0 \\ (p - p_0) \cdot \vec{n} &= 0 \\ (x - x_0, y - y_0, z - z_0) \cdot (a, b, c) &= 0 \\ (x - x_0)a + (y - y_0)b + (z - z_0)c &= 0\end{aligned}$$

This is known as the *scalar equation of the plane*. It is often written as,

$$ax + by + cz = d$$

where $d = ax_0 + by_0 + cz_0$.

As an example, say we want to determine the equation for the plane that contains points $p = (1, -2, 0)$, $q = (3, 1, 4)$, and $r = (0, -1, 2)$. Using these points we can find two vectors that sit on this plane,

$$\begin{aligned}\vec{pq} &= (2, 3, 4) \\ \vec{pr} &= (-1, 1, 2)\end{aligned}$$

Now we can find the vector orthogonal to the plane by finding the cross product of the vectors $\vec{pq}$ and $\vec{pr}$. Note that the cross product of two vectors yields a vector that is orthogonal to these vectors.

The cross product can be found by finding the determinant of the following matrix,

$$\vec{pq} \cdot \vec{pr} = \begin{vmatrix} \vec{i} & \vec{j} & \vec{k} \\ 2 & 3 & 4 \\ -1 & 1 & 2 \end{vmatrix} = \vec{i}2 - \vec{j}8 + \vec{k}5$$

here, $\vec{i}, \vec{j}$, and $\vec{k}$ refer to the standard basis vectors.

Using point p and the orthogonal vector $\vec{n} = (2, -8, 5)$, the equation for the plane is defined as,

$$\begin{aligned}2(x - 1) - 8(y + 2) + 5(z - 0) &= 0 \\ 2x - 8y + 5z &= 18\end{aligned}$$

To summarize, the steps to finding the equation of a plane in $\mathbb{R}^3$ are as follows,

1. Find two vectors on the plane.
2. Using the cross product of the two vectors, find a vector that is orthogonal to the plane.
3. Substitute the orthogonal vector and any point on the plane into the scalar equation of the plane: $a(x - x_0) + b(y - y_0) + c(z - z_0) = 0$ where $\vec{n} = (a, b, c)$ is the orthogonal vector and (x_0, y_0, z_0) is a point on the plane.

3 Differentiation

Consider the function $f : \mathbb{R} \to \mathbb{R}$. In single variable calculus the derivative of $f(x)$ at the point $x = a$, denoted $df(x)/dx$, is given by evaluating the following limit,

$$\frac{df(x)}{dx} = \lim_{x \to a} \frac{f(a) - f(x)}{a - x}$$

if we define $a = x + h$ then we can write the derivative as,

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}$$

In vector calculus, instead of having a single equation represent the derivative of a vector valued function $\mathbf{f}$ at point $\vec{a}$ we will have a matrix. This matrix is called the *Jacobian* matrix and represents the total derivative of $\mathbf{f}$ at $\vec{a}$.

3.1 Functions, Graphs, Level Sets, and Vector Fields

A *function maps* a subset U of $\mathbb{R}^n$ to $\mathbb{R}^m$ if there is a rule $\mathbf{f}$ which associates each vector in U with exactly one vector in $\mathbb{R}^m$. In this context, U is defined as the *domain* of the function $\mathbf{f}$. More formally, we can say that $\mathbf{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$. In order to make notation consistent, $\mathbf{f}$ is used when the function represents a map from $\mathbb{R}^n$ to $\mathbb{R}^m$ when $m > 1$, and f is used when the function represents a map from $\mathbb{R}^n$ to $\mathbb{R}$.

The *graph of a function* $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ is the set of points,

$$(x_1, \ldots, x_n, \ldots, x_{n+m})$$

in $\mathbb{R}^{n+m}$ such that,

$$(x_{n+1}, \ldots, x_{n+m}) = \mathbf{f}(x_1, \ldots, x_n)$$

As an example, consider a function $f : U \subset \mathbb{R} \to \mathbb{R}$ where f is given by,

$$f(x_1) = 3x_1 + 6$$

The graph of the function would be the set of points $(x_1, x_2) \in \mathbb{R}^2$ such that $(x_2) = f(x_1)$.

As another example, consider a function $f : U \subset \mathbb{R}^2 \to \mathbb{R}$ where f is given by,

$$f(x_1, x_2) = x_1^2 + x_2^2$$

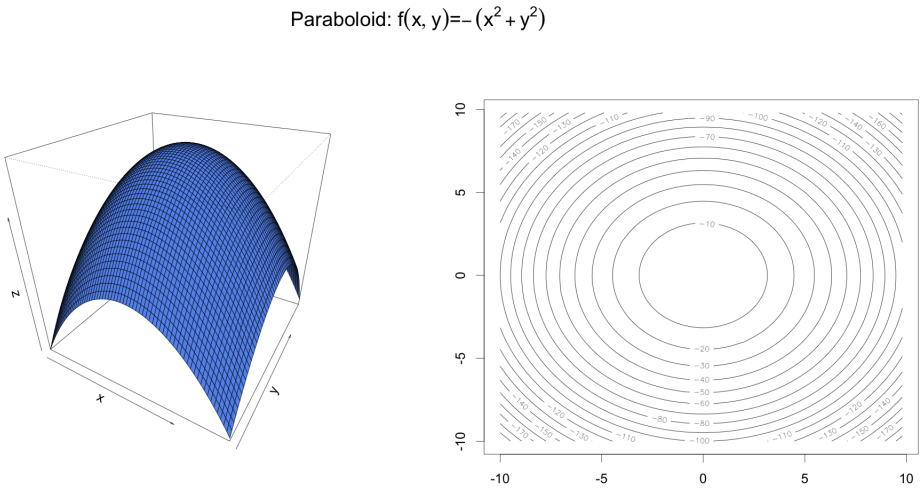
The graph of the function would be the set of points $(x_1, x_2, x_3) \in \mathbb{R}^3$ such that $(x_3) = f(x_1, x_2)$.

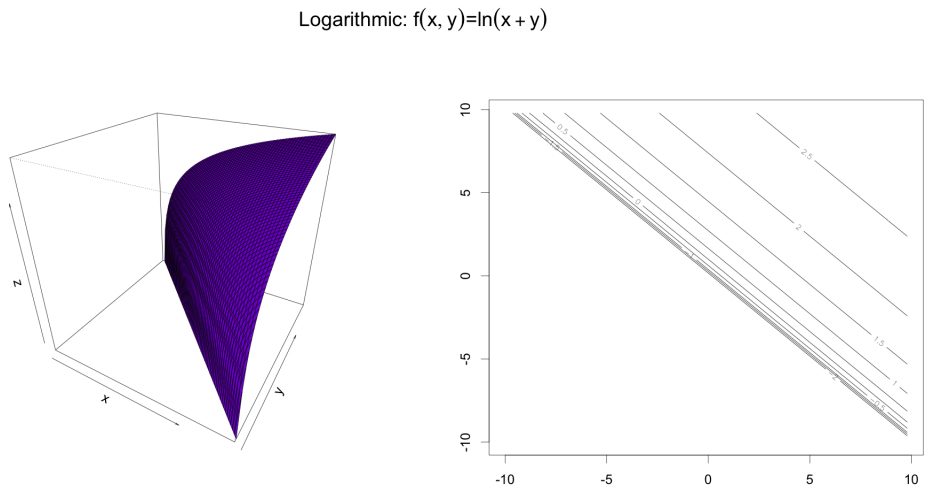
So, if you have n-dimensional inputs and m-dimensional outputs, then the graph of the function that maps the n-dimensional inputs to the m-dimensional outputs will require $m+n$ dimensions. In the examples above, the graph of $f : U \subset \mathbb{R} \to \mathbb{R}$ requires a two-dimensional space and the graph of $f : U \subset \mathbb{R}^2 \to \mathbb{R}$ requires a three-dimensional space. Similarly, the graph of $f : U \subset \mathbb{R}^2 \to \mathbb{R}^2$ requires a four-dimensional space.

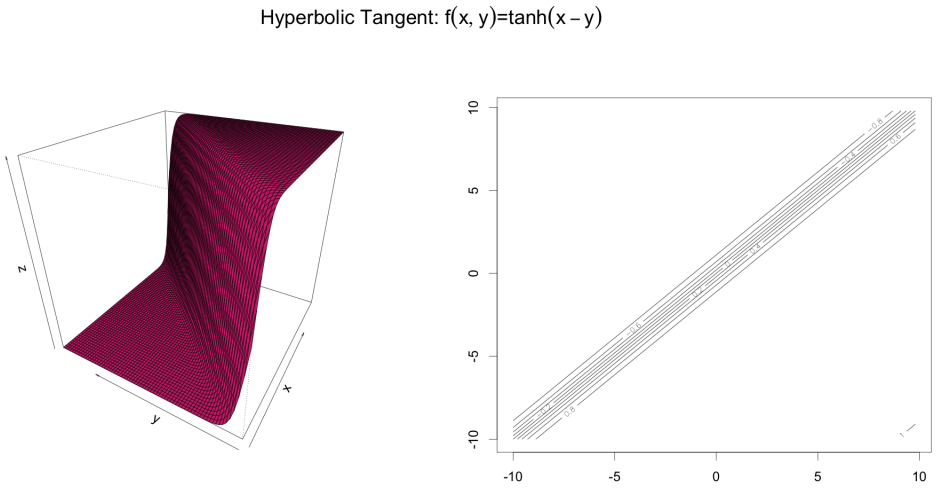
Consider a function $f : U \subset \mathbb{R}^n \to \mathbb{R}$ and a constant $c \in \mathbb{R}$. Then the *level set of* f is the set of points $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ that satisfy the equation $f(x_1, x_2, \ldots, x_n) = c$. Note that for every value of $c \in \mathbb{R}$ we will have a level set, so if we find the level sets for all values of c then we will obtain a set of level sets.

For instance, consider the function $f : U \subset \mathbb{R}^2 \to \mathbb{R}$ and $z \in \mathbb{R}$. Then the set of points (x_1, x_2) that satisfies the equation $f(x_1, x_2) = z$ is a level set of f. In this case, if we find the level sets for all values of $z \in \mathbb{R}$ then we can plot all these level sets. This plot is a contour plot.

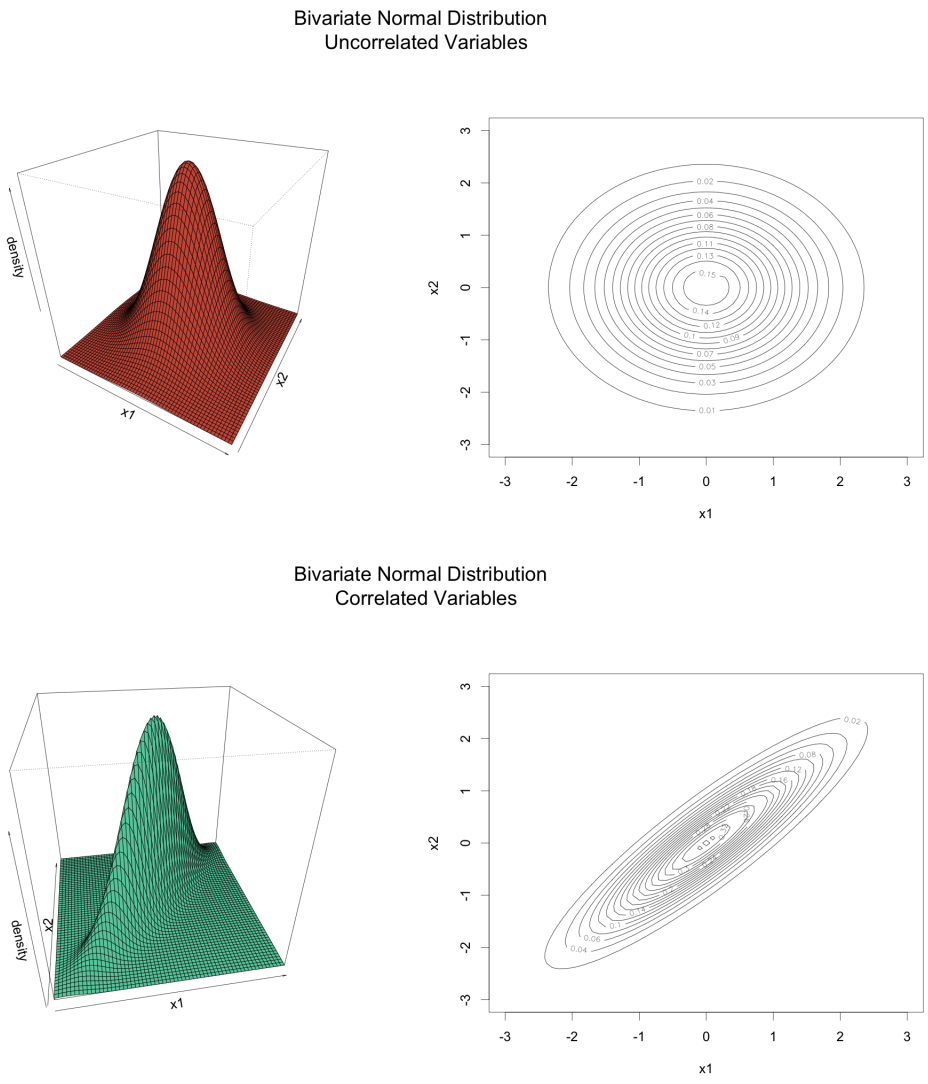
Below are three examples of a three dimensional plot and the corresponding two dimensional level sets of the function $f : U \subset \mathbb{R}^2 \to \mathbb{R}$. The first example represents the function $f(x, y) = -(x^2+y^2)$. The shape that this function takes on is known as a paraboloid. The second example represents The third example represents $f(x, y) = \tanh(x, y)$, which is also known as the hyperbolic tangent function.







We can use level sets to examine the correlation between two random variables using, for example, the bivariate normal distribution. The bivariate normal distribution can be plotted in three-dimensions with the x-axis and y-axis representing the two random variables and the z-axis representing the probability density associated with the x-y pair of random variables.



A *vector field* on $\mathbb{R}^n$ is a function $\mathbf{f} : U \subset \mathbb{R}^n \to \mathbb{R}^n$.

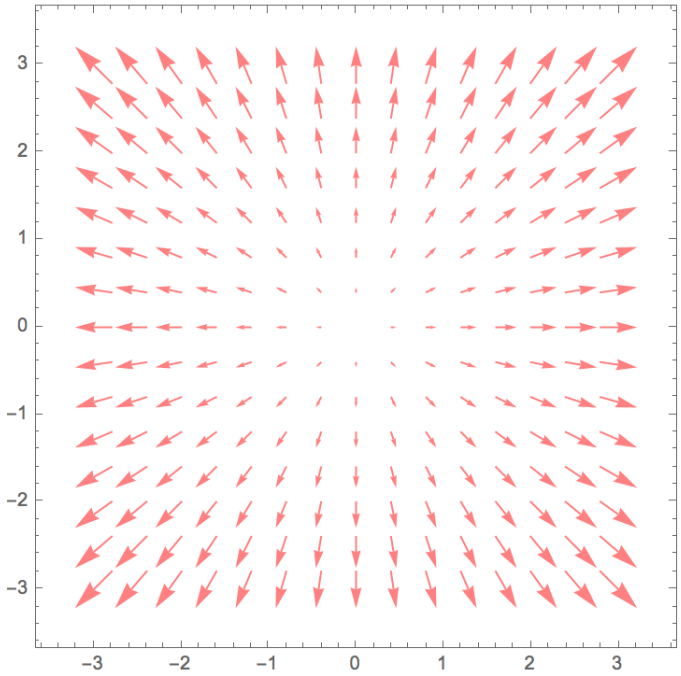
For instance, the function $\mathbf{f} : U \subset \mathbb{R}^3 \to \mathbb{R}^3$ will map three-dimensional vectors into a three-dimensional space. In this case the three-dimensional input vector, (x, y, z), will be the tail of each vector in the vector field and the output $\mathbf{f}(x, y, z)$ will be the head of the vector that starts at (x, y, z).

As a two dimensional example $\mathbf{f} : U \subset \mathbb{R}^2 \to \mathbb{R}^2$, Figure 6 illustrates the graph of $f(x, y) = (x, y)$ where $-3 \leq x, y \leq 3$.

3.2 Partial Derivatives

If we have a function of several variables, then the *partial derivative* of this function represents the rate of change of the function with respect to each of the these variables.

Figure 6: A Vector Field of $f(x, y) = (x, y)$



More formally, the partial derivative of $f : U \subset \mathbb{R}^n \to \mathbb{R}$ with respect to $x_i \in \mathbb{R}$ is the function,

$$\frac{\partial f}{\partial x_i}(x_1, x_2, \ldots, x_n) = \lim_{h \to 0} \frac{f(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)}{h}$$

When applying this definition to a function of several variables we take the partial derivative of the function with respect to x_i by treating all the variables x_{-i} in the function as constants.

Just as with higher-order derivatives in functions of single variables, we also have higher-order derivatives in functions with multiple variables. The *second-order partial derivatives* of a function f with respect to x_j then x_i is defined as,

$$\frac{\partial^2 f}{\partial x_j \partial x_i} = \frac{\partial}{\partial x_j} \frac{\partial f}{\partial x_i}$$

When $i \neq j$ we call the partial derivatives *mixed partials*.

Consider two mixed partial derivatives of the same function where in the first case we differentiate with respect to x_i then x_j and in the second case we differentiate with respect to x_j then x_i. Then, *Clairaut's Theorem* shows that both of these mixed partial derivatives are equal to one another.

More formally, Clairaut's Theorem states that if the mixed partials of $f : D \subset \mathbb{R}^2 \to \mathbb{R}$ are defined in an open disk B about a point $\mathbf{a}$ in D, and if $\partial^2 f / \partial x_j \partial x_i$ and $\partial^2 f / \partial x_i \partial x_j$ are continuous at point $\mathbf{a}$ then,

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{a}) = \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{a})$$

3.3 Differentiation and the Total Derivative

A function $\mathbf{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ is *differentiable* at a point $\mathbf{a} \in U$ if there exists a linear transformation $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^m$ such that,

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

where $\mathbf{T}$ is the *derivative* (or *total derivative*) of $\mathbf{f}$ *at* $\mathbf{a}$.

The total derivative is denoted as $\mathbf{Df}(\mathbf{a})$. So calculating the derivative of the function $\mathbf{f}$ at point $\mathbf{a}$ requires finding the matrix $\mathbf{Df}(\mathbf{a})$.

Formally, we can say that if $f : U \subset \mathbb{R}^n \to \mathbb{R}$ has partial derivatives that are defined on an open ball about point $\mathbf{a} \in U$ and if the partial derivatives of f with respect to $x_1, x_2, \ldots, x_n$ are continuous at $\mathbf{a}$, then f is differentiable at $\mathbf{a}$ and the derivative of f at $\mathbf{a}$ are given by,

$$(Df(\mathbf{a}))(\mathbf{x}) = \left[\frac{\partial f}{\partial x_1}\mathbf{a}, \frac{\partial f}{\partial x_2}\mathbf{a}, \ldots, \frac{\partial f}{\partial x_n}\mathbf{a} \right] \mathbf{x}$$

In order to find the derivative of a vector valued function at point $\mathbf{a}$ we need the *Jacobian matrix of* f *at* $\mathbf{a}$ denoted $Jf(\mathbf{a})$. Given the function $f : U \subset \mathbb{R}^n \to \mathbb{R}$, the Jacobian matrix is,

$$Jf = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right]$$

The extension to a vector valued function $\mathbf{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ should be intuitive. The Jacobian will have n columns and m rows, where each row represents the output of the function $\mathbf{f}$: $f_1, f_2, \ldots, f_m$.

As an example, consider the function $f : U \subset \mathbb{R}^2 \to \mathbb{R}$ such that $f(x_1, x_2) = -x_1^2 - x_2^2$. The Jacobian of $f(x_1, x_2)$ at any point $\mathbf{a} = (x_1, x_2)$ is given by,

$$Jf(x_1, x_2) = \begin{bmatrix} -2x_1 & -2x_2 \end{bmatrix}$$

So at any point $\mathbf{a} = (x_1, x_2)$, the derivative of $f(x_1, x_2)$ is given by,

$$Jf(\mathbf{a})(x_1, y_1) = \begin{bmatrix} -2a_1 & -2a_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

In single variable calculus the linear approximation of a function $f : U \subset \mathbb{R} \to \mathbb{R}$ at point a where $x \in \mathbb{R}$ is given by,

$$L(x) = f(x) + f'(x)(x - a)$$

The extension to a function $f : U \subset \mathbb{R}^n \to \mathbb{R}$ with multiple input variables is given by,

$$L(\mathbf{x}) = f(\mathbf{a}) + Jf(\mathbf{a})(\mathbf{x} - \mathbf{a})$$

The above is the linear approximation of $f(\mathbf{x})$ at point $\mathbf{a}$.

The total derivative of a real-valued function $f : \mathbb{R}^n \to \mathbb{R}$ can be written using *differential notation*,

$$df = \frac{\partial f}{\partial x_1}dx_1 + \frac{\partial f}{\partial x_2}dx_2 + \ldots + \frac{\partial f}{\partial x_n}dx_n$$

where df is called the *total differential* of f.

Let $\mathbf{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ be given by,

$$\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x}))$$

The derivative of $\mathbf{f}$ at point $\mathbf{a}$ is the linear transformation $\mathbf{Df}(\mathbf{a}) : \mathbb{R}^n \to \mathbb{R}^m$ given by,

$$(\mathbf{Df}(\mathbf{a}))(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}\mathbf{a} & \frac{\partial f_1}{\partial x_2}\mathbf{a} & \ldots & \frac{\partial f_1}{\partial x_n}\mathbf{a} \\ \frac{\partial f_2}{\partial x_1}\mathbf{a} & \frac{\partial f_2}{\partial x_2}\mathbf{a} & \ldots & \frac{\partial f_2}{\partial x_n}\mathbf{a} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}\mathbf{a} & \frac{\partial f_m}{\partial x_2}\mathbf{a} & \ldots & \frac{\partial f_m}{\partial x_n}\mathbf{a} \end{bmatrix}$$

example here.

From the definition of differentiability we know that,

$$(\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{a})) - (\mathbf{Df}(\mathbf{a})(\mathbf{x} - \mathbf{a})$$

is approximately 0 for values of $\mathbf{x}$ near $\mathbf{a}$. We can rearrange this to show that,

$$\mathbf{f}(\mathbf{x}) \approx \mathbf{f}(\mathbf{a}) + \mathbf{Df}(\mathbf{a})(\mathbf{x} - \mathbf{a})$$

where $\mathbf{f}(\mathbf{a}) + \mathbf{Df}(\mathbf{a})(\mathbf{x} - \mathbf{a})$ is called the *differential approximation of* $\mathbf{f}$ *near* $\mathbf{a}$.

3.4 The Chain Rule

If $\mathbf{f} : U \subset \mathbb{R}^m \to \mathbb{R}^p$ and $\mathbf{g} : \mathbf{f}(U) \to \mathbb{R}^m$, then the *composition* of $\mathbf{g}$ with $\mathbf{f}$, denoted $\mathbf{g} \circ \mathbf{f}$, is defined by,

$$(\mathbf{g} \circ \mathbf{f})(\mathbf{x}) = \mathbf{g}(\mathbf{f}(\mathbf{x}))$$

This can be thought of as starting with $\mathbf{x} \in \mathbb{R}^m$ and applying the function $\mathbf{f}(\cdot)$, which gives us $\mathbf{f}(\mathbf{x})$. Then we apply the function $\mathbf{g}(\cdot)$ to $\mathbf{f}(\mathbf{x})$, which gives us $\mathbf{g}(\mathbf{f}(\mathbf{x}))$.

We can extend the chain rule to vector valued functions. Let $\mathbf{f} : U \subset \mathbb{R}^n \to \mathbb{R}^p$ and $\mathbf{g} : \mathbf{f}(U) \to \mathbb{R}^m$. If $\mathbf{f}$ is differentiable at $\mathbf{a} \in U$ with derivative $\mathbf{Df}(\mathbf{a})$ and $\mathbf{g}$ is differentiable at $\mathbf{f}(\mathbf{a})$ with derivative $\mathbf{Dg}(\mathbf{f}(\mathbf{a}))$, then the composite function $\mathbf{g} \circ \mathbf{f}$ is differentiable at $\mathbf{a}$ and the derivative of $\mathbf{g} \circ \mathbf{f}$ at $\mathbf{a}$ is the linear transformation $\mathbf{D}(\mathbf{g} \circ \mathbf{f})(\mathbf{a})$ given by,

$$(\mathbf{D}(\mathbf{g} \circ \mathbf{f})(\mathbf{a}))(\mathbf{x}) = \mathbf{Dg}(\mathbf{f}(\mathbf{a}))((\mathbf{Df}(\mathbf{a}))(\mathbf{x}))$$

or in terms of the Jacobian matrix the derivative of $\mathbf{g} \circ \mathbf{f}$ is given by,

$$J(\mathbf{g} \circ \mathbf{f})(\mathbf{a}) = J\mathbf{g}(\mathbf{f}(\mathbf{a}))J\mathbf{f}(\mathbf{a})$$

As an example consider two functions $f : U \subset \mathbb{R}^2 \to \mathbb{R}^2$ and $g : f(U) \to \mathbb{R}^2$. Let, $g(x, y) = (x^2y^3, 3x - y^2)$ and $f(x, y) = (2y, 3x + y^2)$. We are interested in finding the derivative of the composite function $g \circ f$ at $\mathbf{a} = (1, 0)$.

The Jacobian for $f(x, y)$ is given by,

$$Jf(x, y) = \begin{bmatrix} 0 & 2 \\ 3 & 2y \end{bmatrix}$$

We can substitute $g(w, v)$ into $f(x, y)$ to get the function,

$$g(f(x, y)) = (4y^2(3x + y^2)^3, 6y - (3x + y^2)^2)$$

The Jacobian for $g(f(x, y))$ is,

$$Jg(f(x, y)) = \begin{bmatrix} 36y^2(3x + y^2)^2 & y^3(3x + y^2)^2 + 8y(3x + y^2)^3 \\ 6(3x + y^2) & 6 - 4y(3x + y^2) \end{bmatrix}$$

The Jacobians for $Jg(f(x, y))$ and $f(x, y)$ evaluated at $\mathbf{a}$ are,

$$Jg(f(\mathbf{a})) = \begin{bmatrix} 0 & 0 \\ 18 & 6 \end{bmatrix}, Jf(\mathbf{a}) = \begin{bmatrix} 0 & 2 \\ 3 & 0 \end{bmatrix}$$

Finally, the derivative of $g \circ f$ at point $\mathbf{a}$ is,

$$Jg(f(\mathbf{a}))Jf(\mathbf{a}) = \begin{bmatrix} 0 & 0 \\ 18 & 6 \end{bmatrix} \begin{bmatrix} 0 & 2 \\ 3 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 18 & 36 \end{bmatrix}$$

To summarize, the chain rule applied to a vector valued function nested in another vector valued function is just the product of the Jacobian of the outer function evaluated at the inner function and the Jacobian of the inner function. This can be easily extended to multiple nested functions.

4 Applications of Differentiation

4.1 Gradient, Divergence, and Curl

The *gradient* of $f : U \subset \mathbb{R}^n \to \mathbb{R}$ is the vector-valued function,

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \frac{\partial f}{\partial x_2}(\mathbf{x}), \ldots, \frac{\partial f}{\partial x_n}(\mathbf{x}) \right)$$

For example, consider the function $f : U \subset \mathbb{R}^3 \to \mathbb{R}$ where $f(\mathbf{x}) = x_1^2 x_2^2 + 2x_1 + 3x_3^3 x_1$. Then the gradient of $f(\mathbf{x})$ is,

$$\nabla f(\mathbf{x}) = (2x_1 x_2^2 + 2 + 3x_3^3, 2x_1^2 x_2, 9x_3^2 x_1)$$

Let $f : U \subset \mathbb{R}^n \to \mathbb{R}$, let $\mathbf{a} \in U$, and let $\mathbf{u}$ be the unit vector in $\mathbb{R}^n$. The *directional derivative* of f *at* $\mathbf{a}$ *in the direction* $\mathbf{u}$ is,

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

Recall that a unit vector is a vector with length 1 and $\hat{\mathbf{u}} = \frac{\mathbf{u}}{||\mathbf{u}||}$ is the unit vector in the direction of $\mathbf{u}$.

If $f : U \subset \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\mathbf{a}$ and $\mathbf{u}$ is a unit vector then the directional derivative $D_{\mathbf{u}} f(\mathbf{a})$ exists and,

$$D_{\mathbf{u}} f(\mathbf{a}) = \nabla f(\mathbf{a}) \cdot \mathbf{u}$$

Consider the function $f(x_1, x_2) = 3x_1^2 x_2 + x_2^2 x_1^2$. To find the directional derivative $f(x_1, x_2)$ at the point $(-1, 1)$ in the direction $(4, 8)$ we first find the gradient,

$$\nabla f(x_1, x_2) = (6x_1 x_2 + 2x_2^2 x_1, 3x_1^2 + 2x_2 x_1^2)$$

We can define $\mathbf{a} = (-1, 1)$ and $\mathbf{u} = (4, 8)$. A unit vector in the prescribed direction is,

$$\mathbf{u} = \frac{(4, 8)}{\sqrt{16 + 64}} = \frac{(4, 8)}{\sqrt{80}}$$

Then we have,

$$D_{\mathbf{u}} f(\mathbf{a}) \cdot \mathbf{u} = (-8, 5) \cdot \frac{(4, 8)}{\sqrt{80}} = \frac{8}{\sqrt{80}}$$

If $\nabla f(\mathbf{a}) \neq \mathbf{0}$ and the level curve of f through $\mathbf{a}$ has a tangent vector $\mathbf{t}$ at $\mathbf{a}$, then $\nabla f(\mathbf{a})$ is perpendicular to $\mathbf{t}$.

We can see this by considering the function $f : U \subset \mathbb{R}^2 \to \mathbb{R}$. Let $\mathbf{a}, \mathbf{x} \in \mathbb{R}^2$ and $f(x_1, x_2) = c$ be the equation for the level curve of f through $\mathbf{a}$. Parameterize the level curve by $\mathbf{x} = \mathbf{r}(t)$ where $\mathbf{r}(0) = \mathbf{a}$. We can substitute $\mathbf{r}(t)$ into f, which gives,

$$f(\mathbf{r}(t)) = c$$

Differentiating the equation gives,

$$\nabla f(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = 0$$

By definition when $t = 0$,

$$\nabla f(\mathbf{a}) \cdot \mathbf{r}'(0) = 0$$

Recall that two vectors are perpendicular if they are orthogonal (i.e. their dot product equals 0). Also recall that $\mathbf{r}'(0)$ is the tangent vector to the level curve at $\mathbf{a}$. It follows that if $\mathbf{r}'(0) \neq \mathbf{0}$, then $\nabla f(\mathbf{a})$ is perpendicular to the to the vector that is tangent to the level curve at $\mathbf{a}$.

Since $\nabla f(\mathbf{a})$ points in the direction of the largest increase in f, and if we have two level curves where $c_2 > c_1$, then $\nabla f(\mathbf{a})$ will point to $f(x_1, x_2) = c_2$.

These concepts can be extended to higher dimensions. For instance, given the function $f : U \subset \mathbb{R}^3 \to \mathbb{R}$, we will have a level surface (instead of a level curve) and the gradient of f at $\mathbf{a} \in \mathbb{R}^3$ is perpendicular to a plane (instead of a vector) that is tangent to the level surface at $\mathbf{a}$. Note that there exists a bundle of tangent vectors at $\mathbf{a}$ that are each perpendicular to the gradient of f.

To summarize, the gradient at some point $\mathbf{a}$ gives the direction of the greatest increase from point $\mathbf{a}$. It follows that the negative of the gradient will give the direction of the greatest decrease. (Note, moving perpendicular to the gradient will give no change.) The directional derivative is the dot product of the gradient and some specified direction from point a and will give you the rate of change of moving in that direction.

In addition to the total derivative of a vector field, we can specify the divergence and curl of a vector field. Consider the differentiable vector field $\mathbf{F} : U \subset \mathbb{R}^3 \to \mathbb{R}^3$. Also, let ∇ represent the vector $(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z})$

The *divergence* of $\mathbf{F}$ is the dot product,

$$\text{div}(\mathbf{F}) = \nabla \cdot \mathbf{F}$$

and the *curl* of $\mathbf{F}$ is the cross product,

$$\text{curl}(\mathbf{F}) = \nabla \times \mathbf{F}$$

Note that $\text{div}(\mathbf{F})$ is a scalar-valued function and $\text{curl}(\mathbf{F})$ is a vector field.

An example of applying divergence and curl to a vector field.

4.2 The Hessian Matrix

Consider the function $f : U \subset \mathbb{R}^n \to \mathbb{R}$ that has second-order partials at $\mathbf{a} \in U$. The *Hessian matrix* for f at $\mathbf{a}$ is,

$$Hf(\mathbf{a}) = \left[\frac{\partial^2 f}{\partial x_j \partial x_i}\right](\mathbf{a})$$

The *Hessian form for f at* $\mathbf{a}$ is the quadratic form defined by,

$$h(\mathbf{x}) = \mathbf{x}^\top Hf(\mathbf{a})\mathbf{x}$$

Just like the Jacobian is the matrix of first-order derivatives, the Hessian is the matrix of second-order derivatives. For a function of several variables, the Hessian is analogous to the second derivative of single variable function. The Hessian represents the second derivative of a function at point $\mathbf{a}$. We can also say that the Hessian matrix is the Jacobian matrix for the gradient of a function. More formally,

$$Hf(\mathbf{x}) = J\nabla f(\mathbf{x})$$

As an example, consider the function $f : U \subset \mathbb{R}^2 \to \mathbb{R}$ where $f(x, y) = 3x^2 + e^{2y} + xy^3$. The Hessian matrix of $f(x, y)$ for any $\mathbf{a} \in \mathbb{R}^2$ is,

$$Hf(\mathbf{a}) = \begin{bmatrix} 6 & 3y^2 \\ 3y^2 & 4e^{2y} + 6xy \end{bmatrix}(\mathbf{a})$$

The Hessian form is

$$h(x, y) = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 6 & 3y^2 \\ 3y^2 & 4e^{2y} + 6xy \end{bmatrix}(\mathbf{a}) \begin{bmatrix} x \\ y \end{bmatrix}$$

At the point $(-3, 2)$ the Hessian matrix is,

$$Hf(-3, 2) = \begin{bmatrix} 6 & 12 \\ 12 & 4e^4 - 36 \end{bmatrix}$$

and the Hessian form is,

$$\begin{aligned} h(x, y) &= \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 6 & 12 \\ 12 & 4e^4 - 36 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \\ &= \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 6x + 12y \\ 12x + (4e^4 - 36)y \end{bmatrix} \\ &= 6x^2 + 24xy + (4e^4 - 36)y^2 \end{aligned}$$

4.3 Local Extrema

Just like with single variable calculus, we are interested in the points that achieve a maximum or minimum of some function. In single variable calculus we are given some function $f : U \subset \mathbb{R} \to \mathbb{R}$. If $f'(a) = 0$ and $f''(a) > 0$ then $f(a)$ is a local minimum. If $f'(a) = 0$ and $f''(a) < 0$ then $f(a)$ is a

local maximum.

Similarly, for functions of several variables $f : U \subset \mathbb{R}^n \to \mathbb{R}$, if $\nabla f(\mathbf{a}) = \mathbf{0}$ then *a negative definite Hessian matrix means that $f(\mathbf{a})$ is a local maximum and a positive definite Hessian matrix means that $f(\mathbf{a})$ is a local minimum.* An *indefinite Hessian matrix means that $f(\mathbf{a})$ is a saddle point.* If a local extremum exists at point $\mathbf{a}$ then either $\nabla f(\mathbf{a}) = \mathbf{0}$ or the gradient of f is undefined at point $\mathbf{a}$. In this context, the point $\mathbf{a}$ is defined as a *critical point.*

The maximum/minimum values of functions of several variables require and understanding of positive/negative definiteness of the Hessian matrix. Let $\mathbf{M}$ be some real matrix and let $\mathbf{x}$ be a non-zero vector of real numbers. Then $\mathbf{M}$ is a positive definite matrix if the scalar $\mathbf{x}^\top \mathbf{M}\mathbf{x}$ is positive for every vector $\mathbf{x}$. Similarly, $\mathbf{M}$ is a negative definite matrix if the scalar $\mathbf{x}^\top \mathbf{M}\mathbf{x}$ is negative for every vector $\mathbf{x}$.

Just like in single variable calculus, we can specify a second derivative test for functions $f : U \subset \mathbb{R}^n \to \mathbb{R}$. Let $\mathbf{a} \in U$ and $\nabla f(\mathbf{a}) = \mathbf{0}$. We can define submatrices $H_1, H_2, \ldots, H_n$ of the Hessian matrix $Hf(\mathbf{a})$ as follows,

$$H_1 = \begin{bmatrix} h_{11} \end{bmatrix}, H_2 = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix}, \ldots, H_n = \begin{bmatrix} h_{11} & \ldots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{n1} & \ldots & h_{nn} \end{bmatrix}$$

We can apply these "subhessian matrices" to find local extrema as follows,

- f has a local minimum at $\mathbf{a}$ if,

$$|H_1| > 0, |H_2| > 0, \ldots, |H_n| > 0$$

- f has a local maximum at $\mathbf{a}$ if,

$$|H_1| < 0, |H_2| > 0, |H_3| < 0, |H_4| > 0, \ldots$$

- f has a saddle point at $\mathbf{a}$ if $|H_i| \neq 0$ for $i = 1, \ldots, n$ and the signs of the determinants do not follow the patterns for a minimum or maximum (defined above).
- There is no information about f at point $\mathbf{a}$ being a local extrema if $|H_i| = 0$ for any $i = 1, \ldots, n$.

As an example consider the function $f : U \subset \mathbb{R}^2 \to \mathbb{R}$ where $f(x, y) = x^2 + y^2$. To find the local extrema of this function we have to find the Hessian matrix. The Jacobian matrix is,

$$Jf(x, y) = \nabla f(x, y) = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$$

The Hessian matrix is,

$$Hf(x, y) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

It follows that $|H_1| = 2 > 0$ and $|H_2| = 4 > 0$, and the critical point $\mathbf{a} = (0, 0)$ is where the function reaches a minimum.

4.4 Constrained Optimization

Consider the functions $f : U \subset \mathbb{R}^n \to \mathbb{R}$ and $g : U\mathbb{R}^n \to \mathbb{R}$. Let $\mathbf{a}$ and λ^* be solutions to the following system of equations,

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

Also, let $W(\mathbf{x}, \lambda) = Hf(\mathbf{x}) - \lambda Hg(\mathbf{x})$. If $W(\mathbf{a}, \lambda^*)$ is positive definite, then $f(\mathbf{a})$ is a local minimum value of f along $g(\mathbf{x}) = 0$. $W(\mathbf{a}, \lambda^*)$ is negative definite, then $f(\mathbf{a})$ is a local maximum value of f along $g(\mathbf{x}) = 0$.

In information theory, entropy (specifically Shannon Entropy) is a measure of the disorder or uncertainty of information content. It is defined as,

$$E[I(X)] = -\sum_{i=1}^{n} P(x_i) \ln P(x_i)$$

where $P(x_i)$ refers to the probability of the random variable x_i.

Therefore, our constrained optimization problem is the following,

$$\begin{aligned} &\max \ E[I(X)] \\ &\text{where } \sum_{i}^{n} P(x_i) = 1 \end{aligned}$$

The gradients we need to find are,

$$\nabla E[I(X)] = -\begin{bmatrix} \ln P(x_1) + 1 \\ \ln P(x_2) + 1 \\ \vdots \\ \ln P(x_n) + 1 \end{bmatrix}$$

and,

$$\nabla \lambda \left(\sum_{i}^{n} P(x_i) - 1 \right) = \lambda \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

This gives us the following system of equations,

$$\begin{aligned} -\ln P(x_1) + 1 &= \lambda \\ -\ln P(x_1) + 1 &= \lambda \\ &\vdots \\ -\ln P(x_1) + 1 &= \lambda \\ \sum_{i}^{n} P(x_i) &= 1 \end{aligned}$$

So we have $n+1$ unknowns and $n+1$ equations. All but the last equation gives,

$$P(x_1) = P(x_2) = \ldots = P(x_n) = e^{1-\lambda}$$

We can substitute this into the last equation and solve for λ,

$$\begin{aligned} ne^{1-\lambda} &= 1 \\ e^{1-\lambda} &= \frac{1}{n} \\ 1-\lambda &= \ln\left(\frac{1}{n}\right) \\ \lambda &= 1 - \ln\left(\frac{1}{n}\right) \end{aligned}$$

Putting this result into $e^{1-\lambda}$ gives,

$$P(x_1) = P(x_2) = \ldots = P(x_n) = \frac{1}{n}$$

In order to find if this is the maximum of $E[I(X)]$ we need to evaluate the difference in Hessians, $W(\frac{1}{n}, \lambda^*)$. So we have,

$$HE[I(X)] = \begin{bmatrix} -\frac{1}{P(x_1)} & 0 & \ldots & 0 \\ 0 & -\frac{1}{P(x_2)} & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & -\frac{1}{P(x_n)} \end{bmatrix}$$

and,

$$\lambda H\left(\sum_i^n P(x_i) - 1\right) = \lambda \mathbf{0}_{n\times n}$$

It follows that,

$$W\left(\frac{1}{n}, \lambda^*\right) = HE[I(X)] - H\left(\sum_i^n P(x_i) - 1\right) = \begin{bmatrix} -\frac{1}{P(x_1)} & 0 & \ldots & 0 \\ 0 & -\frac{1}{P(x_2)} & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & -\frac{1}{P(x_n)} \end{bmatrix} = \begin{bmatrix} -n & 0 & \ldots & 0 \\ 0 & -n & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & -n \end{bmatrix}$$

This matrix is negative definite. So the values of $P(x_i) = \frac{1}{n}$ maximize entropy. The maximum value of entropy is,

$$E[I(X)] = -\sum_{i=1}^{n} \frac{1}{n} \ln\left(\frac{1}{n}\right) = -\ln\left(\frac{1}{n}\right) = -(\ln 1 - \ln n) = \ln n$$

A two-sided coin, for example, can produce a maximum entropy of $\ln(2) \approx 0.693$. On the other hand a coin that is heads on both sides produces an entropy of $\ln(1) = 0$. In other words, the coin with heads on both sides can be predicted with certainty so the entropy (i.e. measure of unpredictability) is zero. Since we are using the natural logarithm (logarithm with base e) the measurement of entropy is in nats. So the numbers 0-9 produce approximately 2.30 nats.

5 Integration

5.1 Paths and Arclength

A *path* is a continuous function $\mathbf{f}$ from $\mathbb{R}$ to $\mathbb{R}^n$. There are different types of paths. Define a path as $\mathbf{f} : [a, b] \to \mathbb{R}^n$. This means that $\mathbf{f}(t_1) \neq \mathbf{f}(t_2)$ for all t_1 and t_2 in (a,b) such that $t_1 \neq t_2$. A path is simple if it has no points of self-intersection except possible at the end points. A path is closed if $\mathbf{f}(b) = \mathbf{f}(a)$. A path is smooth if $\mathbf{f}'(t)$ is continuous and $\mathbf{f}'(t) \neq \mathbf{0}$. A path is piecewise smooth if there are finitely many subintervals of [a,b] on which the path is smooth.

A *curve* or *arc* is a set of points $C \in \mathbb{R}^n$ if there is a path $\mathbf{f} : I \to \mathbb{R}^n$ such that $C \subset \mathbf{f}(I)$.

The *length* or *pathlength* of a simple smooth path $\mathbf{f}(t)$ where $a \leq t \leq b$ is defined as,

$$\int_a^b ||\mathbf{f}'(t)||dt$$

5.2 Iterated Integrals

Recall that the definition of the definite integral for single variable functions is defined as,

$$\int_a^b f(x)dx = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i)\Delta x_n$$

where, $f : U \subset \mathbb{R} \to \mathbb{R}$ and Δx represents one of the n partitions that we divide the domain space into.

Double integrals are the two dimensional version of definite integrals for single variable functions. An application of double integrals is that allow you to determine the volume of a solid in $\mathbb{R}^3$ space.

We can extend the definition of the single integral to the double integral. Let $f : U \subset \mathbb{R}^2 \to \mathbb{R}$, and let R be a region in U. Then the area of the object defined by f is,

$$\int\int_R f(x, y)dA = \lim_{n,m\to\infty} \sum_{i=1}^{n}\sum_{j=1}^{m} f(x_i, y_j)\Delta x_i \Delta y_j$$

$$\int\int_R f(x, y)dA = \lim_{n\to\infty} \sum_{(x_i,y_i)\in R} f(x_i, y_i)\Delta A_n$$

Here, ΔA is the area differential. Let $R \in [a, b] \times [c, d]$. Also, let $\Delta x = \frac{b-a}{n}$ and let $\Delta x = \frac{d-c}{n}$, where n represents the number of equal subdivisions of the region R. Finally, $A_n = \Delta x \Delta y$.

An iterated integral is simply a series of integrals that you take over the function. In the case of the double integral we take two integrals over the function $f(x, y)$. The iterated integral is defined

with the limits of the integral specified in terms of one of the variables. Specifically, if $f(x, y)$ is continuous and the region $R = \{(x, y)|a \leq x \leq b, g_1(x) \leq y \leq g_2(x)\}$ where $g_1(x)$ and $g_2(x)$ are the lower and upper limits of the integral with respect to variable y, respectively, then,

$$\int\int_R f(x, y)dA = \int_a^n \int_{g_1(x)}^{g_2(x)} f(x, y)dydx$$

For example, in order to evaluate the integral,

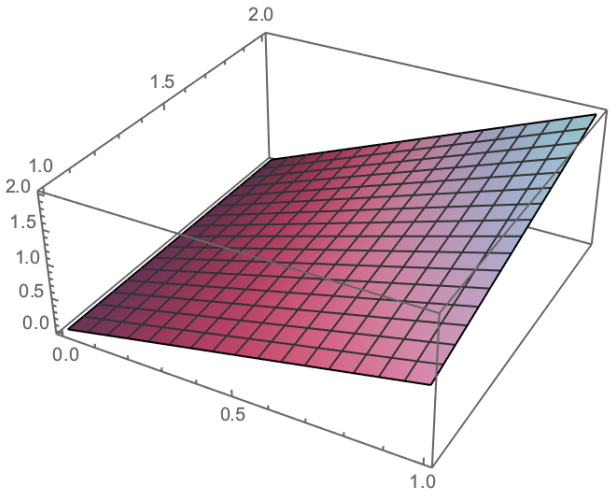
$$\int_0^1 \int_{1-x}^{1+x} xy \; dydx$$

we first evaluate the inner integral $\int_{1-x}^{1+x}$ with respect to y and then the outer integral $\int_0^1$ with respect to x. This will give us,

$$\begin{aligned}\int_0^1 \int_{1-x}^{1+x} xy \; dydx &= \int_0^1 \left[\frac{xy^2}{2}\right]_{1-x}^{1+x} dx \\ &= \int_0^1 2x \; dx \\ &= \left[x^2\right]_0^1 \\ &= 1\end{aligned}$$

Figure 7 is a graph of the function $f(x, y) = xy$ on which the double integral was performed. The double integral evaluates the area under this surface.

Figure 7: Plot of $f(x, y) = xy$



5.3 Example: Area Under Bivariate Normal Distribution

5.4 Line Integrals

We will first define the *line integral* and then, with the examples that follow, the interpretation of what a line integral is will be clear.

Consider a continuous function $f : U \subset \mathbb{R}^n \to \mathbb{R}$ over a smooth path C. The line integral of f along path C is,

$$\int_C f(x, y)ds = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i, y_i)\Delta s_i$$

where $\Delta s_i, \ldots, \Delta s_n$ represents the path C divided into n subarcs. Analogous to the integral of a single variable function, here we are dividing the path into n partitions and taking the limit as these partitions become infinitesimally small. If we're considering a curve in three-dimensional space, the line integral is just the area under this curve.

A more appropriate interpretation of the line integral is how it describes the relationship between the a vector field and a path on a vector field. The line integral is just the sum of the dot products of the vector field (i.e. a given point in the vector field) and the tangent vector along the path prescribed by the function (at that same point).

To formalize this interpretation, consider the vector field $\mathbf{F} : U \subset \mathbb{R}^n \to \mathbb{R}^n$ and the smooth path C in U which is parameterized by $\mathbf{x} = \mathbf{f}(t),\ a \leq t \leq b$. $\mathbf{F}$ takes the value $\mathbf{F}(\mathbf{f}(t))$ at each point on path C. At each change in t the objects change in position on path C is given by $\mathbf{f}'(x)\Delta t$. It follows that

$$\mathbf{F}(\mathbf{f}(t)) \cdot \mathbf{f}'(t)\Delta t$$

is the work done by the vector field $\mathbf{F}$ on an object on path C in the short interval Δt. Here work is defined as the product between the force exerted on an object at time t, which is given by $\mathbf{F}(\mathbf{f}(t))$, and the distance the object travels from time t to $t + \Delta t$, which is given by $\mathbf{f}'(t)\Delta t$.

Figure 8 provides an example of a vector field with a path overlaid. If we drop an object on the path then the properties of the vector field will move it along the path.

If we are interested in finding the total amount of work done by an object on the path then we can simply take the sum over all the changes in t,

$$\sum_{i=1}^{n} \mathbf{F}(\mathbf{f}(t)) \cdot \mathbf{f}'(t)\Delta t_i$$

The sum above is the discrete approximation of the following integral,

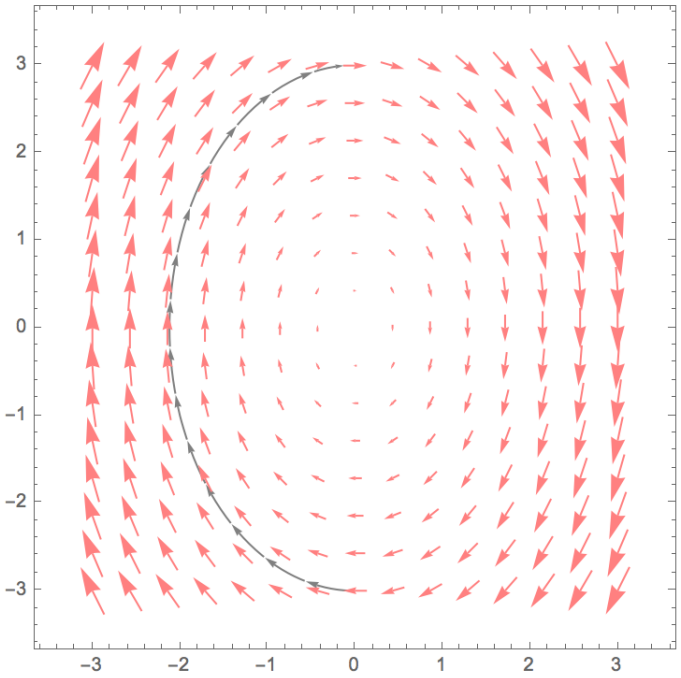
$$\int_{i=1}^{n} \mathbf{F}(\mathbf{f}(t)) \cdot \mathbf{f}'(t)\ dt$$

Formally, the *line integral* or *path integral* of a continuous vector field $\mathbf{F}$ over a smooth path C is,

$$\int_C \mathbf{F} \cdot d\mathbf{x} = \int_a^b \mathbf{F}(\mathbf{f}(t)) \cdot \mathbf{f}'(t)\ dt$$

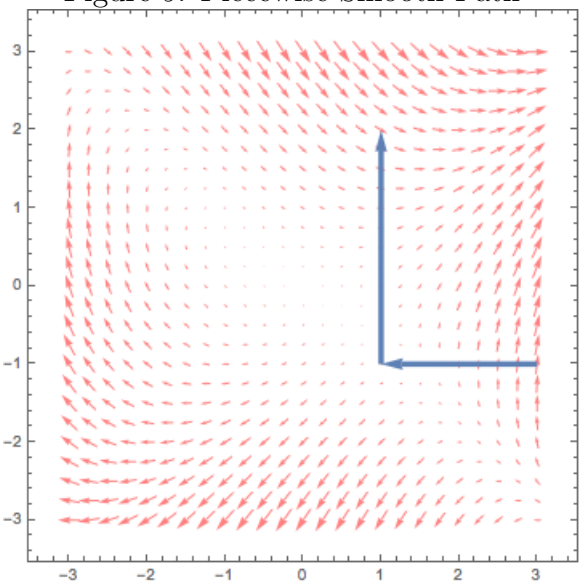
where $\mathbf{f}(t)$ is the parametrization of C, and $a \leq t \leq b$.

Figure 8: Vector Field with a Path



As an example, consider the vector field $\mathbf{F}(x, y) = (x + 2y, x^2 - y^2)$ and let the path consist of the line segments $(3, -1)$ to $(1, -1)$ and $(1, -1)$ to $(1, 2)$. This is illustrated in Figure 9.

Figure 9: Piecewise Smooth Path



In order to evaluate the line integral of the vector field over the path we have to first parameterize the path by a single variable t. The horizontal line segment can be defined as $(x, -1)$ and the vertical line segment can be defined as $(1, y)$. Therefore, for the horizontal line segment h we have the parametrization,

$$L_h(t) = (t, -1)$$

and for the vertical line segment v we have,

$$L_v(t) = (1, t)$$

Using the definition of the line integral (and evaluating each path separately) gives us the following,

$$\int_1^3 \mathbf{F}(L_h(t)) \cdot L_h'(t) = \int_1^3 (t-2, t^2-1) \cdot (1, 0) = \int_1^3 t - 2$$
$$\int_{-1}^2 \mathbf{F}(L_v(t)) \cdot L_v'(t) = \int_{-1}^2 (1-2t, 1-t^2) \cdot (0, 1) = \int_{-1}^2 1 - t^2$$

Taking the integral we have,

$$\int_1^3 t - 2 = \left[\frac{t^2}{2} - 2t\right]_1^3 = 0$$
$$\int_{-1}^2 1 - t^2 = \left[t - \frac{t^3}{3}\right]_{-1}^2 = 0$$

Recall that the line integral of the vector field along the path $(3, -1)$ to $(1, -1)$ to $(1, 2)$ is the sum of the line integral along $(3, -1)$ to $(1, -1)$ and the line integral along $(1, -1)$ to $(1, 2)$. Therefore, the line integral evaluated is zero. In the physics sense the work done by the force field on the object is zero along the path. This makes sense if you identify the orthogonality between the vector field and the tangent vector to the path in Figure 9. As the object travels along the path, the dot product between the tangent vector and the vector defined by the vector field at each point is equal to zero. Therefore, the sum of these dot products (i.e. the line integral) must also be zero.

If the vectors defined by the vector field are not perpendicular to the tangent vectors of the path at each point t then there is either positive work (the vector field is directing the object along the path, or there is negative work (the vector field is directing the object in a direction that opposes the path.

In Figure 10 a new path has been introduced to the vector field and path given in Figure 9. This new path is defined by the function $f(x) = \frac{3}{2}x + \frac{7}{2}$, where $1 \leq x \leq 3$.

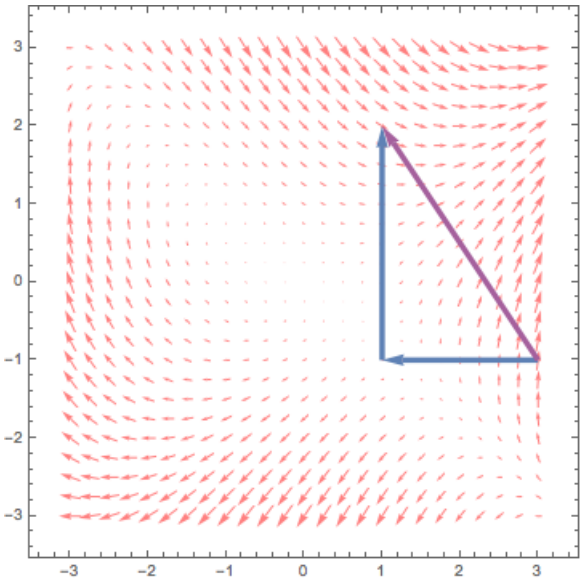
We can evaluate the work done by the vector field along this path. First note that the curve has already been parametrized in terms of one variable: $f(x) = \frac{3}{2}x + \frac{7}{2} \Leftrightarrow f(t) = \frac{3}{2}t + \frac{7}{2}$. The vector field is identical to the previous example and is defined as $\mathbf{F}(x, y) = (x + 2y, x^2 - y^2)$. We can define $f(t)$ as the vector valued function $\mathbf{f}(t) = (t, -\frac{3}{2}t + \frac{7}{2})$. Now we can evaluate the vector field at this function,

$$\mathbf{F}(\mathbf{f}(t)) = \left(-2t + 7, -\frac{5}{4}t^2 - \frac{42}{4}t + \frac{49}{4}\right)$$

and we can evaluate the derivative of the vector valued path function,

$$\mathbf{f}'(t) = \left(1, \frac{3}{2}\right)$$

Figure 10: A Piecewise Smooth Path and a Smooth Path



Now we have all the components required to evaluate the line integral,

$$\begin{aligned}\int_1^3 \mathbf{F}(\mathbf{f}(t)) \cdot \mathbf{f}'(t) &= \int_1^3 \frac{15}{8}t^2 + \frac{110}{8}t + \frac{91}{8}\ dt \\ &= \left[\frac{5}{8}t^3 + \frac{55}{8}t^2 + \frac{91}{8}t\right]_1^3 \\ &= 94\end{aligned}$$

So with regard to this path, the work done by the vector field along the path is positive. Looking at Figure 10 this seems like the appropriate conclusion since there are few points along the path at which the dot product between the vector field and the gradient along the path would be negative or zero.

6 Fundamental Theorems

6.1 Fundamental Theorem of Calculus

From single variable calculus we have the *Fundamental Theorem of Calculus*, which connects integration to differentiation by showing that integration is anti-differentiation.

Let $f(x)$ be a continuous function. From the definition of integration we can define the area under a curve as a function of one of the integral boundaries,

$$A(x) = \int_a^x f(t)dt$$

Using the definition of differentiation we can then define the rate of change of this area,

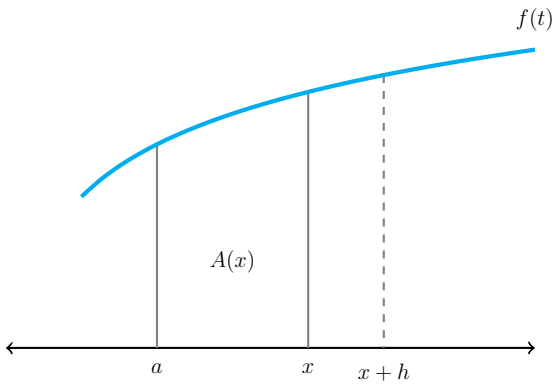
$$\frac{d}{dx}A(x) = \lim_{h\to 0}\frac{A(x+h) - A(x)}{h} = f(t)$$

It follows that,

$$\frac{d}{dx}\int_a^x f(t)dt = f(t)$$

This is the first part of the Fundamental Theorem of Calculus and shows that integration is just anti-differentiation. With respect to the notation above, $A(x)$ is the anti-derivative of $f(x)$. This is illustrated in Figure 11.

Figure 11: Fundamental theorem of Calculus - Part 1



From integration we know that,

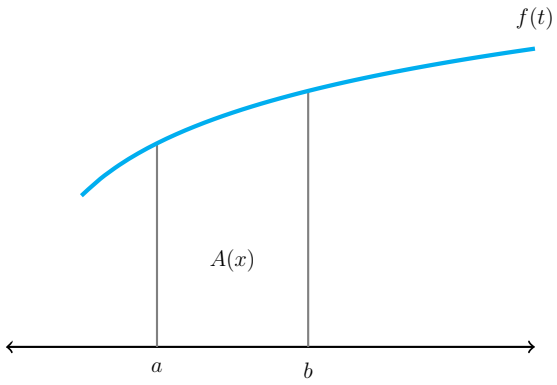
$$\int_a^b f(t)dt = A(b) - A(a)$$

where $a < b$, $A(b)$ and $A(a)$ are indefinite integrals (which is analogous to saying they are anti-derivatives of $f(b)$ and $f(a)$, respectively). Substituting the fact that $A'(x) = f(t)$ from the first part of the fundamental theorem we have,

$$\int_a^b A'(x)dx = A(b) - A(a)$$

This is the second part of the Fundamental Theorem of Calculus. It uses the first part of the fundamental theorem to show that integration evaluates the area under a curve between the two integration bounds. This is illustrated in Figure 12.

Figure 12: Fundamental Theorem of Calculus - Part 2



6.2 Fundamental Theorem of Line Integrals

Intuitively, one might think that some paths that take an object from point A to point B will require more/less work than other paths, as illustrated in Figure 13. However, for certain vector fields this not the case. If the vector field is the gradient of a scalar valued function then it is path independent (i.e. all path that takes you from point A to point B will yield the identical amount of work). This is established in the *Fundamental Theorem of Line Integrals*.

This fundamental theorem requires the property of *connected sets*. A set $A \in \mathbb{R}^n$ is connected if, for any two points $\mathbf{a}$ and $\mathbf{b}$ in A, there exists a continuous path lying in A that joins the points $\mathbf{a}$ and $\mathbf{b}$.

Fundamental Theorem of Line Integrals
Let $U \subset \mathbb{R}^n$ be an open connected set and let $f : U \to \mathbb{R}$ be a function whose gradient is continuous on U. If C is any smooth continuous path lying in U that joins points $\mathbf{a}$ and $\mathbf{b}$ then,

$$\int_C \nabla f(\mathbf{x}) \cdot d\mathbf{x} = f(\mathbf{b}) - f(\mathbf{a})$$

To sketch out the proof, parametrize C by $\mathbf{g}(t) = \mathbf{x}$, where $\alpha \leq t \leq \beta$, and define $\mathbf{g}(\alpha) = \mathbf{a}$ and $\mathbf{g}(\beta) = \mathbf{b}$.

As an aside, recall that parametrization means that we are taking the scalar valued function with multiple variables $f(\mathbf{x})$ and translating it into the vector valued function $\mathbf{g}(t)$ with a single variable. For example we are taking the following function,

$$f(\mathbf{x}) = f(x_1, x_2) = x_1 + 2x_1x_2$$

and parametrizing it to,

$$\mathbf{g}(t) = \begin{bmatrix} t + 2x_2 t & x_1 + 2x_1 t \end{bmatrix}$$

Notice that,

$$\nabla f(\mathbf{x}) = \frac{d}{dt}\mathbf{g}(t) = \begin{bmatrix} 1 + 2x_2 & 2x_1 \end{bmatrix}$$

Back to the proof. Applying the parametrization we have,

$$\int_C \nabla f(\mathbf{x}) \cdot d\mathbf{x} = \int_\alpha^\beta \nabla f(\mathbf{g}(t)) dt$$

Since we are taking the gradient with respect to only one variable, the above is identical to,

$$\int_C \nabla f(\mathbf{x}) \cdot d\mathbf{x} = \int_\alpha^\beta \frac{d}{dt} f(\mathbf{g}(t)) dt$$

Applying the Fundamental Theorem of Calculus (second part) we have,

$$\begin{aligned} \int_C \nabla f(\mathbf{x}) \cdot d\mathbf{x} &= \int_\alpha^\beta \frac{d}{dt} f(\mathbf{g}(t)) dt \\ &= f(\mathbf{g}(\beta)) - f(\mathbf{g}(\alpha)) \\ &= f(\mathbf{b}) - f(\mathbf{a}) \end{aligned}$$

The Fundamental Theorem of Line Integrals relied on the assumption that f is a scalar valued function with a continuous gradient on U. In other words this assumption is saying that the vector field $\mathbf{F}$ must be the gradient of some function, i.e. $\mathbf{F} = \nabla f(\mathbf{x})$. If this is true then the line integral (i.e. the work done by the object along the path) is simply the difference between the endpoints evaluated at the function f (known as the *potential function*) whose gradient is $\mathbf{F}$. Such vector fields are known as *conservative vector fields* or *path independent vector fields*.

For example, consider the vector field,

$$\mathbf{F} = (\underbrace{1 + 2x_2}_{F_1}, \underbrace{2x_1 + 2x_2}_{F_2})$$

We can determine the potential function f of this vector field by finding,

$$\int F_1 dx_1 = x_1 + 2x_1 x_2 + C$$
$$\int F_2 dx_2 = 2x_1 x_2 + x_2^2 + C$$

The common terms for F_1 and F_2 are $2x_1 x_2 + C$. In order to equate F_1 and F_2 we need to add x_2^2 to F_1 and add x_1 to F_2. This gives us the potential function of $\mathbf{F}$

$$f(\mathbf{x}) = x_1 + 2x_1 x_2 + x_2^2$$

As a check notice that $\mathbf{F} = \nabla f$.

This was pretty straightforward since the vector field in the above example was conservative by design. Unfortunately, it is not always easy to determine by inspection whether a vector field is conservative.

The Fundamental Theorem of Line Integrals gives us a quick way to evaluate a line integral. However, it relies on the assumption that the vector field is conservative and we need a way to check that this true.

One check is that if $\oint_C \mathbf{F} \cdot d\mathbf{x} = 0$ for all closed paths C that lie in the domain then the vector field $\mathbf{F}$ is conservative. This is an arduous task since it requires evaluating *all* closed paths in the domain to rule out the possibility that $\oint_C \mathbf{F} \cdot d\mathbf{x} \neq 0$.

The general check is to evaluate the Jacobian of the vector field and determine whether it is symmetric. If it is then the vector field is conservative. Recall that a matrix $\mathbf{M}$ is symmetric if $\mathbf{M} = \mathbf{M}^\top$.

Formally, let $U \in \mathbb{R}^n$ be a simply connected open set and let $\mathbf{F} : U \to \mathbb{R}^n$ be a continuously differentiable vector field. The vector field $\mathbf{F}$ is conservative iff the Jacobian matrix $J\mathbf{F}$ is symmetric, i.e. $J\mathbf{F} = (J\mathbf{F})^T$. (Note that a *simply connected set* is a set that has no holes (i.e. gaps) through it.)

- Path independence
- Irrotational
- Work done by a closed path is zero

Figure 13: Path Independence on a Conservative Vector Field