LECTURE NOTES FOR VECTOR CALCULUS (CALC 2, 3)

ALBERT M. FISHER

Contents

Date: May 16, 2024.

1. Review of Logic and Set Theory

2. Review of Linear Algebra

Let us recall:

Definition 2.1. We write $\mathbb{N}$ for the for the *natural numbers*, $\mathbb{N} = \{0, 1, 2, \dots\}$ and $\mathbb{Z}$ for the *integers*, $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$. $\mathbb{Q}$ denotes the rational numbers, $\mathbb{Q} = \{m/n : \ m, n \in \mathbb{Z}, n \neq 0\}$. $\mathbb{Z}$ and $\mathbb{Q}$ are both *rings*: there are two operations on it, addition and multiplication, denoted by $+$ and $\cdot$, satisfying the following axioms:
(1) There exists a unique number 0 such that $0 + x = x$ *(existence of an identity element for +)*
(2) $x + y = y + x$ *(the commutative law for +)* We note that another name for commutative is *abelian*.
(3) $(x + y) + z = x + (y + z)$ *(the associative law for +)*
(4) $1 \cdot x = x$ *(existence of an identity element for ·)*
(5) $x \cdot (y \cdot z) = (x \cdot y) \cdot z$ *(the associative law for ·)*
(6a) $x \cdot (y + z) = x \cdot y + x \cdot z$
(6b) $(x + y) \cdot z = x \cdot z + y \cdot z$ (the two *distributive laws*).

For a *commutative ring* the multiplication is also commutative.

Remark 2.1. We can either assume the identity element 0 is unique as in (1) and prove from that:
(1') there exists an additive inverse for each x, denoted $-x$, or we can drop the uniqueness assumption and replace it by $(1')$ and then *prove* uniqueness from that.

A *field* is a set K, with two operations $+$ and $\cdot$, which is a commutative ring and such that also:

(8) each $x \neq 0$ has a *multiplicative inverse*: there exists $\tilde{x}$ such that $x \cdot \tilde{x} = 1$. We write x^{-1} for this number.

The basic examples to keep in mind of fields and the only ones we use in this course, are $\mathbb{R}$ (the real numbers) and $\mathbb{C}$ (the complex numbers) Other examples are $\mathbb{Q}$ (the rational numbers) and the finite fields $\mathbb{Z}_p$ for p prime, and *extension fields* of $\mathbb{Q}$ such as $\mathbb{Q}[\sqrt{2}]$ (see any algebra text). There are few fields, and many rings! A *group* G is a set with only one operation which is associative but not necessarily commutative, and with inverses. This is usually called *multiplication* and is written $\cdot$; however if the group is commutative this operation may be written $+$. Thus the fields are commutative groups for $+$ but also their non-zero elements form a commutative group for multiplication. Any ring is a group for addition. $\mathbb{N}$ with addition gives an example of a *semigroup* as it does not have additive inverses.

There are many, many more groups than rings. Other examples are invertible $(n \times n)$ matrices with the operation of matrix multiplication. The *quaternions* are a four-dimensional example of something which is not quite a field, as multiplication is noncommutative, and the *octonions* are an eight-dimensional example which is even worse: multiplication is not even associative! Both of these find uses in Physics, and both can be represented by certian collections of matrices.

Abstract Algebra is the study of all these sorts of things: groups, rings, fields, vector spaces and much, much more. Each of these can be called an *algebraic theory.* For example, *group theory* consists of all the definitions and theorems for groups.

The basic idea of any mathematical theory is that we study certain *objects* (usually sets) satisfying certain properties called *axioms.* The axioms are the most basic true statements. We use calculations and intuition to try to guess at new true statements, and then endeavor to use the rules of logic to prove all these statements, called *theorems.* (Thus the axioms themselves are also theorems!)

One of the most important and useful algebraic theories (in math, as well as in physics) is the theory of *vector spaces*, called *Linear Algebra.*

What is magnificent about the axiomatic approach to math is that any theorem proved is then valid in all examples, which at first sight may seem completely different. So it is a very powerful approach. Also, many things become much clearer when presented this way.

For example, theorems proved for real vector spaces are valid (with the same proofs) for complex vector spaces, and of any dimension, including infinite dimension (although new axioms are introduced in those cases, to handle other issues: Hermitian inner products; convergence of infinite series in the case of infinite dimensions....)

We now give a brief introduction to Linear Algebra. An excellent text is [Axl97].

Definition 2.2. A *vector space* (sometimes called a *linear space*) *over the real numbers* $\mathbb{R}$ (sometimes over a different *field*, such as the complex numbers $\mathbb{C}$ or the rational numbers $\mathbb{Q}$) is a set V with two operations $+, \cdot$ (addition of vectors, multiplication of a vector by a real number) satisfying these properties, called the *vector space axioms.* The elements of the field are also called *scalars.*

(1) There exists a unique vector $\mathbf{0}$ such that $\mathbf{0} + \mathbf{v} = \mathbf{v}$ *(existence of an identity element for +)*

(2) $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$ *(the commutative law for +)*

(3) $(\mathbf{u}+\mathbf{v})+\mathbf{w}=\mathbf{u}+(\mathbf{v}+\mathbf{w})$ *(the associative law for* $+$*)*
(4) $0\cdot\mathbf{v}=\mathbf{0}$ *(multiplication by* 0*)*
(5) $1\cdot\mathbf{v}=\mathbf{v}$ *(existence of an identity element for* $\cdot$*)*
(6a) $a\cdot(\mathbf{v}+\mathbf{w})=a\cdot\mathbf{v}+a\cdot\mathbf{w}$
(6b) $(a+b)\cdot\mathbf{v}=a\cdot\mathbf{v}+b\cdot\mathbf{v}$ *(two distributive laws).*

Note that the associatve law allows us to define the sum of more than two vectors, as $\mathbf{u}+\mathbf{v}+\mathbf{w}+\mathbf{x}=((\mathbf{u}+\mathbf{v})+\mathbf{w})+\mathbf{x}=(\mathbf{u}+\mathbf{v})+(\mathbf{w}+\mathbf{x})$ and so on.

Exercise 2.1. Show that for each $\mathbf{v}\in V$ there exists a vector $\tilde{\mathbf{v}}$ such that $\tilde{\mathbf{v}}+\mathbf{v}=\mathbf{0}$: *(the existence of an additive inverse.)* Show that this is *unique*. See Remark 2.1 above!

Example 1. *Arrows in the plane, based at a point* $\mathbf{0}$*.* This is a purely geometrical definition. We are given a geometrical plane; we know what lines are and what length is. We draw a line segment from a point labelled $\mathbf{0}$ to another labelled $\mathbf{v}$. The arrowhead is located at $\mathbf{v}$. The collection of all such arrows is V. We define our operations: to form $\mathbf{u}=\mathbf{v}+\mathbf{w}$ we move in parallel the second vector $\mathbf{w}$ to the tip of $\mathbf{v}$; this gives the tip of $\mathbf{u}$.

Multiplication by the real number $a>0$ multiplies the length by a, and by -1 gives us the opposite arrow.

Exercise 2.2. A *linear combination* of two vectors $\mathbf{v},\mathbf{w}$ is a sum $a\mathbf{v}+b\mathbf{w}$. A linear combination of vectors $\mathbf{v}_1,\dots,\mathbf{v}_n$ is $a_1\mathbf{v}_1+\cdots+a_n\mathbf{v}_n$. We also write this as $\sum_1^n a_i\mathbf{v}_i$. Draw the picture for arrows, for a linear combination of 2 and of 3 vectors. Verify the axioms for V geometrically (by drawing pictures). No coordinates are used. Note that the same pictures work for arrows in 3 dimensions.

Example 2. *Euclidean space* $\mathbb{R}^n$*.* The points are n-tuples $\underline{x}=(x_1,x_2,\dots,x_n)$. We define addition in $\mathbb{R}^n$ by adding coordinates, similarly for multiplication by a scalar. Thus $(a,b)+(c,d)=(a+c,b+d)$ and $t(a,b)=(ta,tb)$.

Example 3. *$(m\times n)$ matrices, the space* $M_{m\times n}$*.* Addition and multiplication are defined coordinate-by-coordinate.

Exercise 2.3. Show the axioms hold for the last two examples.

Example 4. *A function space.* We define $\mathcal{C}^0([a,b],\mathbb{R})$ to be the collection of all continuous functions from the closed interval $[a,b]$ to $\mathbb{R}$. $\mathcal{C}^n([a,b],\mathbb{R})$ for $n\geq 1$ denotes the function with n derivatives such that the n^{th} derivative $f^{(n)}$ is continuous. We also write $\mathcal{C}$ for $\mathcal{C}^0$.

Exercise 2.4. Define addition and multiplication by a scalar in an appropriate way for the last example, and show the axioms hold.

Definition 2.3. Given two vector spaces V,W, a map (i.e. a function) T from V to W is called a *linear transformation* iff it preserves the operations, that is, iff
(i) $T(\mathbf{v}+\mathbf{w})=T(\mathbf{v})+T(\mathbf{w})$;
(i) $T(a\mathbf{v})=aT(\mathbf{v})$.

We can summarize this by:

(L) $T(a\mathbf{v} + b\mathbf{w}) = aT(\mathbf{v}) + bT(\mathbf{w})$. We call property (L) *linearity.*

Exercise 2.5. Show that the map from the space of arrows V to $\mathbb{R}^2$ defined by sending an arrow $\mathbf{v}$ with tip at location (a, b) to $(a, b) \in \mathbb{R}^2$ is a linear map. Show it is *bijective* i.e. it is $1 - 1$ and onto. This is called an *isomorphism* of the vector spaces. Isomorphic spaces like this are not the same, but we say they can be *identified* via the isomorphism.

Exercise 2.6. Show that $\mathbb{R}^n$ is isomorphic to the space of *column vectors* $M_{n\times 1}$ and also to the space of *row vectors* $M_{1\times n}$.

Show that a linear transformation $T : V \to W$ always sends $\mathbf{0}_V$ (the zero element in V) to $\mathbf{0}_W$.

Definition 2.4. A *subspace* U of a vector space V is a subset that is also a vector space, with the same operations. Show that a subset U is a subspace iff for all $\mathbf{v}, \mathbf{w} \in U$, then $\mathbf{v} + \mathbf{w} \in U$ and for any $a \in \mathbb{R}$, then $a\mathbf{v} \in U$. In other words, any linear combination of elements of U remains in U.

Exercise 2.7. Show that the intersection of two subspaces of V is a subspace. Show that this can be false for unions.

Definition 2.5. Given a collection of vectors $S \subseteq V$ we say the *vector space generated by* S is the smallest vector space that contains S. We denote this by $\langle S\rangle$. Show that this makes sense (hint: use the previous exercise). Show that $\langle S\rangle$ is the collection of all linear combinations of elements of S. We call $\langle S\rangle$ the *span* of S. Given a subspace $U \subseteq V$, we say $S \subseteq V$ *spans* or *generates* U iff $\langle S\rangle = U$.

Definition 2.6. A collection $S = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ of vectors in V is *linearly independent* iff
(def 1) $\sum_1^n a_i\mathbf{v}_i = \mathbf{0} \implies a_i = 0$ for all i.
(def 2) If $\mathbf{u} = \sum_1^n a_i\mathbf{v}_i$ and also $\mathbf{u} = \sum_1^n b_i\mathbf{v}_i$, then $a_i = b_i$ for all i. Thus, any vector $\mathbf{u}$ in the span of the vectors $\mathbf{v}_i$ has unique a_i, called the *coefficients* or *coordinates* of $\mathbf{u}$ with respect to these vectors.

A *basis* for $U \subseteq V$ is an ordered collection S of vectors in U which spans U and is linearly independent. Thus, any $\mathbf{u} \in U$ can be expressed as a linear combination of the vectors $(\mathbf{v}_1, \dots, \mathbf{v}_n)$ in S and this expression is unique. In particular $\mathbf{u}$ has well-defined (i.e. uniquely defined) coordinates in the basis S.

Theorem 2.1. *If V is a vector space and there exists a finite number of vectors which generate V, we say V is* finite-dimensional. *A finite dimensional vector space has a basis. The number of element in this basis is well-defined (does not depend on the choice of the basis).*

Proof. This takes some work: see [Axl97] for a nice proof. □

Definition 2.7. This number is defined to be the *dimension* of V.

Proposition 2.2. *Given a vector space V of dimension n, then V is isomorphic to $\mathbb{R}^n$.*

Proof. We express $\mathbf{u} \in V$ as a linear combination of the basis: $\mathbf{u} = \sum_1^n a_i \mathbf{v}_i$, and define $\Phi : V \to \mathbb{R}^n$ by $\Phi(\mathbf{v}) = (a_1, \dots, a_n)$. (Check that this is an isomorphism). □

Definition 2.8. (Definition of matrix products; see Lectures and handwritten notes.)

Exercise 2.8. (function spaces) Show that the collection of all Riemann integrable functions and the collection of all differentiable functions on $[a, b]$ are subspaces of $\mathcal{C}$. Show that differentiation and definite integrals define linear transformations. Show that the polynomials $\mathcal{P}_n$ of degree $\leq n$, for some $n \geq 0$, form a vector space of dimension $n + 1$. (Hint: find a basis!) Find a matrix which represents the derivative map from $\mathcal{P}_{n+1}$ to $\mathcal{P}_n$. Show the space $\mathcal{C}^0$ has infinite dimension.

Norms and inner products.

Definition 2.9. A *norm* $||\cdot||$ on V is a function with values in $\mathbb{R}$ which satisfies:
(*i*) $||a\mathbf{v}|| = |a| \cdot ||\mathbf{v}||$ (homogeneity);
(*ii*) $||\mathbf{v} + \mathbf{w}|| \leq ||\mathbf{v}|| + ||\mathbf{w}||$ (triangle inequality);
(*iii*) $||\mathbf{v}|| \geq 0$, and $||\mathbf{v}|| = 0$ iff $\mathbf{v} = \mathbf{0}$. (positive definiteness).

Example 5. For V the space of arrows in 2 or 3 dimensional space, the *standard norm* is the length of the line segment. For $\mathbb{R}^n$, the *standard norm* of $\mathbf{a} = (a_1, \dots, a_n)$ is $||\mathbf{a}|| = \left(\sum_1^n a_i^2\right)^{1/2}$. This is also called the *l^2-norm.*

Remark 2.2. The isomorphism from V to $\mathbb{R}^2$ or $\mathbb{R}^3$ preserves the norm.

Having a norm allows us to define a *metric* (a notion of *distance*) on V, with the distance between points defined by $d(\mathbf{v}, \mathbf{w}) = ||\mathbf{w} - \mathbf{v}||$.

Definition 2.10. An *inner product* is a function from $V \times V$ to $\mathbb{R}$, written $\mathbf{v} \cdot \mathbf{w}$ or $\langle \mathbf{v}, \mathbf{w} \rangle$, satisfying the following;
(1) $\mathbf{v} \cdot \mathbf{w} = \mathbf{w} \cdot \mathbf{v}$ (commutative law);
(2) $(a\mathbf{v}) \cdot \mathbf{w} = a(\mathbf{v} \cdot \mathbf{w})$ (associativity of scalar multiplication)
(3) $\mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}$ (distributive law)
(4*a*) $\mathbf{v} \cdot \mathbf{v} \geq 0$ and
(4*b*) If $\mathbf{v} \cdot \mathbf{v} = 0$ then $\mathbf{v} = \mathbf{0}$.

These imply that also:
(2′) $\mathbf{v} \cdot (a\mathbf{w}) = a(\mathbf{v} \cdot \mathbf{w})$.
(3′) $(\mathbf{u} + \mathbf{v}) \cdot \mathbf{w} = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}$.

Properties $(2, 2'), (3, 3')$ tell us that this is a *bilinear form*; (1), (4*a*), and (4*b*) add that the form is *symmetric*, *positive* and *positive definite.* Note that a positive definite bilinear form defines a norm, via

$$||\mathbf{v}|| \equiv (\mathbf{v} \cdot \mathbf{v})^{1/2}.$$

Example 6. The *standard inner product* for the space of *arrows*, V is

$$\mathbf{v} \cdot \mathbf{w} = ||\mathbf{v}||||\mathbf{w}|| \cos \theta$$

where in the plane θ is the angle from $\mathbf{v}$ to $\mathbf{w}$ measured in the counterclockwise direction.

Note that this would give the same number if we went in the opposite direction, since $\cos(\theta) = \cos(-\theta)$. This is lucky since the same definition works in space, where we don't know what "counterclockwise" means!!

The *standard inner product* for $\mathbb{R}^n$ is $(x_1, \ldots, x_n) \cdot (y_1, \ldots, y_n) = \sum_1^n x_i y_i$.

Exercise 2.9. Verify that these each satisfy the axioms, and also that they are equal for our isomorphism from V to $\mathbb{R}^2$ or $\mathbb{R}^3$.

Example 7. (Hilbert space). This is an example which comes up and also in any study of waves.

For this $L^2 = L^p([a,b])$ is defined to be the vector space of all $f : [a,b] \to \mathbb{R}$ such that $\int_{[a,b]} |f|^2 \, \mathrm{d}x < \infty$. We define the norm

$$||f||_2 = \left(\int_{[a,b]} |f|^2 \, \mathrm{d}x \right)^{\frac{1}{2}}.$$

Thus L^2 is the space of all function with finite norm.

We define an inner product by $f \cdot g = \langle f, g \rangle = \int_{[a,b]} fg \, \mathrm{d}x$.

Note that this gives the norm as above.

Hilbert space is infinite-dimensional; a basis is given by all functions of the form (taking $[a,b] = [0, 2\pi]$ for simplicity) $\sin(nx), \cos(nx)$. The expansion is called Fourier series. Here we have to use *infinite* linear combinations, and the key point is that the series mus converge. For this all to work properly, we should use a fancier integral than the Riemann integral: the *Lebesgue* integral, for which we needto study *Real Analysis* and in particular *Measure Theory*.

In applications to Physics (especially Quantum Mechanics) the field is $\mathbb{C}$, then we have to use the complex conjugate and define: $\int_{[a,b]} f\overline{g} \, \mathrm{d}x$. This gives a *Hermitian inner product* which has slightly different axioms.

(TO DO...)

3. Vector Calculus, Part I: Derivatives and the Chain Rule

3.1. **Metrics, open sets, continuity.** Let us recall:

Definition 3.1. A *norm* $||\cdot||$ on V is a function with values in $\mathbb{R}$ which satisfies:
(*i*) $||a\mathbf{v}|| = |a| \cdot ||\mathbf{v}||$ (homogeneity);
(*ii*) $||\mathbf{v} + \mathbf{w}|| \le ||\mathbf{v}|| + ||\mathbf{w}||$ (triangle inequality);
(*iii*) $||\mathbf{v}|| \ge 0$, and $||\mathbf{v}|| = 0$ iff $\mathbf{v} = \mathbf{0}$. (positive definiteness).

Given a set X, a *metric* on X is a function $d : X \times X \to [0, +\infty]$ satisfying:
(*i*) $d(x,y) = d(y,x)$ (symmetry);
(*ii*) $d(x,z) \le d(x,y) + d(y,z)$ (triangle inequality);
(*iii*) $d(x,y) \ge 0$, and $d(x,y) = 0$ iff $x = y$ (positive definiteness).

We then say that $(V, ||\cdot||)$, respectively (X, d), is a *normed space*, respectively a *metric space*.

Having a norm allows us to define a metric on V, with the distance between points defined by $d(\mathbf{v},\mathbf{w}) = ||\mathbf{w}-\mathbf{v}||$.

Exercise 3.1. Verify this!

A *topology* on X is a collection $\mathcal{T}$ of subsets of X (these will be called *open sets*) satisfying:
(i) $\emptyset, X \in \mathcal{T}$;
(ii) an arbitrary union of open sets is open;
(iii) a finite intersection of open sets is open.

A set $C \subseteq X$ is *closed* iff its complement $C^c = X \setminus C = \{x \in X : x \notin C\}$ is open.

The collection $\mathcal{F}$ of closed sets satisfies:
(i) $\emptyset, X \in \mathcal{F}$;
(ii) an arbitrary intersection of closed sets is closed;
(iii) a finite union of closed sets is closed.

These properties of $\mathcal{F}$ are equivalent to the corresponding properties for $\mathcal{T}$ via the laws for unions and intersections of complements of sets:

$(A\cap B)^c = A^c \cup B^c$, $(A\cup B)^c = A^c \cap B^c$

and more generally,

$$(\cap_{i\in I} A_i)^c = \cup_{i\in I}(A_i)^c;$$

$$(\cup_{i\in I} A_i)^c = \cap_{i\in I}(A_i)^c.$$

where I is some index set, for example $\mathbb{N} = \{0, 1, 2, \dots\}$ or even an uncountable set like $\mathbb{R}$.

Exercise 3.2. Verify these statements!

For a metric space a *limit point* of $A \subseteq X$ is $x \in X$ such that for each $r > 0$, there is some point of A in $B_r(x)$. For a metric spaces a set C is closed iff it contains all of its limit points. Thus for example (a, b) is not a closed set as a, b are limit points.

Having a metric allows us to define a *topology* on X, as follows.

Definition 3.2. Given a metric space $(X.d)$, the *(open) ball of radius* $r \in [0,\infty]$ *around* $x \in X$ is $B_r(x) = \{y \in X : d(x, y) < r\}$.

A set $\mathcal{U} \subseteq X$ is *open* iff, equivalently,
(i) $\mathcal{U}$ is a union of open balls;
(ii) for every $x \in \mathcal{U}$, $\exists r > 0$ such that $B_r(x) \subseteq \mathcal{U}$.

Exercise 3.3. Verify that this does give a topology.

Convergence and continuity.

Given a sequence $(x_n)_{\mathbb{N}\in\mathbb{N}}$, we say (x_n) *converges to* x, equivalently written $\lim_{n\to\infty} x_n = x$ or $x_n \to x$, iff for every open set $\mathcal{U}$ containing x, then for n sufficiently large, $x_n \in \mathcal{U}$. For a metric space, equivalently given $r > 0$, $\exists N$ such that for all $n > N$, $d(x_n, x) < r$ (since we can use balls of radius r).

Definition 3.3. Given two metric spaces (X_1, d_1) and (X_2, d_2), then $f : X \to Y$ is *coninuous* iff
(i) if $x_n \to x$ then $f(x_n) \to f(x)$. (This is the usual definition for $f : \mathbb{R} \to \mathbb{R}$.)
Equivalently, iff:
(ii) if $f(x_0) = y_0$, then given $\varepsilon > 0$, $\exists \delta > 0$ such that if $d(x, x_0) < \delta$ then for $y = f(x)$, then $d(y, y_0) < \varepsilon$. (This is the famous "$\varepsilon - \delta$"- definition.)
(iii)the inverse image of every open set is open.
This third definition works for the more general situation of two topological spaces, $(X, \mathcal{T})$ and $(Y, \mathcal{S})$.

Exercise 3.4.
(i) Use each of the three definitions to prove:

Proposition 3.1. *A composition of continuous functions is continuous.*

(ii) Show that (for metric spaces), all three definitions are equivalent. Hint: first prove $(i) \iff (ii)$, then $(ii) \iff (iii)$.

Remark 3.1. For some unusual non- metric topological spaces one has to replace sequences by so-called *nets* or *filters*.

3.2. **Curves.**

Definition 3.4. A *(parametrized) curve* in a vector space V is a function $\gamma : [a, b] \to V$.
The *image* of the curve is the image of this function, i.e. the collection of all values: $\{\gamma(t) : t \in [a, b]\}$. Thus the *parameter* t parametrizes the image.
The simplest example is a *parametrized line*; the curve $l(t) = \mathbf{p} + t\mathbf{v}$ where $\mathbf{p}, \mathbf{v}$ are elements of some vector space V. The image of l is a straight line in V; the parametrized line passes through the point $\mathbf{p}$ going in the direction $\mathbf{v}$.
Note that the image of a curve is different from the *graph*. Here we recall that by definition the *graph* of a function $f : X \to Y$ (where X, Y can be any sets) is $\text{graph}(f) \equiv \{(x, y) \in X \times Y : y = f(x)\} = \{(x, f(x) : x \in X\}$. (Here $X \times Y$ is the *product space*, defined to be the collection of all ordered pairs.)
Thus the graph of the curve γ in V is $\{(t, \gamma(t)) : t \in [a, b]\}$, a subset of $[a, b] \times V$. The image shows where you go on the curve, but not how fast or in what direction you go along this image. We see shortly how you can change the parametrization of a curve, keeping the same image.

Let us suppose that V has a norm defined on it. Then V is a metric space with $d(\mathbf{v}, \mathbf{w}) = ||\mathbf{v} - \mathbf{w}||$, so we know what it means for a curve to be *continuous.* We define the derivative of γ to be

$$\gamma'(t) = \lim_{h \to 0} \frac{\gamma(t + h) - \gamma(t)}{h}$$

if the limit exists; this is also called the *tangent vector* to γ at time t. See Fig. 5.

Lemma 3.2. *We have in coordinates:* $\gamma'(t) = (x_1'(t), \ldots, x_m'(t))$.

Proof. This is immediate from the definition. For example, in $\mathbb{R}^2$, for $\gamma(t) = (x, y)(t) = (x(t), y(t))$ then

$$\frac{\gamma(t+h)-\gamma(t)}{h} = \left(\frac{x(t+h)-x(t)}{h}, \frac{y(t+h)-y(t)}{h}\right) \to (x'(t), y'(t)).$$

□

Note that given a differentiable curve $\gamma : [a,b] \to \mathbb{R}^m$ then γ' is a second curve in $\mathbb{R}^m$. We can keep going and define the higher derivatives $\gamma'' = (\gamma')'$ and so on, all curves in $\mathbb{R}^m$.

The most common interpretation of the tangent vector of a curve comes from physics. There we interpret t as time and γ as position:

Definition 3.5. If $\gamma(t)$ represents the position of a particle at time t, then the derivative γ' (the tangent vector) gives the *velocity* of the particle $\mathbf{v}(t) = \gamma'(t)$, and the *acceleration* at time t is the vector $\mathbf{a}(t) = \mathbf{v}'(t) = \gamma''(t)$. Note that all of these are *vector quantities*, having both a magnitude and a direction. The *speed* is the magnitude of the velocity vector, the *scalar quantity* $||\mathbf{v}||$.

Now we prove some basic facts about curves and their derivatives:

Proposition 3.3. *(Leibnitz' Rule for curves) Given two differentiable curves $\gamma, \eta : [a,b] \to \mathbb{R}^m$, then $(\gamma \cdot \eta)' = \gamma' \cdot \eta + \gamma \cdot \eta'$.*

Proof. We just write the curves in coordinates, and apply Leibnitz' Rule (the Product Rule) for functions from $\mathbb{R}$ to $\mathbb{R}$. □

Proposition 3.4. *Let γ be a differentiable curve in $\mathbb{R}^m$ such that $||\gamma|| = c$ for some constant c. Then $\gamma \perp \gamma'$.*

Proof.
We use Leibnitz' Rule. We have $c = ||\gamma||^2 = \gamma \cdot \gamma$ so for all t,

$$0 = (\gamma \cdot \gamma)' = \gamma' \cdot \gamma + \gamma \cdot \gamma' = 2\gamma \cdot \gamma'$$

using commutativity of the inner product.

□

The meaning of $||\gamma|| = c$ is intuitively clear: for $\mathbb{R}^2$ this says that the curve is in a circle; for $\mathbb{R}^3$ that the image of the curve is in a sphere, and the statement is that the tangent vector to the curve is tangent to the sphere as it is perpendicular to the position vector. See Fig. 5.

Corollary 3.5. *If $\gamma : [a,b] \to \mathbb{R}^m$ is twice differentiable then if $||\gamma'||$ is constant, we have $\gamma' \perp \gamma''$.*

Proof. We just apply the Proposition to the curve γ'. □

Corollary 3.6. *If $\gamma : [a,b] \to \mathbb{R}^m$ is twice differentiable and represents the position of a particle at time t, then if the speed $||\gamma'||$ is constant, the acceleration is perpendicular to the curve (i.e. $\mathbf{a} \perp \mathbf{v}$).*

In other words if you are driving a car at a constant speed around a track, the only acceleration you will feel is side-to-side. If you apply the brakes or the accelerator pedal, a component vector of acceleration tangent to the curve will be added to this.

If we reparametrize a curve to have speed 1, then the magnitude of the acceleration vector can be used to measure how much it curves: we explain this next.

Proposition 3.4 allows us to make the following definition.

Definition 3.6. The *curvature* of a twice differentiable curve γ in $\mathbb{R}^n$ at time t is the following. For its unit-speed parametrization $\widehat{\gamma}(s)$ we define the curvature at time s to be $\widehat{\kappa}(s) = ||\widehat{\mathbf{a}}(s)||$; for γ the curvature at time t is $\kappa(t) = (\widehat{\kappa} \circ l)(t) = \kappa(t)$

For example, the curve $\gamma_r(t) = r(\cos t/r . \sin t/r)$ has velocity $\gamma_r'(t) = (-\sin t/r, \cos t/r)$ which has norm one; the acceleration is $\gamma_r''(t) = \frac{1}{r}(\cos(t/r). \sin(t/r)) = -\frac{1}{r^2}\gamma_r(t)$, with norm $\frac{1}{r}$. The curvature is therefore $\frac{1}{r}$. So if the radius of the next curve on the race track is half as much, you will feel twice the force, since by Newton's law, $F = m\mathbf{a}$! This is the physical (and geometric) meaning of the curvature.

3.3. **Arc length of a curve.** Given a curve $\gamma_1 : [c, d] \to \mathbb{R}^n$, by a *reparametrization* γ_2 of the curve we mean the following: we have an invertible differentiable function $h : [a, b] \to [c, d]$ with $h'(t) \neq 0$ for all t, such that $\gamma_2 = \gamma_1 \circ h$. Note that γ_1 and γ_2 have the same image, and that the tangent vectors are multiples: $\gamma_2'(t) = \gamma_1 \circ h'(t) = \gamma_1'(h(t))h'(t)$. We call this a *positive* or *orientation–preserving parameter change* if $h'(t) > 0$, *negative* or *orientation–reversing* if < 0.

Definition 3.7. We define the *arc length* of γ to be:

$$\int_a^b ||\gamma'(t)||\mathrm{d}t.$$

We introduce a special formula for this:

$$\int_\gamma \mathrm{d}s = \int_a^b ||\gamma'(t)||\mathrm{d}t.$$

As we shall explain below, ds is interpreted to mean *integration with respect to arc length*, and "$\int_\gamma$" is read "the integral over the curve γ", so all together this is read "the integral over the curve γ with respect to arc length".

For an example we already know from Calculus I, consider a function $g : [a, b] \to \mathbb{R}$, we consider its graph $\{(x, g(x)) : a \leq x \leq b\}$. We know the arc length of this graph is

$$\int_a^b \sqrt{1 + (g'(t))^2}\mathrm{d}x.$$

We claim that the new formula includes this one: parametrizing the graph as a curve in the plane $\gamma(t) = (t, g(t))$. Then $\gamma'(t) = (1, g'(t))$ so $||\gamma'(t)|| = \sqrt{1 + (g'(t))^2}$, whence indeed the arc length is $\int_\gamma \mathrm{d}s = \int_a^b \sqrt{1 + (g'(t))^2}\mathrm{d}x$ as claimed.

Proposition 3.7.
(i) The arc length of a curve is unchanged for any change of parametrization, independent of orientation. That is,

$$\int_{\gamma_1} ds = \int_{\gamma_2} ds.$$

Proof. (*i*) We have $h : [a,b] \to [c,d]$ so writing $u = h(t)$, then $\gamma_2 = \gamma_1 \circ h$ with $\gamma_1 : [c,d] \to \mathbb{R}^n$ and $\gamma_2 : [a,b] \to \mathbb{R}^n$. So $\gamma_2 \equiv \gamma_1 \circ h$ with $\gamma_2(t) = (\gamma_1 \circ h)(t) = \gamma_1(h(t)) = \gamma_1(u)$, so $\gamma_2'(t) = (\gamma_1 \circ h)'(t) = \gamma_1'(h(t))h'(t)$. Then $\mathrm{d}u = h'(t)\mathrm{d}t$, and using the Chain Rule, we have:

$$\int_{\gamma_2} \mathrm{d}s \equiv \int_{t=a}^{t=b} ||\gamma_2'(t)||\,\mathrm{d}t = \int_{t=a}^{t=b} ||(\gamma_1'(h(t))h'(t)||\,\mathrm{d}t$$

Assuming first that $h' > 0$, this equals (1)

$$= \int_{t=a}^{t=b} ||(\gamma_1'(u)||\,h'(t)\mathrm{d}t = \int_{u=c}^{u=d} ||(\gamma_1'(u)||\,\mathrm{d}u = \int_{\gamma_1} \mathrm{d}s.$$

If instead $h' < 0$, then we have as before

$$\int_{\gamma_2} \mathrm{d}s \equiv \int_{t=a}^{t=b} ||\gamma_2'(t)||\,\mathrm{d}t = \int_{t=a}^{t=b} ||(\gamma_1'(h(t))h'(t)||\,\mathrm{d}t$$

and now because, since $h' < 0$, $h(b) = c, h(a) = d$,

$$= -\int_{t=a}^{t=b} ||(\gamma_1'(u)||\,h'(t)\mathrm{d}t = -\int_{u=d}^{u=c} ||(\gamma_1'(u)||\,\mathrm{d}u = \int_{u=c}^{u=d} ||(\gamma_1'(u)||\,\mathrm{d}u = \int_{\gamma_1} \mathrm{d}s. \tag{2}$$

□

We next see how this can be used to give a *unit speed parametrization* of a curve $\gamma : [a,b] \to \mathbb{R}^n$. Set $l(t) = \int_a^t ||\gamma'(r)||\,\mathrm{d}r$, so $l(t)$ is the arclength of γ from time a to time t. Let us denote the length of γ by L. Thus function l maps $[a,b]$ to $[0,L]]$. Note that l is a primitive (antiderivative) for $||\gamma'||$ so $l'(t) = ||\gamma'(t)||$. We shall assume that $||\gamma'(t)|| > 0$ for all t; in this case, the function l is invertible. Our parameter change will be given by its inverse, $h(t) = l^{-1}(t)$; then h' is also positive.

Proposition 3.8. *Assume that* $||\gamma'(t)|| > 0$ *for all* t. *Then the reparametrized curve* $\widehat{\gamma} = \gamma \circ h$ *has speed one.*

Proof. Now $(l \circ h)(t) = t$ so $1 = (l \circ h)'(t) = l'(h(t)) \cdot h'(t)$. We have (by the Fundamental Theorem of Calculus) that $l'(t) = ||\gamma'(t)||$ so $l'(h(t)) = ||\gamma'||(h(t)) = ||\gamma'(h(t))||$ since $h'(t) > 0$. Thus

$$||\widehat{\gamma}'(t)|| = ||(\gamma \circ h)'(t)|| = ||\gamma'(h(t)) \cdot h'(t)|| = ||\gamma'(h(t))|| \cdot h'(t) = l'(h(t)h'(t) = 1.$$

□

The function l maps $[a,b]$ to $[0,L]]$ whence the parameter-change function h maps $[0,L]$ to $[a,b]$. We keep t for the variable in $[a,b]$ and define $s = l(t)$, the arc length up to time t, so now s is the variable in $[0,L]$ and $h(s) = t$.

The change of parameter gives $\widehat{\gamma}(s) = (\gamma \circ h)(s) = \gamma(h(s)) = \gamma(t)$. This indeed parametrizes the curve $\widehat{\gamma}$ is by arc length s.

Note further that

$$\int_{\gamma} \mathrm{d}s \equiv \int_a^b ||\gamma'(t)||\,\mathrm{d}t = \int_0^{l(b)} ||\widehat{\gamma}'(s)||\mathrm{d}s \equiv \int_{\widehat{\gamma}} \mathrm{d}s$$

From $s = l(t)$ we have $\mathrm{d}s = l'(t)\mathrm{d}t = ||\gamma'(t)||\mathrm{d}t$. Now we understand rigorously what is $\mathrm{d}s$: it represents the infinitesimal arc length; this helps explain the notation $\int_\gamma \mathrm{d}s$ for the total arc length.

3.4. **Level curves of a function.** We would like to visualize a function $F : \mathbb{R}^2 \to \mathbb{R}$. We do this in two main ways, by drawing the **graph** of the function (the subset $\{(x, y, z) : z = F(x, y)\}$) or by drawing the **level curves** of the function. The level curve of level $c \in \mathbb{R}$ is the following subset of the plane $\mathbb{R}^2$:

$$\{(x, y) : F(x, y) = c\}.$$

Remark 3.2. In geography, a *topographic map* of a region X shows the level curves of the altitude function $F(x, y)$ with $F : X \to \mathbb{R}$. See Fig. 1.

For a function $F : \mathbb{R}^3 \to \mathbb{R}$ we can no longer draw the graph (we would need four dimensions!) but can still draw the analogue of the level curves. These are the level *surfaces*. An example is $F(x, y, z) = x^2 + y^2 + z^2$ for which the level surfaces of level c^2 are the spheres of radius c. See §3.13.

Remark 3.3. In weather maps we see curves which could indicate constant pressure or temperature. These actual functions are defined on space (since height above the ground is also a variable) so are the part close to earth of these level surfaces; if the Earth were perfectly flat, these would be the level curves of G defined by $G(x, y) = F(x, y, 0)$, in other words, the level surfaces for F meet sea level $z = 0$ in the level curves for G.

3.5. **Partial derivatives; directional derivative.** Given a map $F : \mathbb{R}^2 \to \mathbb{R}$, choosing a point $\mathbf{p} \in \mathbb{R}^n$ then the *directional derivative* of F at $\mathbf{p}$ in the direction $\mathbf{u}$ is the following. Here we assume $||\mathbf{u}|| = 1$, i.e. $\mathbf{u}$ is a *unit vector*.

Above (Def. 3.4) we have defined the parametrized line $l(t) = \mathbf{p} + t\mathbf{u}$: the curve which starts at $\mathbf{p}$ and moves in the direction $\mathbf{u}$ at unit speed.

Now $f(t) = F(l(t))$ is a function from $\mathbb{R}$ to $\mathbb{R}$. We define

$$D_{\mathbf{u}}(F)|_{\mathbf{p}} = f'(0).$$

This gives the amount of increase of the function F in the direction $\mathbf{u}$ at that point.

A special case is for $\mathbf{u} = (1, 0)$. We define

$$\frac{\partial F}{\partial x}(\mathbf{p}) = D_{\mathbf{u}}(F)|_{\mathbf{p}}.$$

Similarly for $\mathbf{u} = (0, 1)$ we define

$$\frac{\partial F}{\partial y}(\mathbf{p}) = D_{\mathbf{u}}(F)|_{\mathbf{p}}.$$

See Fig. 4.

It is very easy to calculate the partial derivatives. For the partial with respect to x, we fix the variable y and find the derivative with respect to x alone.

For example, when $F(x, y) = x^2y^3$, then $\frac{\partial F}{\partial x} = 2xy^3$ while $\frac{\partial F}{\partial y} = 3x^2y^2$.

For $F : \mathbb{R}^n \to \mathbb{R}$ the definitions are similar.

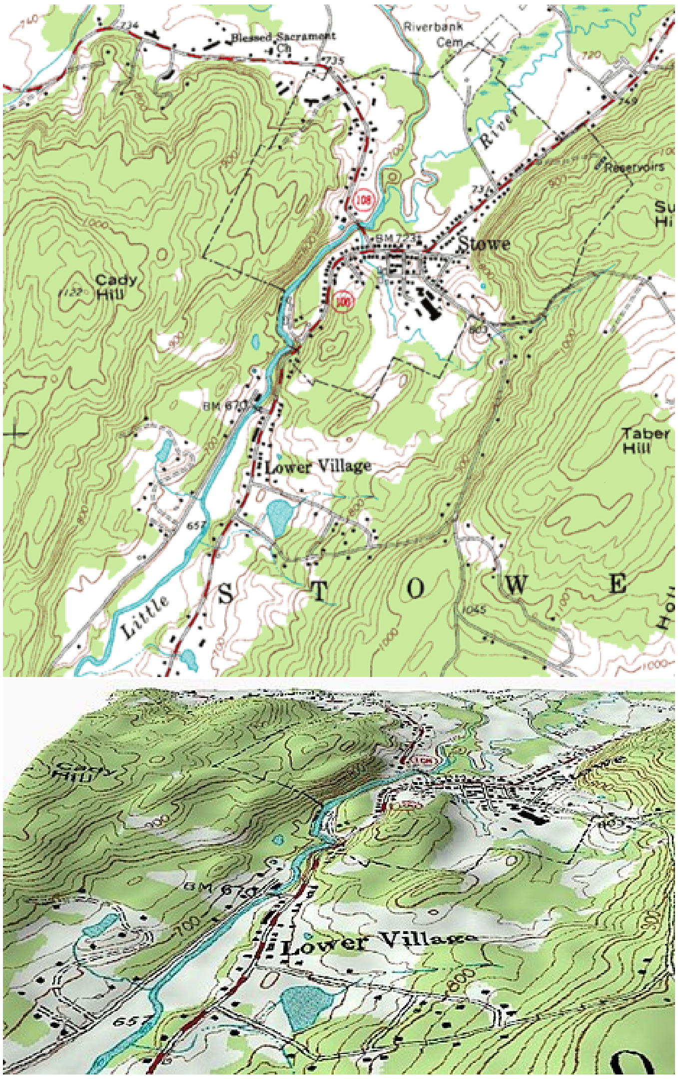


FIGURE 1. From Wikipedia, Topographic Map: a topographic map of the ski area of Stowe, Vermont and a shaded version of the map which helps to visualize the landscape.

3.6. **Properties of the gradient of a function.** Given a map $F : \mathbb{R}^m \to \mathbb{R}$ we define a vector at each point $\mathbf{p} \in \mathbb{R}^m$, $\nabla F = (\frac{\partial F}{\partial x_1}, \dots, \frac{\partial F}{\partial x_m})$, called the *gradient* of F at $\mathbf{p}$.

The term may be related to, for example, a road going up a steep *grade*.

As we shall see in the next sections, the gradient has the following important properties:
(1) This defines a *vector field*, called the *gradient vector field* of F.
(2) The gradient vector field is everywhere orthogonal to the *level sets* of F. These are level *curves* for $F : \mathbb{R}^2 \to \mathbb{R}$ and level *surfaces* for $F : \mathbb{R}^3 \to \mathbb{R}$. We prove this via the Chain Rule, see §3.12. In general, the level sets are *submanifolds*, i.e. differentiable

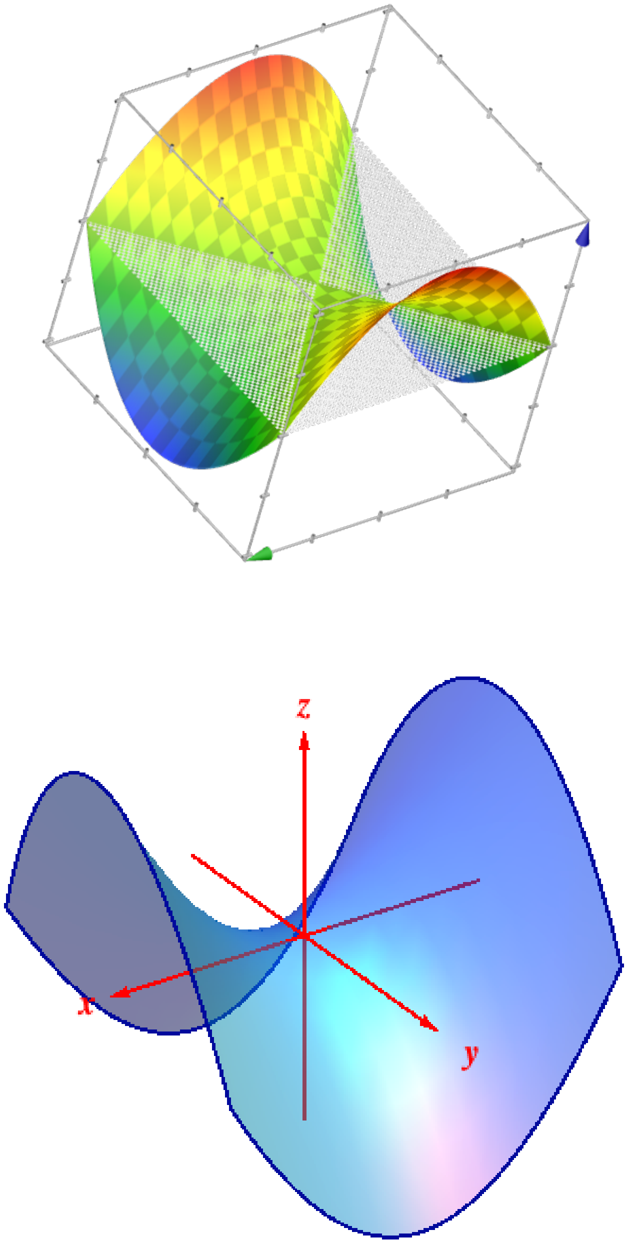


FIGURE 2. Graph of the function $F(x,y) = x^2 - y^2$ (a *parabolic hyperboloid*). From Google, search "$x^2 - y^2$" (rotatable image there) and from https://web.ma.utexas.edu/users/m408m/Display12-6-2.shtml. Horizontal slices (these project to the *level curves*) give a family of hyperbolas in the plane. Sliced vertically parallel to the x and y axes gives parabolas, sliced parallel to the lines $x = \pm y$ one way gives hyperbolas.

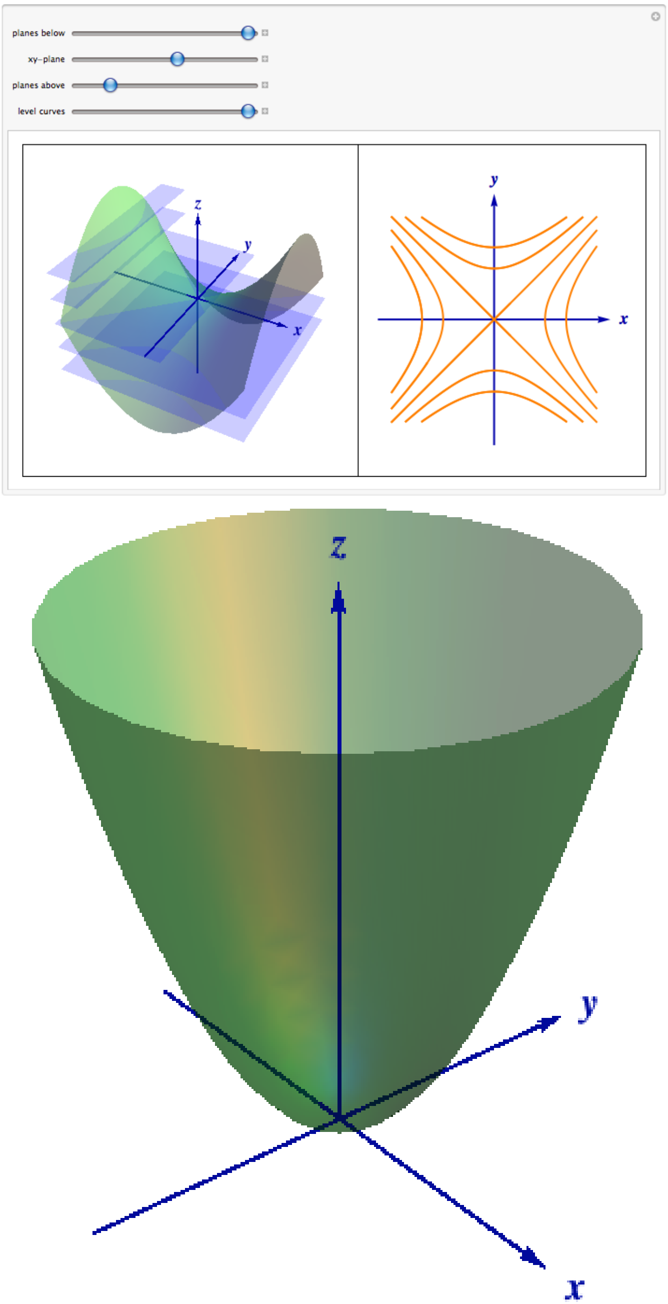


FIGURE 3. Graphs of parabolic hyperboloid with level curves (a family of hyperbolas), and of the paraboloid $F(x,y) = x^2 + y^2$).
From https://web.ma.utexas.edu/users/m408m/Display12-6-2.shtml.

subsets, of $\mathbb{R}^n$, of dimension $(n-1)$; this is a consequence of the *Implicit Function Theorem*, §3.19. (Here we have to assume that $\mathbf{n} \neq \mathbf{0}$).
(3) The gradient vector points in the direction of steepest increase of the function F at the point $\mathbf{p}$; its magnitude is the amount of increase in that direction.

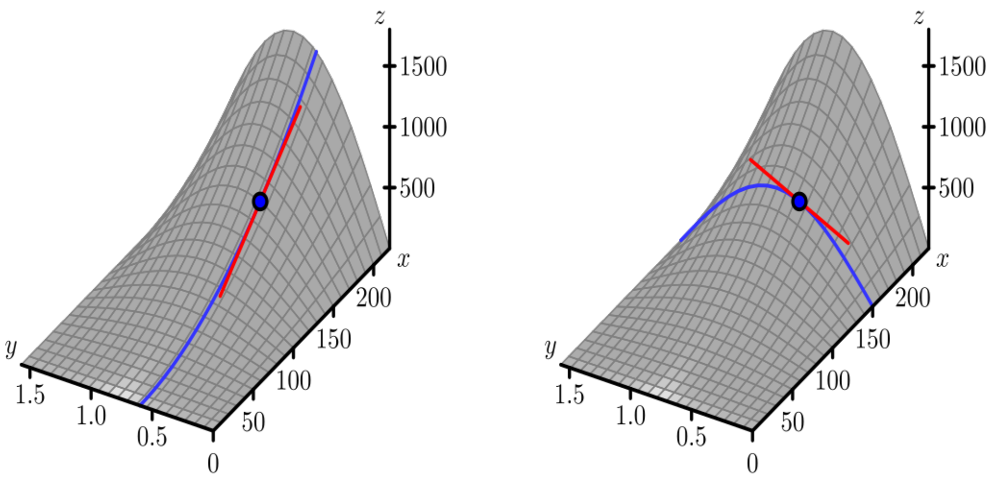


FIGURE 4. Partial derivatives; figure from https://activecalculus.org/vector/

(4) The directional derivative of F at $\mathbf{p}$, in the direction of the unit vector $\mathbf{u}$, is given simply by the inner product:

$$D_{\mathbf{u}}(F)|_{\mathbf{p}} = \nabla F|_{\mathbf{p}} \cdot \mathbf{u}.$$

(5) The gradient ∇F is the vector form of the *derivative* DF of the function F.
(6) The gradient can be used to easily write the equation of the *tangent line* to a level curve of $F : \mathbb{R}^2 \to \mathbb{R}$ at the point $\mathbf{p} = (x, y)$. For $F : \mathbb{R}^3 \to \mathbb{R}$, the gradient can be used to write the equation of the *tangent plane* to a level *surface* at a point $\mathbf{p} = (x_0, y_0, z_0)$. We explain this below in §3.12.

3.7. **Three types of curves and surfaces.** In the course we actually encounter three different (but related) types of curves and surfaces. First, recall:

Definition 3.8. For $f : [a, b] \to \mathbb{R}$ then its *graph* is

$$\text{graph}(f) = \{(x, f(x)) : x \in [a, b]\}$$

which is the subset of the plane we usually draw for this. Similarly, for $F : \mathbb{R}^2 \to \mathbb{R}$ then $\text{graph}(F) = \{(\mathbf{v}, F(\mathbf{v})) : \mathbf{v} = (x, y) \in \mathbb{R}^2\}$. Thus $\text{graph}(F) = \{(x, y, z) : z = F(x, y)\}$.

The different types of curves are:
(i) the graph of function $f : \mathbb{R} \to \mathbb{R}$;
(ii) a level curve of a function $F : \mathbb{R}^2 \to \mathbb{R}$;
(iii) a parametrized curve $\gamma : \mathbb{R} \to \mathbb{R}^m$.

Note that the first two are curves in the plane, while the second is a curve in $\mathbb{R}^m$ for any dimension m.

For surfaces we have, similarly:
(i) the graph of function $F : \mathbb{R}^2 \to \mathbb{R}$;
(ii) a level surface of a function $F : \mathbb{R}^3 \to \mathbb{R}$;
(iii) a parametrized surface $S : \mathbb{R}^2 \to \mathbb{R}^m$.

Now the first two are surfaces in $\mathbb{R}^3$, while the last is a surface inside of m-dimensional space so is much more general. Usually we will require that these functions be continuous and in the case of (iii) that the derivative DS exists, is continuous, and is surjective at every point. This guarantees the the image surface does not have creases or folds.

We also have *non*-parametrized curves and surfaces: the image of one of the functions in parts *(iii)*.

In both situations, curves and surfaces, these are all related, with (i) being a special case of (ii) and (ii) a special case of (iii). Regarding the passage from (ii) to (iii) see Exercise 3.7.

In all three cases it is important to first consider the linear (or affine) situation. That is because, first, it is the simplest case, and secondly, because these will describe the tangent line and the tangent plane, of a curve or surface, in all cases. These are exactly the affine lines or planes which bext approximate the curve or surface at a chosen point.

Here are the affine versions in all cases:
(i) An affine function $f : \mathbb{R} \to \mathbb{R}$ is $f(x) = y$ where $y = ax + b$.

An affine function $F : \mathbb{R}^2 \to \mathbb{R}$ has the form $F(x, y) = z$ where $z = ax + by + c$.

Note that the graph of f is a line in the plane, while the graph of F is a plane in $\mathbb{R}^3$.
(ii) The level curve of an affine function $F : \mathbb{R}^2 \to \mathbb{R}$: given $F(x, y) = ax + by$ then the level curve of level c is the line in the plane,

$$ax + by = c,$$

equivalently

$$Ax + By + C = 0,$$

for $A = a, B = b, C = -c$. This is now in the form of the *general equation of a line in the plane in* $\mathbb{R}^3$.

Given the affine function $F : \mathbb{R}^3 \to \mathbb{R}$ defined by given $F(x, y, z) = ax + by + cz$ then the level surface of of level d is the plane in $\mathbb{R}^3$:

$$ax + by + cz = d$$

or equivaelntly the *general equation of a plane*

$$Ax + By + Cz + D = 0,$$

for $A = a, B = b, C = c, D = -d$.
(iii) A parametrized affine curve $l : \mathbb{R} \to \mathbb{R}^m$: this is a parametrized line,

$$l(t) = \mathbf{p} + t\mathbf{v}.$$

A parametrized affine surface is $L : \mathbb{R}^2 \to \mathbb{R}^m$: this is a parametrized plane:

$$L(s, t) = \mathbf{p} + s\mathbf{v} + t\mathbf{w}.$$

Exercise 3.5.
$\widetilde{(i)}$ Which lines in the plane, or planes in $\mathbb{R}^3$, can (or cannot) be written as the graph of an affine function as in (i)?

$(\widetilde{ii})$ For which values of a, b or a, b, c in (ii) do you get a line or plane?
$(\widetilde{iii})$ For which vectors $\mathbf{v}$ and $\mathbf{v}, \mathbf{w}$ in (iii) do you get a line or plane?
$(\widetilde{iv})$ Make sure you know how to go from one type of line (or plane) to the other, whenever possible (see the Linear Algebra lecture notes and exercises)!
$(\widetilde{v})$ Write each type of line or plane in *matrix form.*
Solution: We explain $(\widetilde{ii})$. We claim that the equation $Ax + By + C = 0$ gives a line in the plane $\mathbb{R}^2$ exactly when not both A, B are 0. Here we have to understand the meaning of "gives the equation of a line in the plane."

There are two important points:
(1) This means that we are *in the plane*, this is our *Universe of Discourse* (we are talking only about points in the plane $\mathbb{R}^2$, not about $\mathbb{R}$ or $\mathbb{R}^3$).
(2) "gives the equation of a line" means that the collection of all solutions to the equation forms a line.

That is,
$$\{(x, y) \in \mathbb{R}^2 : Ax + By + C = 0\}$$
is a geometrical line in $\mathbb{R}^2$.

It makes a huge difference what is our Universe of Discourse (i.e. what we are talking about). For example, the equation $x = 2$ in $\mathbb{R}$ is a point, in $\mathbb{R}^2$ it is a vertical line, $\{(x, y) : x = 2\}$, in $\mathbb{R}^3$ it is a vertical *plane.*

Now for $Ax + By + C = 0$ to be a line means that
$$\{(x, y) \in \mathbb{R}^2 : Ax + By + C = 0\}$$
is a line. Let us consider the case where $B \neq 0$. Then this equation is *equivalent*, i.e. it has the same solutions:
$$y = -A/By - C/B = ax + b$$
which we know is a line.

Next suppose A, B are both 0. Then we have
$$\{(x, y) \in \mathbb{R}^2 : 0x + 0y + C = 0\}$$
equivalently
$$\{(x, y) \in \mathbb{R}^2 : C = 0\}$$
and there are two cases:
(i) $C = 0$: this statement is *true*, hence is true for all (x, y), so the solution set is all of $\mathbb{R}^2$;
(i) $C \neq 0$: this statement is *false*, hence is false for all (x, y), so there are no solutions, and the solution set is the empty set.

This proves the Claim.

Planes are handled similarly.

Exercise 3.6.
(i) Given vector spaces V, W and a linear transformation $T : V \to W$, prove that:

Proposition 3.9. *The image of T, Im(T) and the kernel (null space) of T,* $\ker(T)$ *are (vector) subspaces of W, V respectively.*

(ii) Interpret these statements for $T : \mathbb{R}^2 \to \mathbb{R}^m$ (for the image), respectively $T : \mathbb{R}^3 \to \mathbb{R}$ (for the kernel) as matrix equations, and describe the connection to the parametric and general equations of a plane.

Finding the tangent line or plane: as the graph of a function.

Next, for each of the three cases of curves and surfaces, we show how to find the tangent line, respectively the tangent plane. These are exactly the affine lines or planes which bext approximate the curve or surface at the point.

First, given a function $f : \mathbb{R} \to \mathbb{R}$ the formula for the tangent line to its graph is

$$l(x) = f(p) + f'(p)(x - p).$$

(Draw a picture!)

Note that $l(t)$ is itself a curve, and that it satisfies:
(i) it is an affine function, that is, linear plus a vector;
(ii)$l(p) = f(p)$;
(iii) $l'(p) = f'(p)$.

The formula for the tangent line to a (parametrized) curve $\gamma : \mathbb{R} \to \mathbb{R}^m$ is nearly identical:

$$l(t) = \gamma(p) + (t - p)\gamma'(p).$$

This now satisfies similar properties:
(i) it is an affine function;
(ii)$l(p) = \gamma(p)$;
(iii) $l'(p) = \gamma'(p)$.

For a level curve there are two ways to approach finding the tangent line. The first is to parametrize the level curve somehow and apply the previous case of a parametrized curve.

Exercise 3.7. For $F(x, y) = x^2 + y^2$, the curve of level 1 is the unit circle, the solutions of the equation (i.e. all pairs (x, y) which satisfy the equation)

$$x^2 + y^2 = 1.$$

Find parametrizations for this level curve, and use that to find the tangent line at a point.

Solution: We can parametrize this for example by the variable x. Then

$$y = \pm\sqrt{1 - x^2}$$

and we have two parametrized curves, with $t = x$, so $\gamma(t) = \pm\sqrt{1 - t^2}$. This works at all points except where $y'(x) = \infty$, that is, $x = \pm 1$. If we instead parametrize it by y then this works except where $x'(y) = 0$, that is, for $y = \pm 1$. We can also parametrize the entire curve at once, by the angle θ, with

$$\gamma(t) = (\cos t, \sin t)$$

and $t = \theta$.

When we parametrize by the variable x, we say the functions $f(x) = \sqrt{1-x^2}$, $\widetilde{f}(x) = -\sqrt{1-x^2}$, are *defined implicity* by the equation $x^2 + y^2 = 1$.

That is, they are *explicit* functions which are "implied" by the equation.

The *Implicit Function Theorem*, §3.19 describes when this can be done, basically when (partial) derivatives become infinite as above.

Given this, we can apply the formulas for the graph of a function, or for a curve in the plane, to find the tangent line of the level curve.

Finding the tangent line or plane: using the normal vector to find the tangent space. The second way to find the tangent line to a level curve is to find a normal vector to the curve. We explain this in §3.12.

Definition 3.9. Given $F : \mathbb{R}^n \to \mathbb{R}$ the *tangent space* $T_\mathbf{p}$ to the level set at the point $\mathbf{p} = (x_1, x_2, \dots, x_n)$, for level $c = F(\mathbf{p})$, is an affine subset of $\mathbb{R}^n$, all vectors $\mathbf{v}$ such that $(\mathbf{v} - \mathbf{p})$ is othogonal to the gradient, $\mathbf{n} = \nabla F_\mathbf{p}$.

We consider first the case of $n = 2$. We write the equation of the tangent line, recalling that given a point $\mathbf{p}$ and a normal vector $\mathbf{n} = (A, B)$ then the line passing through $\mathbf{p}$ and perpendicular to $\mathbf{n}$ is the collection of all $\mathbf{x} = (x, y)$ such that

$$\mathbf{n} \cdot (\mathbf{x} - \mathbf{p}) = 0.$$

Thus for $\mathbf{n} = (A, B)$ and $\mathbf{x} = (x, y)$ and $\mathbf{p} = (x_0, y_0)$ then

$$(A, B) \cdot (x - x_0, y - y_0) = 0$$

giving the general equation for the line,

$$Ax + By + C = 0$$

where $C = -\mathbf{n} \cdot \mathbf{p} = -(Ax_0 + By_0)$.

Since $\nabla F = \mathbf{n} = (\frac{\partial F}{\partial x}|_\mathbf{p}, \frac{\partial F}{\partial y}|_\mathbf{p})$ this gives the formula for the tangent line as

$$z_0 + \frac{\partial F}{\partial x}|_\mathbf{p}(x - x_0) + \frac{\partial F}{\partial y}|_\mathbf{p}(y - y_0) = 0 \tag{3}$$

We know the formula for the tangent line to the graph of a function $f : \mathbb{R} \to \mathbb{R}$ is $l(x) = f(p) + f'(p)(x - p)$. We can also use the normal vector method to find this formula in a second way. To do this we define $F : \mathbb{R}^2 \to \mathbb{R}$ by

$$F(x, y) = f(x) - y.$$

Then the level curve of level 0 gives $f(x) - y = 0$, so $y = f(x)$ which is the graph of f.

(Consider a simple example like $f(x) = x^2$ to understand what is going on!)

Note that at the point $\mathbf{p} = (p, f(p)$ is $\nabla F_\mathbf{p} = (f'(p), -1)$ so the formula (25) gives

$$\mathbf{n} \cdot (\mathbf{x} - \mathbf{p}) = 0$$

where $\mathbf{p} = (p, f(p))$ so we have

$$(f'(p), -1) \cdot ((x - p, y - f(p)) = 0$$

so

$$f'(p)(x - p) - (y - f(p)) = 0$$

so

$$y = f(p) + f'(p)(x - p)$$

as claimed.

Exercise 3.8. See Exercise 3.9 below.

3.8. **The gradient vector field; the matrix form of the tangent vector and of the gradient.** The gradient ∇F of a function $F : \mathbb{R} \to \mathbb{R}^m$ gives an important example of a *vector field*. In general, a *vector field* V *on* $\mathbb{R}^m$ is a function V from $\mathbb{R}^m$ to $\mathbb{R}^m$.

As we mentioned above and shall prove in Proposition 3.15, *the level curves of a function F are orthogonal to the gradient vector field*, so the gradient can help us understand the level curves of F.

We draw the vector $\mathbf{w_v} = V(\mathbf{v})$ *based at each point* $\mathbf{v}$. See Fig. 12.

The tangent vector gives the first definition of derivative of a curve $\gamma : \mathbb{R} \to \mathbb{R}^m$, the vector form of the derivative; the second definition, the *matrix form* of the derivative, is the $(n \times 1)$ matrix, i.e. the column vector with those same entries:

$$D\gamma = \begin{bmatrix} x_1'(t) \\ \vdots \\ x_n'(t)) \end{bmatrix}.$$

Thus $D\gamma : \mathbb{R} \to \mathcal{M}_{n\times 1}$.

For a function $F : \mathbb{R}^n \to \mathbb{R}$, the vector form of its derivative is the gradient ∇F. This has a matrix form, the row vector i.e. $(n \times 1)$ matrix with the same entries:

$$DF|_{\mathbf{x}} = \begin{bmatrix} \frac{\partial F}{\partial x_1} & \cdots & \frac{\partial F}{\partial x_n} \end{bmatrix}.$$

Given $\gamma : \mathbb{R} \to \mathbb{R}^m$ and $F : \mathbb{R}^m \to \mathbb{R}$ then the composition is $F \circ \gamma : \mathbb{R} \to \mathbb{R}$, so we can take its derivative $(F \circ \gamma)'(t)$. The *Chain Rule* says we can compute this in a second way. In vector notation it states:

$$(F \circ \gamma)'(t) = \nabla F_{\gamma(t)} \cdot \gamma'(t).$$

This is even simpler to remember in matrix notation, as we have the product of a row vector and a column vector. For example, with $\gamma : \mathbb{R} \to \mathbb{R}^3$ and $F : \mathbb{R}^3 \to \mathbb{R}$, we have

$$D(F \circ \gamma(t)) = \begin{bmatrix} F_x \, F_y \, F_z \end{bmatrix} |_{\gamma(t)} \begin{bmatrix} x_1'(t) \\ x_2'(t) \\ x_3'(t)) \end{bmatrix}.$$

Exercise 3.9. $F(x, y) = x^2y^3$; $\gamma(t) = (e^t, e^{t^2})$.
(1)Find $F \circ \gamma'(0)$.

First method (directly): $f(t) = F \circ \gamma(t) = e^{2t}e^{3t^2}$. $f'(t) = 2e^{2t}e^{3t^2} + 6t \cdot e^{2t}e^{3t^2}$. $f'(0) = 2$.

Second method (Chain Rule): $\nabla F = (2xy^3, x^2 3y^2)$. $\gamma'(t) = (e^t, t^2e^{t^2})$.
$\gamma(0) = (1, 1)$. $\nabla F(1, 1) = (2, 3)$. $\gamma'(0) = (1, 0)$.
So $F \circ \gamma'(0) = (2, 3) \cdot (1, 0) = 2$.

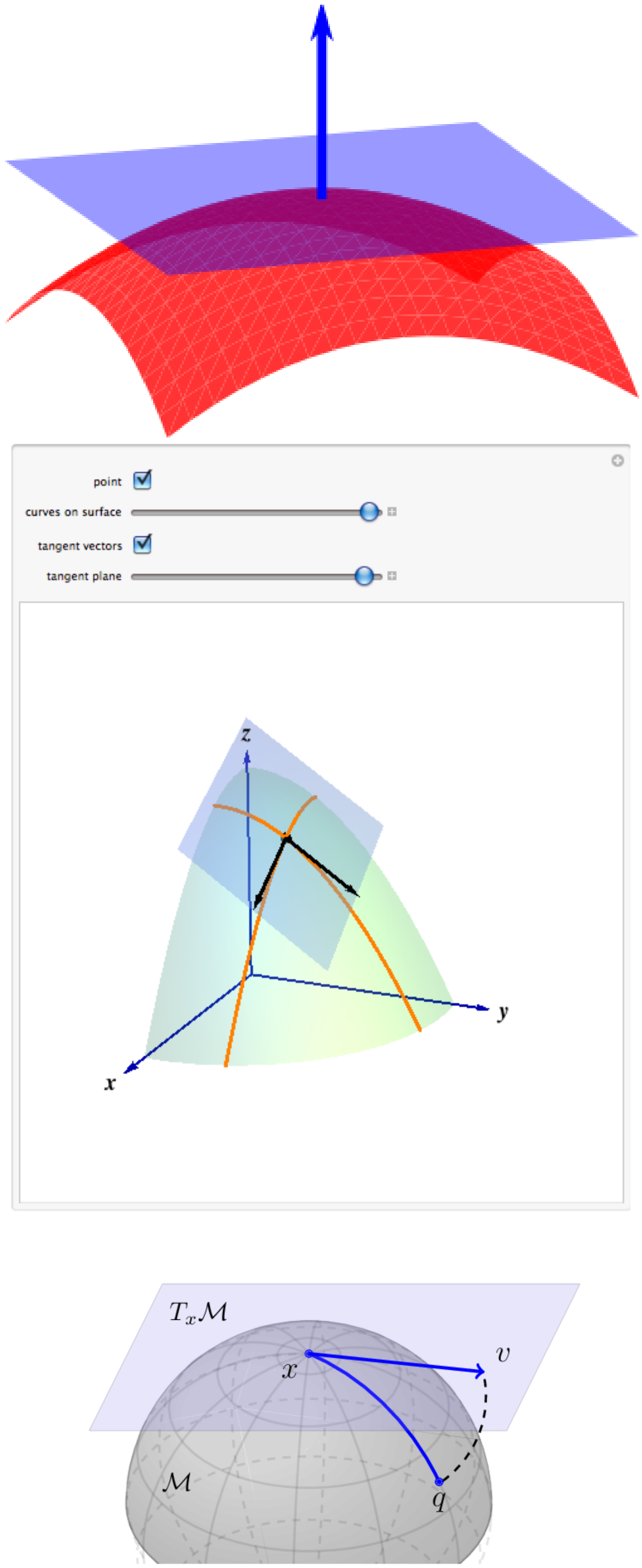


FIGURE 5. The normal vector to the surface is normal (orthogonal) to the tangent plane at that point. Tangent plane to the graph of a function defined on the plane, showing the meaning of the partial derivatives at the point. A tangent vetor to a curve in a surface is in the tangent plane. From https://web.ma.utexas.edu/users/m408m/Display14-4-2.shtml From https://www.researchgate.net/figure/Figure-S3-Geometric-illustration-of-tangent-vector-tangent-space-curve-and... From: Wikipedia, Normal (geometry)

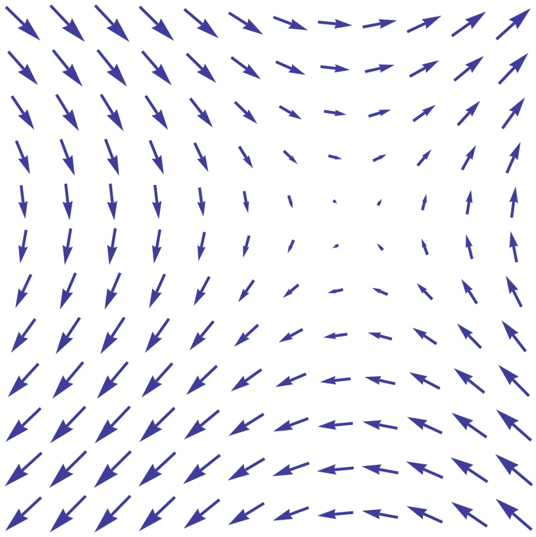


FIGURE 6. A vector field in the plane, from Wikipedia. Compare to the pictures of curves below!

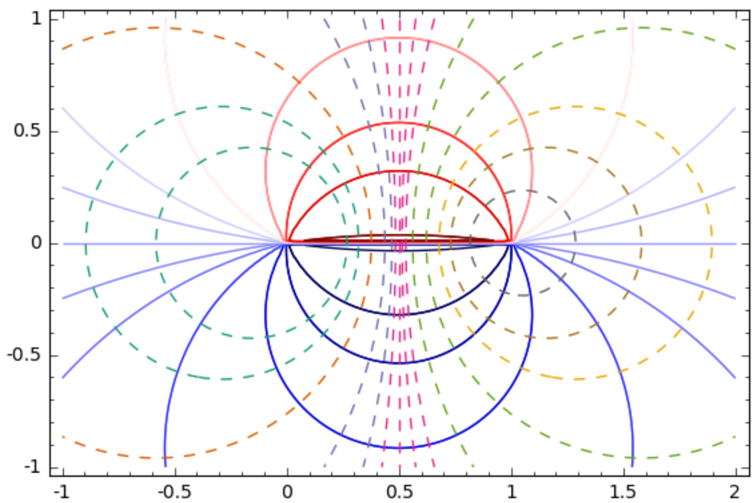


FIGURE 7. Equipotential curves for the electrostatic field of two opposite charges in the plane. Colors indicate different levels of the potential. This can also be interpreted as a gravitational field, where the potential function is height above sea level, and the positive charge is a mountain top while the negative charge is a valley. Orthogonal to the equipotentials are the lines of force; these are tangent to the gradient vector field of the potential function. One can imagine flowing along the lines of force from positive to negative charge, as in a fluid, although this is a force not a velocity field. (Because of this analogy with fluids, they are also called the *lines of flux* of the electrostatic field.

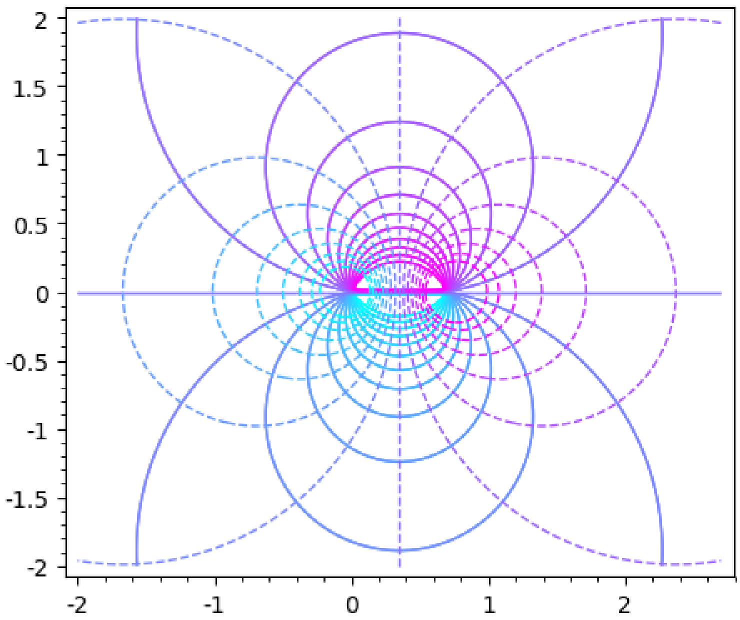


FIGURE 8. Equipotential curves and lines of force for the electrostatic field of two opposite charges in the plane, now closer together.

Which is easier depends on the problem!
(2)Find ∇F at $\mathbf{p} = (2, 5)$.
(3) Find the equation of the tangent plane to the graph of F at that point.

3.9. **General definition of derivative of a map.** Both the matrix version of tangent vector and of gradient are special cases of the general notion of derivative of a map from $F : \mathbb{R}^n \to \mathbb{R}^m$. We state this more generally for normed vector spaces.

Definition 3.10. Let V, W be Banach spaces (a vector space, possibly infinite-dimensional, on which we have a complete norm defined; *complete* here means that there are "no holes" as Cauchy sequences converge; this only can be an issue in infinite dimensions. The reader can think of $\mathbb{R}^n$ with the standard inner product and norm to get the basic idea). We say a function (or **map**) $F : V \to W$ is **differentiable** at the point $\mathbf{p} \in V$ iff there exists a linear transformation $L : V \to W$ such that

$$\lim_{\mathbf{h}\to\mathbf{0}} ||F(\mathbf{x}+\mathbf{h}) - F(\mathbf{x}+\mathbf{h}) - L(\mathbf{h})||/||\mathbf{h}|| = 0.$$

We write $DF_{\mathbf{p}} = L$ for this transformation, called the *derivative* of F at $\mathbf{p}$.

The idea here is that the derivative DF should give the "best linear approximation" at each point.

What this actually means is the linear part of the best *first-order approximation*. The best 0^{th}-order approximation at $\mathbf{x} \in \mathbb{R}^n$ is the constant map with the value at that point, thus the map $\mathbf{x} \mapsto \mathbf{p} = F(\mathbf{x})$. If $L = DF|_{\mathbf{x}}$, then the best first-order approximation will be the linear map shifted by this value, thus the affine map $(\mathbf{x} + \mathbf{v}) \mapsto \mathbf{p} + L\mathbf{v}$. (Literally what the term "first-order approximation" is "making use of the first derivative"; 0^{th}-order means approximating the function near $\mathbf{x}$ by its *value*, which is the stupidest approximation; note that this only helps at all for *continuous* functions. Similarly when the function has k continuous derivatives, we can use the derivatives at $\mathbf{x}$ from 0 to k to make a k^{th}-order approximation near that point. This gets better and better as k increases, as guaranteed by Taylor series at the point. See SS3.10 and §3.24.

Let us relate the above formula to the usual definition for a function $f : \mathbb{R} \to \mathbb{R}$, that is,

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

This definition still works for curves, giving us the tangent vector. However for V of dimension larger than 1 this makes no sense, as *we cannot take the ratio of two vectors.*

Remark 3.4. Or nearly. Consider the following: given a linear map $L : V \to V$, so $L\mathbf{v} = \mathbf{w}$, then in some sense

$$\frac{\mathbf{w}}{\mathbf{v}} = L :$$

the ratio "should be" a linear transformation!!

However L is not well-defined by this: many linear maps will solve the equation $L\mathbf{v} = \mathbf{w}$; it is only well-defined if V has dimension 1. What the definition of derivative requires is that L works for *all* directions $\mathbf{h}$, and this does make L well-defined.

Remark 3.5. Let us see what happens to the general definition for $f : \mathbb{R} \to \mathbb{R}$. Then

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

iff for each $\varepsilon > 0$, there exists $\delta > 0$ such that for $|h| < \delta$,

$$\left| \frac{f(x+h) - f(x)}{h} - c \right| < \varepsilon$$

or

$$\frac{\left| f(x+h) - f(x) - ch \right|}{|h|} < \varepsilon$$

And this is now a special case of the general formula.

We introduce the notation $\mathcal{L}(V, W)$ for the collection of all linear transformations from V to W. If we choose a basis for $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$, then $L : \mathbb{R}^n \to \mathbb{R}^m$ is represented by an $(m \times n)$ matrix. Then $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ can be identified with the matrices $\mathcal{M}_{mn} \sim \mathbb{R}^{mn}$, so $DF : \mathbb{R}^n \to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m) \sim \mathbb{R}^{mn}$.

When considering $F : \mathbb{R}^n \to \mathbb{R}^m$ then both $\mathbf{x} \in \mathbb{R}^n$ and $F(\mathbf{x}) \in \mathbb{R}^m$ can be written in components, with $\mathbf{x} = (x_1, \dots, x_n)$ and $F(\mathbf{x}) = (F_1(\mathbf{x}), \dots, F_m(\mathbf{x}))$. We write the components of F as a column vector, so

$$F(\mathbf{x}) = \begin{bmatrix} F_1(\mathbf{x}) \\ \vdots \\ F_m(\mathbf{x}) \end{bmatrix}.$$

Each component F_k is a function from $\mathbb{R}^n$ to $\mathbb{R}$ so has a gradient, $\nabla F_k = (\frac{\partial F_k}{\partial x_1}, \dots, \frac{\partial F_k}{\partial x_n})$. Now DF_k is by definition is the corresponding row vector so

$$DF_k = \begin{bmatrix} \frac{\partial F_k}{\partial x_1} & \dots & \frac{\partial F_k}{\partial x_n} \end{bmatrix}.$$

We define the *matrix of partials* of $F : \mathbb{R}^n \to \mathbb{R}^m$ to be the $(m \times n)$ matrix with the k^{th} row this gradient vector. Let us write $[\nabla F_k]$ for the row vector DF_k.

Then the ij^{th}-matrix entry is the partial derivative

$$(DF)_{ij} = \frac{\partial F_i}{\partial x_j}$$

and so we have the $(m \times n)$ matrix

$$DF|_{\mathbf{x}} = \begin{bmatrix} [\nabla F_1] \\ \vdots \\ [\nabla F_m] \end{bmatrix} = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \dots & \frac{\partial F_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial F_m}{\partial x_1} & \dots & \frac{\partial F_m}{\partial x_n} \end{bmatrix}.$$

The most basic cases are $f : \mathbb{R}^1 \to \mathbb{R}^m$ and $\mathbb{R}^n \to \mathbb{R}^1$. The first is a curve, discussed above, and usually written $\gamma : \mathbb{R} \to \mathbb{R}^m$. The general formula then gives the matrix form of the tangent vector; since γ is a column vector, with

$$\gamma(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_m(t)) \end{bmatrix}$$

then $D\gamma$ is the $(m \times 1)$ matrix with the same entries as the tangent vector $\gamma'(t) = (x_1', , \dots, x_m')(t)$, so

$$D\gamma|_t = \begin{bmatrix} x_1'(t) \\ \vdots \\ x_m'(t)) \end{bmatrix}.$$

The second type of map $F : \mathbb{R}^n \to \mathbb{R}$ we call simply a *function*. The general formula above then gives the $(1 \times n)$ matrix:

$$DF|_{\mathbf{x}} = \begin{bmatrix} \frac{\partial F}{\partial x_1} & \dots & \frac{\partial F}{\partial x_n} \end{bmatrix}_{\mathbf{x}}.$$

This row vector is the *matrix form* of the gradient ∇F, since as explained above, $\nabla F = (\frac{\partial F}{\partial x_1}, \dots, \frac{\partial F}{\partial x_n})$.

As we shall see in Proposition 3.15, *the level curves of a function F are orthogonal to the gradient vector field.*

An example of level curves is seen in Fig. 21.

One can think of the matrix of partials for a function $F : \mathbb{R}^n \to \mathbb{R}^m$ as consisting of lined-up column vectors (tangent vectors) or row vectors (gradients) respectively. We have explained this regarding the rows. To understand this for the columns, writing a vector in the domain as $\mathbf{x} = (x_1, \dots, x_n)$ then fixing say $x_2, \dots, x_n$ and setting $t = x_1$ we have a curve $\gamma(t) = F(t, x_2, \dots, x_n)$; note that the first column of the derivative matrix DF is the derivative of this curve, the column tangent vector $D\gamma$.

We have described how the derivative at a point defines a matrix of partial derivatives. The converse is:

Lemma 3.10. *A differentiable map $F : V \to W$ is continuous. For the case $F : \mathbb{R}^n \to \mathbb{R}^m$, the map is differentiable with a continuous derivative iff the partial derivatives exist and are continuous.*

For proof we refer to e.g. §11.2 of [Gui02].

Another basic theorem regarding derivatives is the relation to the matrix of partials:

Theorem 3.11. *If for $F : \mathbb{R}^n \to \mathbb{R}^m$, all the partial derivatives $\partial F_i/\partial x_j$ exists and is continuous at $\mathbf{p}$, then F is differentiable at $\mathbf{p}$, and its derivative is the linear map given by the matrix of partials.*

For a proof see Theorem 6.4 of Marsden's book [Mar74]. The of derivatives is very clearly carried out on pp. 158-185 of Marsden.

3.10. **Best affine approximation: tangent line and plane.** Given a map $F : \mathbb{R}^n \to \mathbb{R}^m$, the terminology "$k^{\text{th}}$-order approximation" to F at a point $\mathbf{x} \in \mathbb{R}^n$ comes from the Taylor polynomials and Taylor series. For a function $f : \mathbb{R}^n \to \mathbb{R}^m$, the best k^{th}-order approximation at $\mathbf{x} \in \mathbb{R}^n$ is the polynomial of degree k which best fits the map near that point. This is the polynomial (in k variables!) which has all the same partial derivatives at that point, up to order k. See §3.24.

Thus the best 0^{th}-order approximationof F at $\mathbf{p} \in \mathbb{R}^n$ is the constant map with the value at that point: the map $\mathbf{x} \mapsto F(\mathbf{p})$. To get the best first-order approximation we add on the linear map given by the derivative matrix $DF|_{\mathbf{p}}$.

This is the affine map

$$\mathbf{x} \mapsto F(\mathbf{p}) + DF|_{\mathbf{p}}(\mathbf{x} - \mathbf{p}),$$

which we mentioned above for the case of $\mathbb{R}^2$, where this gives the equation of the tangent plane to the graph of F.

Example 8. **(Derivative of a linear or affine map)** What is the best linear approximation to a linear map? Answer: it should be the linear map itself!

Let us understand this precisely.

Consider the matrix

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

Acting on column vectors, this defines the linear transformation

$$\begin{bmatrix} x \\ y \end{bmatrix} \mapsto A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax + by \\ cx + dy \end{bmatrix}$$

or written as vectors,

$$T(x, y) = (ax + by, cx + dy).$$

We want to compute the matrix of partials $DT_{\mathbf{p}}$ at a point $\mathbf{p} = (x_0, y_0)$. Now the components of T are $T = (T_1, T_2)$ where

$$T_1(x, y) = ax + by$$

$$T_2(x, y) = cx + dy.$$

Then

$$\nabla T_1 = (a, b)$$
$$\nabla T_2 = (c, d)$$

for each point $\mathbf{p}$. Thus at each point $\mathbf{p}$,

$$DT_{\mathbf{p}} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

This shows:

Proposition 3.12. *For an affine map $F : \mathbb{R}^n \to \mathbb{R}^m$, defined by $F(\mathbf{v}) = \mathbf{v}_0 + T(\mathbf{v})$ where $T : \mathbb{R}^n \to \mathbb{R}^m$ is linear, the derivative is*

$$DF_{\mathbf{p}} = T.$$

That is to say, the derivative is constant*, is constantly equal to the linear part of the map.*

Remark 3.6. To really understand this, consider the case of $\mathbb{R}_\theta$, the rotation counterclockwise of the plane by angle θ.

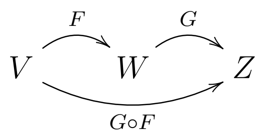
3.11. **The general Chain Rule.** The main theorem involving derivatives is the:

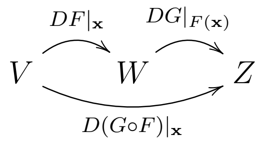
Proposition 3.13. (Chain Rule) *A composition of differentiable maps is differentiable, and the derivative is the composition of the corresponding linear maps.*

That is, for $F : V \to W$ and $G : W \to Z$ then for $G \circ F : V \to Z$ we have:

$$D(G \circ F)|_{\mathbf{p}} = DG|_{f(\mathbf{p})} \circ DF_{\mathbf{p}}.$$

Thus for the finite-dimensional case the chain rule is stated using the product of matrices.





The first example is $\gamma : \mathbb{R} \to \mathbb{R}^3$ and $F : \mathbb{R}^3 \to \mathbb{R}$, where have seen the Chain Rule above; in matrix notation it is:

$$D(F \circ \gamma(t)) = \begin{bmatrix} F_x \, F_y \, F_z \end{bmatrix} |_{\gamma(t)} \begin{bmatrix} x_1'(t) \\ x_2'(t) \\ x_3'(t)) \end{bmatrix}$$

The product gives a (1×1) matrix, whose entry is a number.

In vector notation the Chain Rule is:

$$(F\circ\gamma)'(t) = \nabla F_{\gamma(t)}\cdot\gamma'(t).$$

This number is the same as the entry of the (1×1) matrix above.

Now we can give a second proof of Proposition 3.4 above, which we repeat here:

Proposition 3.14. *Let γ be a differentiable curve in $\mathbb{R}^n$ such that $||\gamma|| = c$ for some constant c. Then $\gamma\perp\gamma'$.*

Proof. (Second Proof, using gradient) We define a function $F:\mathbb{R}^n\to\mathbb{R}$ by $F(\mathbf{x}) = ||\mathbf{x}||^2 = \mathbf{x}\cdot\mathbf{x} = \sum_{i=1}^n x_i^2$. Then since $||\gamma|| = c$ is constant, $c^2 = ||F\circ\gamma||$ whence by the Chain Rule,

$$0 = (F\circ\gamma)'(t) = (\nabla F(\gamma(t))\cdot\gamma'(t)$$

but $F(\mathbf{x}) = F(x_1,\dots,x_n) = x_1^2+\dots+x_n^2$ whence $\nabla F(\mathbf{x}) = 2(x_1,\dots,x_n) = 2\mathbf{x}$. Thus $0 = 2\gamma(t)\cdot\gamma'(t)$, as claimed.

□

Directional derivative and the gradient.

The gradient gives us a simple way of calculating the directional derivative. Given $F:\mathbb{R}^n\to\mathbb{R}$, with gradient vector field ∇F, and given a unit vector $\mathbf{u}$, then the directional derivative of F in direction $\mathbf{u}$ is given simply by the inner product:

$$D_{\mathbf{u}}(F)|_{\mathbf{p}} = (\nabla F(\mathbf{p}))\cdot\mathbf{u}.$$

Exercise 3.10. Check this on the standard basis vectors and compare to the partial derivatives! What is the direction of steepest increase of F at a point $\mathbf{p}$? Of steepest decline? What is the rate of increase of F if in a direction tangent to a level curve?

3.12. **Level curves and parametrized curves.** There are two very distinct types of curves we encounter here: the curves of this section, which are *parametrized curves* (with *parameter* t = time), and the *level curves* of a function. Next we describe a link between the two:

Proposition 3.15. *Let $G:\mathbb{R}^2\to\mathbb{R}$ be differentiable and suppose $\gamma:[a,b]\to\mathbb{R}^2$ is a curve which stays in a level curve of G of level c. Then $\gamma'(t)$ is perpendicular to the gradient of G.*

Proof. We have that $G(\gamma(t)) = c$ for all t. Hence $G(\gamma(t))' = 0$ for all t. Then by the chain rule, this equals $0 = D(G\circ\gamma)(t) = DG|_{\gamma(t)}D\gamma|_t$. The derivatives here are matrices, with DG a (1×2) matrix (a row vector) and $D\gamma$ a column vector; in vector notation, these are the gradient and tangent vector, so this gives $0 = (G\circ\gamma)'(t) = (\nabla G)(\gamma(t))\cdot\gamma'(t)$, so $\nabla G|_{\gamma(t)}\cdot\gamma'(t) = 0$, telling us that the gradient is perpendicular to the tangent vector of the curve, as claimed. □

Example 9. (Dual hyperbolas) See Fig. 21, depicting level curves of the functions $F(x,y) = (x^2-y^2)$ and $G(x,y) = 2xy$.

Exercise 3.11. Plot the level curves of F for levels $0,1,-1$ and for G of levels $0,2,-2$. Compute the gradient vector fields and find their matrices (they are linear!) Compare to the earlier examples of linear vector fields.

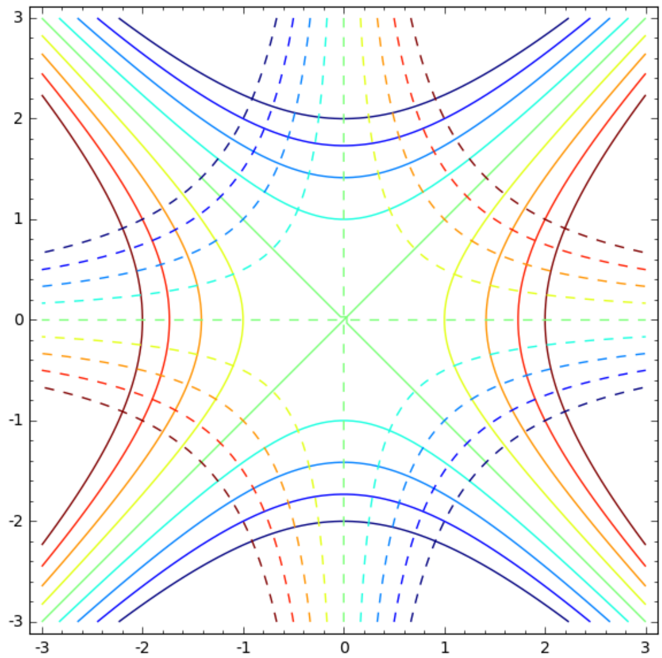


FIGURE 9. Dual families of hyperbolas: Level curves for the functions $F(x,y) = (x^2 - y^2)$ and $G(x,y) = 2xy$. Note that in this special example the level curves of F are orthogonal to the level curves of G. In fact, the gradient vector field of F is orthogonal to the level curves of F, and is tangent to the level curves of G, and vice-versa!

These functions are related algebraically by a change of variables, $u = \frac{1}{\sqrt{2}}(x-y)$, $v = \frac{1}{\sqrt{2}}(x+y)$ and geometrically by a rotation $R_{\pi/4}$. To verify this we define the function $H : \mathbb{R}^2 \to \mathbb{R}^2$ by $H(x,y) = (u,v) = \frac{\sqrt{2}}{2}(x-y, x+y)$ then

$$G \circ H(x,y) = 2 \cdot \frac{1}{2}(x-y)(x+y) = x^2 - y^2 = F(x,y)$$

so $F = G \circ H$.

Now H is a linear transformation of $\mathbb{R}^2$ given by

$$\frac{\sqrt{2}}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} u \\ v \end{bmatrix}$$

and so by the matrix

$$\frac{\sqrt{2}}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}$$

which is indeed rotation counterclockwise by $\pi/4$.

We next check Proposition 3.15 for this example.

The gradient of F is $\nabla F = (\frac{\partial F}{\partial x}, \frac{\partial F}{\partial y}) = (2x, -2y)$ and of G is $\nabla G = (\frac{\partial G}{\partial x}, \frac{\partial G}{\partial y}) = (y, x)$. Note that $(2x, -2y) \cdot (y, x) = 0$ so these vector fields are orthogonal. Furthermore we can find tangent vectors to the level curves as follows. Let us parametrize the level curve $F(x,y) = x^2 - y^2 = c$ by the variable x. Then $y = y(x)$, so the curve is $\gamma(x) = (x, y(x)$ with tangent vector $(1, y'(x))$taking the derivative of the equation with respect to x gives $2x - 2yy' = 0$, so $y' = x/y$. Thus $\gamma'(x) = (1, x/y)$. This is proportional to the vector $(y, x) = \nabla G$ which as we have already noted is orthogonal to the gradient of F at that point.

3.13. **Level surfaces, the gradient and the tangent plane.** In Proposition 3.15 of §3.12 we showed that the gradient vector field of $F : \mathbb{R}^2 \to \mathbb{R}$ is orthogonal to the level curves of F. In fact something similar is true for any dimension. For the case of $\mathbb{R}^3$ we get a new formula for the tangent plane, as we now explain.

Proposition 3.16. *Let $G : \mathbb{R}^3 \to \mathbb{R}$ be differentiable and suppose $\gamma : [a,b] \to \mathbb{R}^3$ is a curve such that the image of γ remains inside the level surface of level c, $\{(x,y,z) : G(x,y,z) = c\}$. That is, for all t, $G(\gamma(t)) = c$. Then $\gamma'(t)$ is perpendicular to the gradient of G.*

More generally this is true for higher dimensions, $G : \mathbb{R}^n \to \mathbb{R}$.

Proof. We have that $G(\gamma(t)) = c$ for all t. Hence $G(\gamma(t))' = 0$ for all t. Now by the chain rule, $0 = D(G \circ \gamma)(t) = DG(\gamma(t)D\gamma(t)$. DG is now a $(1 \times n)$ matrix and $D\gamma$ a $(n \times 1)$ column vector; in vector notation, these are the gradient and tangent vector, so this gives $0 = \frac{\mathrm{d}}{\mathrm{d}t}c = (G \circ \gamma)'(t) = (\nabla G)(\gamma(t)) \cdot \gamma'(t) = 0$. □

Exercise 3.12. First we have a review problem from Linear Algebra: Recall that the general equation for a plane in $\mathbb{R}^3$ is:

$$Ax + By + Cz + D = 0$$

where not all three of A, B, C are 0. Given a point $\mathbf{p} = (x_0, y_0, z_0)$ and a vector $\mathbf{n} = (A, B, C)$ then find the general equation of the plane through $\mathbf{p}$ and perpendicular to $\mathbf{n}$.

Solution: We know that the plane is the collection of all $\mathbf{x} = (x, y, z)$ such that

$$\mathbf{n} \cdot (\mathbf{x} - \mathbf{p}) = 0,$$

so for $\mathbf{n} = (A, B, C)$ and $\mathbf{x} = (x, y, z)$ and $\mathbf{p} = (x_0, y_0, z_0)$ then

$$(A, B, C) \cdot (x - x_0, y - y_0, z - z_0) = 0$$

giving the general equation for the plane,

$$Ax + By + Cz + D = 0$$

where $D = -\mathbf{n} \cdot \mathbf{p} = -(Ax_0 + By_0 + Cz_0)$.

See also Exercise 3.5.

Exercise 3.13. Given the function $F(x, y, z) = x^2 + y^2 + z^2$, find the tangent plane to this sphere at the point $(1, 2, 3)$.

Solution: Note that $F(1, 2, 3) = 14$. Therefore this point is on the level surface of F of level 14. (This is the sphere about the origin of radius $\sqrt{14}$).

Now the gradient of F is $\nabla F(x, y, z) = (2x, 2y, 2z)$. We know the gradient is orthogonal to the sphere hence to the tangent plane. This *normal vector* (to both) is $\mathbf{n} = \nabla F(1,2,3) = (2,4,6)$. We are in the situation of the previous exercise: the equation of the plane is

$$Ax + By + Cz + D = 0$$

where the normal vector is $\mathbf{n} = (A, B, C) = (2, 4, 6)$ and the plane passes through the point $\mathbf{p} = (1, 2, 3)$.

The equation of the plane is therefore

$$\mathbf{n} \cdot (\mathbf{x} - \mathbf{p}) = 0$$

or

$$\mathbf{n} \cdot ((x, y, z) - \mathbf{p}) = 0$$

so

$$(A, B, C) \cdot (x - 1, y - 2, z - 3) = 0$$

giving

$$2x + 4y + 6z + D = 0$$

where

$$D = -\mathbf{n} \cdot \mathbf{p} = -(2, 4, 6) \cdot (1, 2, 3) = -(2 + 8 + 18) = -28,$$

so we have the plane with general equation

$$2x + 4y + 6z + -28 = 0.$$

Exercise 3.14. We solve an exercise requested by a student.

Guidorizzi Vol 2 # 2 of §11.3, p. 204. [Gui02],

Find the equation of a plane which passes through the ponts $(1, 1, 2)$ and $(-1, 1, 1)$and which is tangent to the graph of the function $f(x, y) = xy$.

Solution. We use normal vectors, as follows. The graph of f is

$$\{(x, y, z) : z = f(x, y)\}$$

which equals

$$\{(x, y, z) : z = xy\}$$

equivalently written

$$\{(x, y, z) : xy - z = 0\}.$$

This is the level surface of $F : \mathbb{R}^3 \to \mathbb{R}$ defined by $F(x, y, z) = xy - z$. This has gradient vector $\nabla F = (y, x, -1)$. Let $\mathbf{p} = (x_0, y_0, z_0)$ denote the point where the plane meets the graph. Then at the point $\mathbf{p}$ we have $\nabla F_{\mathbf{p}} = (y_0, x_0, -1)$. We know that the gradient is orthogonal to the level surfaces, in other words it is orthogonal to the tangent plane to the surface at that point. So $\mathbf{n} = \nabla F_{\mathbf{p}}$ is a normal vector to the tangent plane of the level surface at $\mathbf{p}$. This gives us the equation for the tangent plane

$$\mathbf{n} \cdot (\mathbf{x} - \mathbf{p}) = 0$$

so

$$(y_0, x_0, -1) \cdot ((x, y, z) - (x_0, y_0, z_0)) = 0$$

$$(y_0, x_0, -1) \cdot (x - x_0, y - y_0, z - z_0) = 0$$

so

$$y_0x - x_0y - z + z_0 = 0$$

Now $z_0 = x_0y_0$ since $\mathbf{p}$ is also on the graph of the function. This gives

$$y_0x - x_0y - z + x_0y_0 = 0$$

We need to find x_0, y_0. The two points are on this plane so satisfy the equation.

Substituting $(x, y, z) = (1, 1, 2)$ and $(-1, 1, 1)$ gives us the equations

$$y_0 - x_0 - 2 + x_0y_0 = 0$$

$$-y_0 - x_0 - 1 + x_0y_0 = 0$$

Subtracting,

$$2y_0 - 1 = 0$$

$$y_0 = 1/2$$

We now have from the first equation,

$$y_0 - x_0 - 2 + x_0y_0 = 0$$

so

$$1/2 - x_0 - 2 + x_0 1/2 = 0$$

multiplying by 2,

$$1 - 2x_0 - 4 + x_0 = 0$$

$$x_0 = 3$$

Thus $z_0 = 3/2$ giving the equation of the plane:

$$\mathbf{n} \cdot (\mathbf{x} - \mathbf{p}) = 0$$

with $\mathbf{n} = (y_0, x_0, -1) = (1/2, 3, -1)$ and $\mathbf{p} = (x_0, y_0, z_0) = (3, 1/2, 3/2)$. Finally in the form

$$Ax + By + Cz + D = 0$$

we have

$$1/2x + 3y - z - 3/2 = 0$$

or equivalently

$$x + 6y - 2z - 3 = 0.$$

To check our numbers we can verify that the three points are indeed on this plane.

Remark 3.7. In these notes we have emphasized the role of three distinct ways of presenting, or of viewing, the same object: for example a curve may be the graph of a function, a level curve, or a parametrized curve. We wish to indicate how this fits into a larger context, in other parts of mathematics.

First, here is a solution to part of Exercise 3.6: to write the image and kernel in matrix form.

Consider

$$\begin{bmatrix} 2 & 3 \\ 1 & 2 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} = s \begin{bmatrix} 2 \\ 1 \\ 4 \end{bmatrix} + t \begin{bmatrix} 3 \\ 2 \\ 5 \end{bmatrix} \tag{4}$$

Thus if we write the columns of a (3×2) matrix as $\mathbf{v} = (v_1, v_2, v_3)$, $\mathbf{w} = (w_1, w_2, w_3)$ we have more generally

$$\begin{bmatrix} v_1 & w_1 \\ v_2 & w_2 \\ v_3 & w_3 \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} = s \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} + t \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \tag{5}$$

defining the map $T : \mathbb{R}^2 \to \mathbb{R}^3$ where $T(s,t) = s\mathbf{v} + t\mathbf{w}$. This is a parametrized plane, which in this case passes through $\mathbf{0}$.

Given a parametrized plane in $\mathbb{R}^3$, we should be able to find the general equation. To do this we bring in the *vector product*, which we next explain. But first, a few words about the *determinant*!

3.14. **Elementary row and column operations.** A good reference on matrix operations in Linear Algebra is Strang's text [Str12].

For the next section we recall:

Definition 3.11. Given an $(n \times m)$ matrix M, a *basic elementary row operation* has two types:
(1) We exchange two rows;
(2) We replace a chosen row by itself added to a multiple of a different row.

In many texts, a third operation is permitted:
(3) We multiply a row by a constant.

But we don't need this and it will be more convenient for us to not include this one.

An *elementary row operation* is the result of applying the basic operations 1 and 2 finitely many times.

Elementary column operations are defined similarly.

A *basic elementary matrix* is a $(d\times d)$ matrix formed by carrying out an elementary row operation on the identity matrix I.

Lemma 3.17.
(i) A basic elementary matrix can also be formed by carrying out an elementary column *operation on I.*
(ii) Given an $(n \times m)$ matrix M, let $\widetilde{M}$ be the matrix which results after carrying out an elementary row operation. Then $\widetilde{M} = EM$ where E is the $(n \times n)$ elementary matrix formed by carrying out the same elementary row operation on I.

The same is true for column operations except now E is $(m \times m)$ and $\widetilde{M} = ME$.

Proof. First we consider the basic elementary matrices. (i) If we form E by exchanging row i and row k, then E has a 1 in place i of row k and place k of row i. Thus $E_{ki} = 1$ and $E_{ik} = 1$. If we exchange column i and column k then E has a 1 in place i of column k and place k of column i. Thus $E_{ik} = 1$ and $E_{ki} = 1$, the same thing.

If we form E by replacing row i with a times row k, then E is the identity matrix except for $E_{ik} = a$. If we form E by replacing column k with a times row i, then E is the identity except for $E_{ik} = a$, the same thing!

(ii) For $\mathbf{e}_k^t$ the standard basis row vector, so it is $(1 \times n)$ with $(\mathbf{e}_k)_k = 1$, then $\mathbf{e}_k^t M$ gives the $(1 \times m)$ row vector which is the k^{th} row of M. It follows from this that if an

$(n \times n)$ matrix N has $\mathbf{e}_k^t$ as its j^{th} row, then row j of NM is the k^{th} row of M. This implies the claim, but it is much better to see this by trying out some examples; see (8), (6), (7) below.

The case of operation (2) is similar.

General elementary matrices: by what we have just shown, a general elementary matrix E is a product $E = E_l \dots E_2 E_1$, proving (i). Part (ii) follows.

For column operations, we take the transpose, using the fact that $(EM)^t = M^t E^t$. □

Pictorially,

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \longrightarrow & \mathbf{u} & \longrightarrow \\ \longrightarrow & \mathbf{v} & \longrightarrow \\ \longrightarrow & \mathbf{w} & \longrightarrow \end{bmatrix} = \begin{bmatrix} \longrightarrow & \mathbf{v} & \longrightarrow \\ \longrightarrow & \mathbf{u} & \longrightarrow \\ \longrightarrow & \mathbf{w} & \longrightarrow \end{bmatrix} \tag{6}$$

and

$$\begin{bmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \longrightarrow & \mathbf{u} & \longrightarrow \\ \longrightarrow & \mathbf{v} & \longrightarrow \\ \longrightarrow & \mathbf{w} & \longrightarrow \end{bmatrix} = \begin{bmatrix} \longrightarrow & \mathbf{u} & \longrightarrow \\ \longrightarrow & a\mathbf{u} + \mathbf{v} & \longrightarrow \\ \longrightarrow & \mathbf{w} & \longrightarrow \end{bmatrix} \tag{7}$$

3.15. **Geometrical meaning of elementary matrices: reflection and sliding.** For this, we first consider elementary row operations. Let us consider the (2×2) basic elementary matrix $E = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. In its action on column vectors it interchanges $\mathbf{i}$ and $\mathbf{j}$. We have

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} y \\ y \\ x \end{bmatrix}.$$

Geometrically, this is a reflection in the line $y = x$.

Next consider (3×3) matrix $E = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Now we have

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} y \\ x \\ z \end{bmatrix}$$

which is reflection in the plane $y = x$.

So these elementary matrices of the first type give reflections. Next, the matrix given by the second type of basic elementary operation gives what we shall call a *sliding transformation*; a more common name is a *skewing transformation.* We explain why we call this "sliding": For an example note that

$$\begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x + ty \\ y \end{bmatrix}$$

which moves the unit square to a parallelogram, sliding the top horizontally by distance t.

From (7) we see that left multiplication by $E = \begin{bmatrix} 1 & 0 & 0 \\ -a & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ transforms the parallelopiped $P(\mathbf{i}, a\mathbf{i}+\mathbf{j}, \mathbf{k})$ to $P(\mathbf{i}, \mathbf{j}, \mathbf{k})$ which slides the top of it over so it is the unit cube.

By applying successively these operations we can turn any parallelopiped in $\mathbb{R}^3$ into a rectangular solid. In algebraic terms, this says (equivalently) that we can apply the elementary row operations on a (3×3) invertible matrix M and end up with a diagonal matrix D. We note that by the way we have defined elementary matrices, they all have determinant 1, explained in the next section.

3.16. **Two definitions of the determinant.**

Algebraic definition: Let A be an $(n \times n)$ real or complex matrix. We begin with the usual algebraic definition, which is inductive on n. For $n = 1$, $A = [a] = [A_{11}]$ and det$A$ is just the number a. For $n = 2$, $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, and we set $\det(A) = ad - bc$.

This is extended as follows: we define a matrix with entries $S_{ij} \in \{1, -1\}$ as follows: $S_{ij} = (-1)^{i+j}$. To visualize this, we write simply the corresponding signs, in a checkerboard pattern:

$$S = \begin{bmatrix} + & - & + & - \\ - & + & - & + \\ + & - & + & - \\ - & + & - & + \end{bmatrix}$$

The *ij minor* $A(ij)$ of A is defined to be the $(n-1) \times (n-1)$ matrix formed by removing the i^{th} row and j^{th} column of A.

Then we *expand along the top row* by forming the sum of $(\pm 1)\det A(1j)$, where the signs are given by the top row of S, i.e.

$$\det(A) = \sum_{j=1}^{n} (-1)^{1+j} \det A(1j).$$

Similarly we define the expansion along, say, the i^{th} row to be $\sum_{j=1}^{n} (-1)^{i+j} \det A(ij)$ or indeed along any column.

Lemma 3.18. *These are equal, giving the same number whatever row or column chosen!*

Proof. The proof depends on the following two facts:
Fact 1: If $\widetilde{A}$ is the matrix formed from A by exchanging two rows, then $\det\widetilde{A} = (-1)\det A$; and
Fact 2: $\det A^t = \det A$.

Now if we expand along the second row of A, the formula we get is that for expanding along the first row of $\widetilde{A}$, times (-1), which is $-\det A$, and the two sign changes cancel.

If we expand along the first column of A, this is the same formula as expanding along the first row of A^t. This completes the proof.

∎

Note that this algorithm also works for the (2×2) case!

Geometric definition:

Definition 3.12. Let M be an $(n \times n)$ real matrix. Then

$$\det M = (\pm 1)(\text{factor of change of volume})$$

where we take $+1$ if M preserves orientation, -1 if that is reversed. (Here this is *n-dimensional volume* and so is length, area in dimensions $1, 2$).

Theorem 3.19. *The algebraic and geometric definitions are equivalent.*

Proof. For a (2×2) matrix A, note that the factor of change of volume is the area of the image of the unit square, that generated by the standard basis vectors $(1, 0)$ and $(0, 1)$, which equals the area of the parallelogram with sides the matrix columns, (a, c) and (b, d).
Case 1: $c = 0$. Then the matrix is upper triangular and its determinant algebraically is ad. But the parallelogram area is (base)(height)= ad as well.

The formula area(parallelogram)= (base)(height) is usually proved by cutting off a triangle vertically and shifting it to the other side, thus forming a rectangle of the same base and height. Here is a different way to picture this: imagine the parallelogram is a pile of horizontal layers, like a stack of cards, and straighten the pile to a vertical pile by *sliding* the cards, ending up with the same $(a \times d)$ rectangle. See §3.15 where we define sliding transformations.
General Case: We reduce to Case 1 as follows, *not* by rotating (also possible!) but by sliding the far side of the parallelogram along the direction (b, d). A simple computation shows the area is indeed $ad - bc$.
Higher dimensions: We note that the above "sliding" operations can be done algebraically by an operation of column reduction, equivalently, multiplying on the right by an elementary matrix of determinant one. This reduces to the upper diagonal case, and beyond to the diagonal case if desired.

We observe that the same procedure works in $\mathbb{R}^3$ and beyond.

We can visualize that a parallelopiped $P(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ in $\mathbb{R}^3$ can be transformed by sliding parallel to these three vectors into a rectangular solid, with sides parallel to the axes. To prove this we again note that the operation of sliding is given algebraically by an elementary column operation, or equivalently by right multiplication by an elementary matrix (of determinant ± 1). And we know that column-reduction (just like row-reduction except the transpose) of an invertible matrix can always be done to end up with a diagonal matrix.

□

3.17. **Orientation.** We may be accustomed to thinking of a certain basis as having positive orientation and another negative, but this has no intrinsic meaning: what does make sense is to say that two given bases have the *same* or *opposite* orientation. As we shall explain, there are only two choices for this.

A basis $\mathcal{B}_1$ of $\mathbb{R}^n$ is an n-tuple $(\mathbf{v}_1, \dots, \mathbf{v}_n)$ of linearly independent vectors which generate. (Recall that in fact, n L.I. vectors will always generate). From this we form the matrix $M_{\mathcal{B}}$ with those columns. This is an invertible matrix. (Exercise: verify this!)

We let $\widehat{\mathcal{B}}$ denote the collection of all bases of $\mathbb{R}^n$.

The change from one basis $\mathcal{B}_1$ to another $\mathcal{B}_2$ is given by an invertible matrix A. What we mean by this is:

$$M_{\mathcal{B}_1} A = M_{\mathcal{B}_2}.$$

An example is to simply change the order of the basis, defining $\mathcal{B}_2 = (\mathbf{v}_2, \mathbf{v}_1, \mathbf{v}_3, \dots, \mathbf{v}_n)$. Thus to get the matrix $M_{\mathcal{B}_2}$ we have performed on $M_{\mathcal{B}_1}$ the elementary column operation of switching the first two columns. This is given by right multiplication by the elementary matrix E which is I with the first two columns switched: (!!!)

$$\begin{bmatrix} a & e & * & * \\ b & f & * & * \\ c & g & * & * \\ d & h & * & * \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} e & a & * & * \\ f & b & * & * \\ g & c & * & * \\ h & d & * & * \end{bmatrix} \tag{8}$$

where $*$ stands for any number.

Given our invertible matrix A, we know it is a product of elementary matrices E_i times a diagonal matrix D , since by column reduction we have: $AE_1E_2 \dots E_m = D$. Thus the change of basis can always be given in this way.

Now by definition $GL(n, \mathbb{R})$ is the collection of invertible $(n \times n)$ matrices. The collection of those with $\det A > 0$ is called GL^+; these are the *orientation-preserving* matrices. (From the point of view of Group Theory, GL^+ is a subgroup of index 2 of GL, and its coset is GL^-, the collection (not a subgroup!) of *orientation-reversing* matrices). Letting GL act on the bases $\widehat{\mathcal{B}}$, we define two bases $\mathcal{B}_1, \mathcal{B}_2$ to have the *same orientation* iff one is taken to the other by an element of GL^+. Since this subgroup has index 2, there are only these two choices, and the second case is expressed by saying they have *opposite* orientation.

Then, choose one basis $\mathcal{B}_1$ we declare (arbitrarily) that this has *positive* orientation. The image of this by applying all elements of GL^+ defines $\widehat{\mathcal{B}^+}$, the bases with positive orientation, and the complement defines $\widehat{\mathcal{B}^-}$, the bases with *negative orientation.* Note that $\widehat{\mathcal{B}^-}$ is the GL^+-image of any $\mathcal{B}_2$ not in $\widehat{\mathcal{B}^+}$.

Example 10. We give some examples. Suppose $\mathcal{B}_1 = (\mathbf{v}_1, \mathbf{v}_2 \dots, \mathbf{v}_n)$ has a certain orientation (say, positive). Then switching the order of two of them, $(\mathbf{v}_2, \mathbf{v}_1, \mathbf{v}_3, \mathbf{v}_4 \dots, \mathbf{v}_n)$ has the opposite orientation. Also, as just shown above for the (4×4) case, this change is give by multiplication by the matrix A with columns $(\mathbf{e}_2, \mathbf{e}_1, \mathbf{e}_3, \dots, \mathbf{e}_n)$ in other words I with the first two columns interchanged. Note A has determinant -1.

For $\mathbb{R}^3$, this gives: $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ has the opposite orientation from $(\mathbf{j}, \mathbf{i}, \mathbf{k})$.

For a second example, $(-\mathbf{v}_1, \mathbf{v}_2 \dots, \mathbf{v}_n)$ also has opposite orientation from $\mathcal{B}_1$. (Exercise: what is A in this case?)

Geometrically, this is a *reflection* in the subspace generated by the vectors $\{\mathbf{v}_2 \dots, \mathbf{v}_n\}$.

For example, reflection in the xy-plane is the map $R(a,b,c) = (a,b,-c)$ which is given by the matrix $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$ (times column vectors), and that sends the standard basis $(\mathbf{i},\mathbf{j},\mathbf{k})$ to $(\mathbf{i},\mathbf{j},-\mathbf{k})$.

Exercise: write the formula for reflection in a *line*, for example in the $x-$ axis!

How can you find from this the formula for reflection in a general plane, or in a general line? For reflection in the *point* $\mathbf{0}$?

Definition 3.13. The *standard orientation* for $\mathbb{R}^n$ is defined by the choice that $(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n)$ has positve orientation. Thus for $\mathbb{R}^1$ this is $(\mathbf{e}_1)$, for $\mathbb{R}^2$ is $\mathbf{e}_1, \mathbf{e}_2) = (\mathbf{i},\mathbf{j})$ and for $\mathbb{R}^3$ this is $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) = (\mathbf{i},\mathbf{j},\mathbf{k})$.

Theorem 3.20.
(i) $det(AB) = det(A)det(B)$.
(ii) $det(B^{-1}AB) = det(A)$.

Proof. Part (i) can be proved algebraically, but it is much easier to use the geometric definition of determinant, that $\det(A) = (\pm 1)\cdot$ (factor of change of volume). (Since $(AB)\mathbf{v} = A(B\mathbf{v})$- this is a multiplication of matrices, and we have the associative law- multiplication of the volume by b and then by a changes it by the factor ab).

Now we have the factor of 1 if A preserves orientation, -1 if not. This again works for the product; changing the orientation twice leaves it fixed, and $(-1)(-1) = 1$.

Part (ii) follows from this. □

3.18. **Three definitions of the vector product.** The vector product $\mathbf{v}\wedge\mathbf{w}$ is defined only on $\mathbb{R}^3$, and gives a vector in $\mathbb{R}^3$. (In $\mathbb{R}^2$, we make the special definition $(a,b)\wedge(c,d) = \mathbf{v}\wedge\mathbf{w} \in \mathbb{R}^3$ where $\mathbf{v} = (a,b,0)$ and $\mathbf{w} = (c,d,0)$). Here we present three equivalent definitions of the vector product. We write $\mathbf{i},\mathbf{j},\mathbf{k}$ for the standard basis vectors in $\mathbb{R}^3$. We write $P(\mathbf{v},\mathbf{w},\mathbf{z})$ for the paralelopiped spanned by $\mathbf{v},\mathbf{w},\mathbf{z} \in \mathbb{R}^3$, that is, $P(\mathbf{v},\mathbf{w},\mathbf{z}) \equiv \{a\mathbf{v} + b\mathbf{w} + c\mathbf{z} : a,b,c \in [0,1]\}$, and $P(\mathbf{v},\mathbf{w})$ for the parallelogram spanned by $\mathbf{v},\mathbf{w} \in \mathbb{R}^3$, so $P(\mathbf{v},\mathbf{w}) \equiv \{a\mathbf{v} + b\mathbf{w} : a,b \in [0,1]\}$.

Theorem 3.21. *The following definitions are equivalent.*
(1) (Via the "determinant" formula)*:*

$$\mathbf{v}\wedge\mathbf{w} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} = \mathbf{i}\begin{vmatrix} v_2 & v_3 \\ w_2 & w_3 \end{vmatrix} - \mathbf{j}\begin{vmatrix} v_1 & v_3 \\ w_1 & w_3 \end{vmatrix} + \mathbf{k}\begin{vmatrix} v_1 & v_2 \\ w_1 & w_2 \end{vmatrix}$$

(2) (The geometric definition)*:*

$\mathbf{v}\wedge\mathbf{w}$ *satisfies the following properties:*
(i) $\mathbf{z} = \mathbf{v}\wedge\mathbf{w}$ *is perpendicular to* $\mathbf{v}$ *and to* $\mathbf{w}$*;*
(ii) The norm of $\mathbf{z}$ *is equal to the area of the parallelogram* $P(\mathbf{v},\mathbf{w})$*; thus*

$$||\mathbf{v}\wedge\mathbf{w}|| = ||\mathbf{v}|| ||\mathbf{w}|| \cdot |\sin(\theta)|.$$

(iii) If $\mathbf{z} \neq \mathbf{0}$*, then* $(\mathbf{v},\mathbf{w},\mathbf{z})$ *forms a positively oriented basis for* $\mathbb{R}^3$.

(3) (The algebraic definition) *: The vector product is a bilinear operation such that* $\mathbf{i} \wedge \mathbf{j} = \mathbf{k}$*,* $\mathbf{j} \wedge \mathbf{k} = \mathbf{i}$*,* $\mathbf{k} \wedge \mathbf{i} = \mathbf{j}$*.*

Remark 3.8. (1) is the usual definition given in texts.

Regarding (2), we remark that θ is the angle *from* $\mathbf{v}$ *to* $\mathbf{w}$, where in the plane this would mean measured in the counterclockwise sense from $\mathbf{v}$ to $\mathbf{w}$; in $\mathbb{R}^3$, together with an orientation, "counterclockwise" is defined by looking down along the thumb for the right-hand rule. Note that since the modulus is taken, this is the same for the angle $-\theta$ from $\mathbf{w}$ to $\mathbf{v}$ and in any case is positive as a norm should be.

The formula in (3) is easy to remember as it follows a circle from $\mathbf{i}$ to $\mathbf{j}$ to $\mathbf{k}$.

Proof. To prove that (1) $\implies$ (2) we note that for any vector $\mathbf{u}$, we used the *mixed product*, a mixture of the inner and vector products $\mathbf{u} \cdot (\mathbf{v} \wedge \mathbf{w})$, and note that:

$$\mathbf{u} \cdot (\mathbf{v} \wedge \mathbf{w}) = \begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} \tag{9}$$

Taking $\mathbf{u} = \mathbf{v}$ in (9) it follows that $\mathbf{v} \cdot \mathbf{z} = 0$, similarly for $\mathbf{w}$, proving (i). Recall that $\begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} = \pm$ (volume of the paralellopiped spanned by $\mathbf{u}, \mathbf{v}, \mathbf{w}$), using the fact that $\det M = \det M^t$, where the sign is $+$ iff the map preserves orientation, since the parallelopiped is the image of the unit cube, and since from Theorem 3.21 we know the determinant gives $\pm$ (factor of change of volume).

Now taking in (9) $\mathbf{u} = \mathbf{z} = \mathbf{v} \wedge \mathbf{w}$, then $||\mathbf{z}||^2 = \mathbf{z} \cdot \mathbf{z} = \begin{vmatrix} z_1 & z_2 & z_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} \geq 0$ so the orientation of $(\mathbf{z}, \mathbf{v}, \mathbf{w})$ is positive. Using this, from the geometric definition of the determinant,

$$\begin{vmatrix} z_1 & z_2 & z_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} = \mathrm{vol}(\mathbf{z}, \mathbf{v}, \mathbf{w})$$

where this means the volume of the parallelopiped spanned by the basis (if linearly independent) $(\mathbf{z}, \mathbf{v}, \mathbf{w})$. Here we use the fact that we can exchange rows for columns as $\det A = \det A^t$. But since $\mathbf{z}$ is orthogonal to the base parallelogram, this volume is (base area)(height).

This gives

$$||\mathbf{z}||^2 = (\text{base area})(\text{height}) = (\text{base area})||\mathbf{z}||$$

so $||\mathbf{z}|| =$ (base area) as claimed. This concludes the proof that Def. (1) implies Def. 2.

It is clear that both Defs. (1), (2) imply Def. (3), but knowing Def. (3) for the basis vectors $\mathbf{i}, \mathbf{j}, \mathbf{k}$ determines $\mathbf{v} \wedge \mathbf{w}$ for all $\mathbf{v}, \mathbf{w}$, by bilinearity. Hence all three are equivalent.

□

Corollary 3.22. *We have the nice (and useful!) formula*

$$||\mathbf{v} \wedge \mathbf{w}||^2 = (\mathbf{v} \cdot \mathbf{v})(\mathbf{w} \cdot \mathbf{w}) - (\mathbf{v} \cdot \mathbf{w})^2.$$

Proof. From Theorem 3.21 we know that

$$||\mathbf{v} \wedge \mathbf{w}||^2 = (\text{area})^2 = (||\mathbf{v}||||\mathbf{w}||| \sin \theta|)^2$$

and this is

$$||\mathbf{v}||^2||\mathbf{w}||^2(1 - \cos^2 \theta) = ||\mathbf{v}||^2||\mathbf{w}||^2 - (||\mathbf{v}||||\mathbf{w}|| \cos \theta)^2 = ||\mathbf{v}||^2||\mathbf{w}||^2 - (\mathbf{v} \cdot \mathbf{w})^2.$$

□

We shall next see how the vector product satisfies three important properties, the first two of which we have already proved:

Definition 3.14. A *Lie bracket* $[x, y]$ on a vector space V is an operation on V (a function from $V \times V$ to V) which satisfies the axioms:
– bilinearity;
–anticommutativity: $[y, x] = -[x, y]$;
–the *Jacobi identity*

$$[x, [y, z]] + [y, [z, x]] + [z, [x, y]] = \mathbf{0}$$

Proposition 3.23. *The vector product* $\mathbf{v} \wedge \mathbf{w}$ *on* $\mathbb{R}^3$ *is a Lie bracket, setting* $[\mathbf{v}, \mathbf{w}] = \mathbf{v} \wedge \mathbf{w}$.

Proof. We have shown the first two properties.

Now from (3) we have an exceptionally easy proof of the Jacobi identity, since by bilinearity it is enough to check this on the basis vectors, and for example

$$[\mathbf{i}, [\mathbf{j}, \mathbf{k}]] + [\mathbf{j}, [\mathbf{k}, \mathbf{i}]] + [\mathbf{k}, [\mathbf{i}, \mathbf{j}]] = \mathbf{0}$$

since each term is $\mathbf{0}$, and similarly for the other cases. □

General equation of a plane; matrix form. Going back to the parametric equation of a plane in (33), we had the linear transformation $T : \mathbb{R}^2 \to \mathbb{R}^3$ where $T(s, t) = s\mathbf{v} + t\mathbf{w}$. Writing H for the (3×2) matrix and $\mathbf{z}$ for the (2×1) column vector $\mathbf{z} = \begin{bmatrix} s \\ t \end{bmatrix}$, then in matrix form this is

$$H\mathbf{z} = \begin{bmatrix} v_1 & w_1 \\ v_2 & w_2 \\ v_3 & w_3 \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} = s \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} + t \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \tag{10}$$

The image of the map H is a plane which passes through $\mathbf{0}$. The plane parallel to this which passes through some point $\mathbf{p}$ is the image of the function $H_{\mathbf{p}} : (s, t) \mapsto T(s, t) = s\mathbf{v} + t\mathbf{w} + \mathbf{p}$. Note that H is a linear transformation, while $H_{\mathbf{p}}$ is affine but not linear (unless $\mathbf{p}$ happens to lie on the plane $\text{Im}(H)$).

Given this parametric equation we can find the general equation of the plane $\text{Im}(H_{\mathbf{p}})$ as follows: we take our normal vector to be $\mathbf{n} = (A, B, C)$ where $\mathbf{n} = \mathbf{v} \wedge \mathbf{w}$. Then points (x, y, z) in the plane $T(s, t) = s\mathbf{v} + t\mathbf{w} + \mathbf{p} = (x, y, z)$ satisfy the equation

$$(A, B, C) \cdot ((x, y, z) - \mathbf{p}) = 0$$

so giving

$$Ax + By + Cz + D = 0$$

where $D = -\mathbf{n} \cdot \mathbf{p}$.

We have explained this above, in Exercise 3.12.

In matrix form this is

$$M\mathbf{v} = \begin{bmatrix} A & B & C \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -D \end{bmatrix}. \tag{11}$$

Defining $S : \mathbb{R}^3 \to R$ to be the function $S(x, y, z) = Ax + By + Cz$ then the plane is the level surface of level $-D$ of S.

Putting these two maps together, we have the composition of maps, with the two matrices acting on column vectors:

$$\mathcal{M}_{2,1} \xrightarrow{H} \mathcal{M}_{3,1} \xrightarrow{M} \mathcal{M}_{1,1}$$

or as linear transformations:

$$\mathbb{R}^2 \xrightarrow{T} \mathbb{R}^3 \xrightarrow{S} \mathbb{R}^1 \tag{12}$$

Restricting to the image of T, the geometrical plane $P \subseteq \mathbb{R}^3$, we have:

$$\mathbb{R}^2 \xrightarrow{T} P \xrightarrow{S} \{-D\} \tag{13}$$

The plane P is a set of points (x, y, z), which on the one hand is the image of the map T, and on the other is a translate of the kernel of the map S by the vector $\mathbf{p}$.

Level surfaces of different levels (that is, planes which are parallel, with different constants D) fit together as described by Equation (5.10).

Remark 3.9. **The important point in this is the following:** The plane P is, by itself, simply a subset of points, a two-dimensional subspace of $\mathbb{R}^3$. However, Equation (13) gives us two very different ways of viewing P: via the map T or the map S.

Summarizing, P is the *image* of $\mathbb{R}^2$ via the map T. That is, the map T *parametrizes* P; thus via this map P becomes the parametrized plane $L_{\mathbf{p}}(s, t) = s\mathbf{v} + t\mathbf{w} + \mathbf{p}$.

On the other hand, via the map S, P is the *preimage* (inverse image) of a constant value $-D$. Thus it is is seen to be a *level surface* of the map (of level $-D$). Thus it is only one of a family of parallel planes, of different levels.

This also gives us insight as to the meaning of the diagram: it says something about the object (in this case the plane P) in the middle, from two different perspectives, given by the two maps.

Again, this just reflects the difference between our two ways of understanding a plane, as a parametrized plane, see Equation (33), or as the solution set of its general equation. And this latter is, geometrically, a plane which passes through a point and has a certain normal vector, $\mathbf{n} = (A, B, C)$.

This is the simplest case, of a line in the plane or a plane in space. The general situation comes from these fundamental results of Linear Algebra:

Theorem 3.24. *Given finite-dimensional vector spaces V, W let $T : V \to W$ be a linear transformation. Then:*
i) the null space $\mathcal{N}(T)$ is a vector subspace of V;
ii) the image $Im(T)$ is also; and
(iii) $dim(\mathcal{N}(T)) + dim(Im(T)) = dim(V)$.

Exercise 3.15. Prove (i), (ii)! See Exercise 3.6.

Corollary 3.25. *If T above is surjective, then $dim(\mathcal{N}(T)) = dim(V) - dim(W)$.*

Before we describe the proof, we write it as a diagram, of linear transformations on vector spaces:

$$K \xrightarrow{I} V \xrightarrow{T} W$$

Here the first map I is an *injection* $I(\mathbf{v}) = \mathbf{v}$, which just means that it is a $1-1$ function (just the identity map in this case). Its image is the subspace $\text{Im}(I) = K$ which is the kernel of T, and the image of T is W. That is, the map T is onto.

For the previous example, the map I represents the plane K as a parametrized subspace, while the map T gives its general equation.

In Algebra, a diagram of maps where the image of one map is the kernel of the following map is called an *exact sequence*. In fact, the above diagram of vector spaces extends to

$$\{\mathbf{0}\} \xrightarrow{I} K \xrightarrow{T} V \xrightarrow{S} W \xrightarrow{\pi} \{\mathbf{0}\}$$

where I is the *injection* and π is the projection $\pi(\mathbf{v}) = \mathbf{0}$. This extended diagram is also exact: exactness of the first part

$$\{\mathbf{0}\} \xrightarrow{I} K \xrightarrow{T} V$$

says that I is injective ($1-1$ to its image) since the kernel of T is then $\{\mathbf{0}\}$, while exactness of the second part

$$V \xrightarrow{S} W \xrightarrow{\pi} \{\mathbf{0}\}$$

tells us that the map S is onto (surjective) as the kernel of π is all of W, which by exactness is the image of S.

Back to the proof of the theorem, part (iii) can be proved by writing the map as a matrix and solving the system of linear equations.

For example when $m = 3$ and $n = 2$, we have the following.

Given a matrix

$$M = \begin{bmatrix} A & B & C \\ D & E & F \end{bmatrix}$$

we have the matrix equation

$$M\mathbf{v} = \mathbf{w}$$

where $\mathbf{q} = \mathbf{w}$ is fixed, and M is fixed, and by the *solution set* of this equation we mean the collection of all $\mathbf{v}$ which satisfy this equation. Writing $\mathbf{w} = (s, t)$ we have

$$\begin{bmatrix} A & B & C \\ D & E & F \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} s \\ t \end{bmatrix}. \tag{14}$$

The multiplication $\mathbf{v} \mapsto M\mathbf{v}$ defines a linear transformation $T : \mathbb{R}^3 \to \mathbb{R}^2$. Note that $\text{Im}(T)$ is equal to the *column space* of M, the subspace of $\mathbb{R}^3$ generated by the columns of the matrix. This is simply because for a standard column basis vector $\mathbf{e}_k$, $M\mathbf{e}_k$ gives the k^{th} column of M.

Note that the matrix equation (35) is equivalent to the "system of two linear equations in three unknowns":

$$\begin{cases} Ax + By + Cz = s \\ Dx + Ey + Fz = t \end{cases}$$

This system has *full rank* iff the rows are linearly independent, iff the dimension of the image $\text{Im}(T)$ is the maximum possible, in this case 2.

From Linear Algebra we can find the solution set explicitly by *row reduction.*

Example 11. For a concrete example, after row reduction we may have the matrix equation

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} s \\ t \end{bmatrix}. \tag{15}$$

which gives the system

$$\begin{cases} x + y = s \\ 2z = t \end{cases}$$

and we are free to choose y (for this reason known as a "free variable") but then no longer free to choose x or z as these are determined, since $x = -y + s$ and $z = t/2$.

Thus we have for a solution

$$(x, y, z) = (-y + s, y, t/2) = (s, 0, t/2) + y(-1, 1, 0) = \mathbf{p} + y\mathbf{v} = l(y)$$

which is a parametrized line passing through the point $\mathbf{p}$ in the direction of $\mathbf{v}$.

If we change s, t we get lines parallel to this one. In particular, if $\mathbf{p} = (0, 0)$ then the solution set is the parametric line $l(y) = y\mathbf{v}$, and this is the kernel of the map T, of dimension 1.

In conclusion, the dimension of the solution set is the number of free variables, so in this case of full rank this is, by Cor. 3.25, $3 - 2 = 1$, indeed a line.

Fixing the constants in the system to be $s = 1, t = 4$, the parametric line has the matrix form

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = y \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}. \tag{16}$$

The composition of the two maps is

$$\begin{aligned}
[y] &\mapsto \begin{bmatrix}1 & 1 & 0\\ 0 & 0 & 2\end{bmatrix}\left(y\begin{bmatrix}-1\\ 1\\ 0\end{bmatrix}+\begin{bmatrix}1\\ 0\\ 2\end{bmatrix}\right)\\
&= y\left(\begin{bmatrix}1 & 1 & 0\\ 0 & 0 & 2\end{bmatrix}\begin{bmatrix}-1\\ 1\\ 0\end{bmatrix}\right)+\left(\begin{bmatrix}1 & 1 & 0\\ 0 & 0 & 2\end{bmatrix}\begin{bmatrix}1\\ 0\\ 2\end{bmatrix}\right)\\
&= \begin{bmatrix}0\\ 0\end{bmatrix}+\begin{bmatrix}1\\ 4\end{bmatrix}=\begin{bmatrix}1\\ 4\end{bmatrix}
\end{aligned} \tag{17}$$

which is correct, since the every point on the parametric line is in the solution set of the system.

The geometrical way to think of this is that each of the equations in the system gives a plane, having as normal vector the corrsponding row of the matrix, so the solutions for the pair of equations is the intersection of two planes which is a line, which is perpendicular to both row vectors. The full rank condition means that these planes are not parallel, since their normal vectors are the rows of M, which are linearly independent.

3.19. **The Inverse and Implicit Function Theorems.** What we will see next is how this same point of view applies for the much more general situation of a differentiable, but *nonlinear*, map.

We know that given $F : \mathbb{R}^m \to \mathbb{R}^n$ continuously differentiable, the derivative matrix $DF|_{\mathbf{p}}$ well–approximates the function at the point $\mathbf{p}$. That means that certain properties of F near $\mathbf{p}$ should be reflected in the linear map $DF|_{\mathbf{p}}$ and vice-versa.

Important examples are given by these two theorems, which are closely related. Indeed one can choose to prove either one first, then deducing the other from that.

The *Inverse Function Theorem* states that $F : \mathbb{R}^m \to \mathbb{R}^m$ is invertible near $\mathbf{p}$ iff the matrix $DF|_{\mathbf{p}}$ is invertible, which is true iff $\det(DF|_{\mathbf{p}}) \neq 0$. First we need:

Definition 3.15. $F : \mathbb{R}^m \to \mathbb{R}^n$ is *continuously differentiable* (of class $\mathcal{C}^1$) iff the derivative DF at each point $\mathbf{p}$ exists and the matrix $DF_{\mathbf{p}}$ is a continuous function of $\mathbf{p}$).

Given an open set $\mathcal{U} \subseteq \mathbb{R}^m$, a function $F : \mathcal{U} \to \mathcal{V} = F(\mathcal{U})$ is *invertible* iff there exists $\widetilde{F}$ defined on $\mathcal{V}$ such that $\widetilde{F} \circ F$ is the identity on $\mathcal{U}$ and $F \circ \widetilde{F}$ is the identity on $\mathcal{V}$.

A *parametrized submanifold* $M \subseteq \mathbb{R}^m$ of dimension $d < m$ is the following: there exists $\mathcal{U} \subseteq \mathbb{R}^d$ and $\Phi : \mathcal{U} \to \mathbb{R}^d$ such that Φ is $\mathcal{C}^1$, M is the image $\Phi(\mathcal{U})$, and such that at each point of $\mathcal{U}$,$D\Phi$ is injective to a linear subspace of $\mathbb{R}^m$ of dimension d.

Example 12. A $\mathcal{C}^1$ curve γ in $\mathbb{R}^m$ is a parametrized submanifold of $\mathbb{R}^m$ of dimension one. A parametrized surface in $\mathbb{R}^m$ is a $\mathcal{C}^1$ function S from an open subset $\mathcal{U} \subseteq \mathbb{R}^2$ to its image $\mathbb{S} \subseteq \mathbb{R}^m$ such that the derivative matrix $DS|_{\mathbf{p}}$ is injective for every $\mathbf{p} \in \mathcal{U}$.

Examples are given by cylindrical and spherical coordinates, for a cylinder and sphere respectively:

Cylindrical coordinates on $\mathbb{R}^3$ are given by the map $\alpha : \mathbb{R}^3 \to \mathbb{R}^3$ with $\alpha(r, \theta, z) = (x, y, z)$ where

$$\begin{aligned} x &= r\cos(\theta) \\ y &= r\sin(\theta) \\ z &= z \end{aligned} \tag{18}$$

The cylinder of radius a and height c is the image by α of the subset $r = a, \theta \in [0, 2\pi), z \in [0, c]$. Defining the map S on $[0, 2\pi) \times [0, c]$ by $S(\theta, z) = (a\cos\theta, a\sin\theta, z)$ gives the cylinder as a parametrized surface.

Exercise 3.16. Check that the derivative matrix is:

$$D\alpha = \begin{bmatrix} x_r & x_\theta & x_z \\ y_r & y_\theta & y_z \\ z_r & z_\theta & z_z \end{bmatrix} = \begin{bmatrix} \cos\theta & -r\sin\theta & 0 \\ \sin\theta & r\cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{19}$$

Calculate the determinate of this matrix.

Spherical coordinates on $\mathbb{R}^3$ are given by the map $\beta : \mathbb{R}^3 \to \mathbb{R}^3$ with $\widetilde{S}(\rho, \theta, \varphi) = (x, y, z)$ where $r = \rho\sin(\varphi)$ and so

$$\begin{aligned} x &= r\cos(\theta) = \rho\sin(\varphi)\cos(\theta) \\ y &= r\sin(\theta) = \rho\sin(\varphi)\sin(\theta) \\ z &= \rho\sin(\varphi) \end{aligned} \tag{20}$$

The sphere of radius a is the image by α of the subset $\rho = a, \theta \in [0, 2\pi), \varphi \in [0, \pi]$. Defining the map S on $[0, 2\pi) \times [0, \pi]$ by $\widetilde{S}(\theta, \varphi) = (a, \theta, \varphi)$ gives the sphere as a parametrized surface.

Exercise 3.17. Calculate the derivative matrix and its determinant.

Theorem 3.26. *(Inverse Function Theorem) Let $F : \mathbb{R}^m \to \mathbb{R}^m$ be $\mathcal{C}^1$. Suppose the matrix $DF_{\mathbf{p}}$ is invertible. Then there exists an open set $\mathcal{U}$ containing $\mathbf{p}$ such that F is $\mathcal{C}^1$ and invertibl e on $\mathcal{U}$.*

Proof. See [Mar74], p. 206 and p. 230 or (for a stronger statement, with estimates) [HH15] p. 264 ff. □

Remark 3.10. Note that by the Chain Rule, we then know that for all points $\mathbf{x} \in \mathcal{U}$, with $\mathbf{y} = F(\mathbf{x})$, $\widetilde{F} \circ F(\mathbf{x}) = \mathbf{x}$ so $I = D(F \circ \widetilde{F})(\mathbf{x}) = (DF)_{\widetilde{F}(\mathbf{x})} DF_{\mathbf{x}}$ whence the inverse of the matrix $DF_{\mathbf{x}}$ is $(DF_{\mathbf{x}})^{-1} = D\widetilde{F}_{\mathbf{y}}$.

The *Implicit Function Theorem* states that for a $\mathcal{C}^2$ function $F : \mathbb{R}^m \to \mathbb{R}^n$ with $m \geq n$, then if the derivative matrix at a point $\mathbf{p}$ is surjective (onto; of full rank) then the inverse image set $F^{-1}(\mathbf{q})$ for $F(\mathbf{p}) = \mathbf{q}$ behaves like the inverse image of a point by the matrix: it is a *submanifold* of dimension $m - n$. For the linear (matrix) case see Cor. 3.25.

A submanifold of dimension 1 is a parametrized curve; of dimension 2 is a *parametrized surface*. Note that the case $m = n$ says the following: the inverse image of a point $F^{-1}(\mathbf{q})$ is a manifold of dimension $m - n = 0$, in other words a single point. Thus F is an invertible function, which is just the statement of the Inverse Function Theorem!

In the case $n = 1$, the Implicit Function Theorem moreover gives conditions when given an equation

$$F(x_1, \dots, x_n) = 0$$

we can solve for one of the variables, and use the rest as our parameters.

Example 13. For the simplest example, $F : \mathbb{R}^2 \to \mathbb{R}$ with $F(x, y) = x^2 + y^2$, the curve of level 1 is the unit circle, the solutions of the equation (i.e. all pairs (x, y) which satisfy the equation)

$$x^2 + y^2 = 1.$$

Solving for y gives $y = \pm\sqrt{1 - x^2}$. See Exercise 3.7.

Theorem 3.27. *(Implicit Function Theorem for $\mathbb{R}^m$) Let $F : \mathbb{R}^m \to \mathbb{R}^n$ be $\mathcal{C}^1$, with $m \geq n$. Suppose the matrix $DF_{\mathbf{p}}$ is surjective. Then for $\mathbf{q} = F(\mathbf{p})$, the set $F^{-1}(\mathbf{q})$ is a submanifold of $\mathbb{R}^m$ of dimension $m - n$. That is, for $\mathbf{p} \in F^{-1}(\mathbf{q})$, there exists an open subset $\mathcal{U} \subseteq \mathbb{R}^{m-n}$ and $H : \mathcal{U} \to \mathbb{R}^m$ $\mathbb{C}^1$ such that $F \circ G(\mathbf{p}) = \mathbf{q}$ for all $\mathbf{p} \in \mathcal{U}$.*

Proof. See [War71] Theorem 1.38, p. 31. □

For example, if $F : \mathbb{R}^3 \to \mathbb{R}$, then the level surface $F^{-1}(q)$ is a parametrized surface: its parametrization is given near the point $\mathbf{p}$ by the map H.

Remark 3.11. Similarly to the linear case as explained in Remark 3.9, in the smooth case, the same set (an embedded manifold) is viewed in two different ways, by means the two maps, one where it is the image, one the domain. The first parametrizes the manifold, the second places it as a level curve, surface or manifold of a map on the higher-dimensional space, and thus shows how it is but one of a family of such "parallel" manifolds. This is a special mathematical object known as a *foliation*. Thus the level surfaces of a function $F : \mathbb{R}^3 \to \mathbb{R}$ foliate $\mathbb{R}^3$, and in the special case of F linear, F is given by the inner product with a normal vector, and the foliation consists of all those parallel planes.

Thus a *parametrized m-dimensional manifold in $\mathbb{R}^m$* is a map $\alpha : \mathcal{U} \to M \subset \mathbb{R}^m$ where $\mathcal{U}$ is a connected open subset of $\mathbb{R}^m$, α is differentiable and invertible with image M. (For the definition of connectedness see Def. 5.4).

The higher dimensional version of level curves and surfaces can be stated as follows. Given $f : \mathbb{R}^{m+1} \to \mathbb{R}$ then if f is differentiable and surjective and Df is everywhere onto (one says Df is of *maximal rank*) then for any point $q \in \mathbb{R}$ the set $M = f^{-1}(q)$ is locally a parametrized m-dimensional manifold. Moreover this holds when that condition holds at some point (not necessarily all points): for any $y = f(\mathbf{p})$ such that $Df|_{\mathbf{p}}$ is of maximal rank; q is then called a *regular value.*

Here is a statement of the

Proposition 3.28. *(Implicit Function Theorem for differentiable manifolds) (Lemmas 1,2 of Chapter 2 of* [MW97]*) If $f : M \to N$ is a smooth map between manifolds of dimension $m \geq n$, and if $\mathbf{q} \in N$ is a regular value, such that it is a value of the map (i.e. $f^{-1}(\mathbf{q})$ is nonempty) then the set $f^{-1}(\mathbf{q})$ is a smooth manifold of dimension $m - n$. The null space of $Df_{\mathbf{p}} : TM_{\mathbf{p}} \to TN_{\mathbf{q}}$ is the tangent space of this submanifold, and its orthogonal complement is mapped onto $TN_{\mathbf{q}}$.*

This says that we have a diagram

$$K \xrightarrow{\alpha} M \xrightarrow{f} N$$

where the first map is injective and the second is surjective. When one considers the derivative maps then one gets the exact diagram for the linear case of Example 11; here $\alpha(\mathbf{x}) = \mathbf{p}$. See Cor. 3.25:

$$\{\mathbf{0}\} \longrightarrow K \xrightarrow{D\alpha|_{\mathbf{x}}} M \xrightarrow{Df_{\mathbf{p}}} N \longrightarrow \{\mathbf{0}\}$$

There are many versions of these theorems. For an introduction see Lemmas 1,2 of Chapter 2 of [MW97] and for surfaces Proposition 2 of Chapter 2, p. 59 of [DC16]. For a simple and beautiful general statement see Theorem 1.39 of [War71]. More on the Implicit and related Inverse Function Theorems are given e.g. in §7.2-4 of [Mar74], and in Chapter 2.10 and on p. 729 of [HH15].

Theorem 3.27 raises implicitly the question of, for a given $F : \mathbb{R}^m \to \mathbb{R}^n$, *which* of the variables $x_1, \dots, \mathbf{x}_m$ can be used to parametrize the submanifold. The basic principle is that the answer is the same as for the linear algebra; that is, we look at the matrix $DF_{\mathbf{p}}$. For a modification of Example 11 above,

$$\begin{bmatrix} 1 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \mathbf{q} \tag{21}$$

we have $3 - 2 = 1$ for the dimension of the solution set $K = F^{-1}(\mathbf{q})$, and could take x or y as the free variable, but not z. The geometrical reason is that the solution set is $l(y) = \mathbf{p} + y\mathbf{v}$ where $\mathbf{p} = (1, 0, 1)$ and $\mathbf{v} = (-1, 1, 0)$. This is a line in the plane $z = 0$ so altough the solution set is one-dimensional, it can't be parametrized by z, but only (in terms of the standard coordinates) by x or y.

Proposition 3.29. *If the* (2×3) *matrix*

$$A = \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} \tag{22}$$

has rank 2*, then the solution set has dimension* $3 - 2 = 1$*; and moreover if the* (2×2) *submatrix*

$$\begin{bmatrix} b & c \\ e & f \end{bmatrix}$$

is invertible, then we can take x *as the free variable (the parameter), and similarly for the other* (2×2) *submatrices.*

A nice simpler version of the Implicit Function Theorem, with examples like this, is given on pp. 239-240 of Vol. II of [Gui02]. See also my handwritten *Notas de Aula*. For a general statement like this example, see Hubbard's book Thm. 2.10.14.

3.20. **Higher derivatives.** So far, since §3.9, we have been studying the derivative of a map $F : \mathbb{R}^m \to \mathbb{R}^n$, using the matrix of partial derivatives. This gives the best linear (or affine) approximation to the map, which is also called the best *first-order* approximation. Thus we have the tangent line to the graph of $f : \mathbb{R} \to \mathbb{R}$, the tangent line to a curve $\gamma : \mathbb{R} \to \mathbb{R}^n$, and the tangent plane to the graph of $F : \mathcal{F}^2 \to \mathbb{R}$.

For the best *second-order* approximation we add a term involving the second-order partial derivatives and so on. This gets more and more complicated as we describe next.

The map F is called $\mathcal{C}^0$ iff it is continuous, and $\mathcal{C}^k$ iff the k^{th} derivative exists and is continuous. For this we need to define higher derivatives.

Writing $L(V, W)$ for the collection of linear transformations from V to W, then this is a Banach space with the operator norm. Now for $F : V \to W$, the derivative DF is, at each point $\mathbf{x}$, a linear map from V to W, thus an element of the vector space $L(V, W)$. In other words, given $F : V \to W$, then DF is a map (*nonlinear* in general) from V to $L(V, W)$. Since $DF : V \to L(V, W)$, then the second derivative at the point $\mathbf{x}$ must be a *linear map* $D^2F_{\mathbf{x}} : V \to L(V, W)$. This means that $D^2F_{\mathbf{x}}$ is in the collection of linear maps from V to $L(V, W)$, in other words it is a nonlinear map $D^2F : V \to L(V, L(V, W))$.

Since we can represent the derivative at a point by a matrix, we see that these are increasing in size: If $V = \mathbb{R}^m$ and $W = \mathbb{R}^n$, then $DF_{\mathbf{x}}$ is $(n \times m)$, $D^2F_{\mathbf{x}}$ is $(n \times (nm))$, $D^3F_{\mathbf{x}}$ is $(n \times (n^2m))$; so $D^kF_{\mathbf{x}}$ is a matrix with n^km entries, getting more and more complicated quickly.

The only exception is when $n = 1$, for a curve $\gamma : [a, b] \to \mathbb{R}^m$: in this case (as noted above) then γ' is also curve in $\mathbb{R}^m$, thus so is $\gamma'' = (\gamma')'$ etcetera. By contrast for a function $F : \mathbb{R}^n \to \mathbb{R}$ then the gradient $\nabla F : \mathbb{R}^n \to \mathbb{R}^n$ is a vector field, so $DF : \mathbb{R}^n \to \mathbb{R}^{n^2}$, and then the second derivative is no longer a vector field as $D^2F : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^{n^2} \sim \mathbb{R}^{n^3}$, and so on, with the matrices getting larger and larger.

A *domain* is an open subset of $\mathbb{R}^n$. A *vector field* on a domain $\mathcal{U}$ is simply a such a map defined only on the subset $\mathcal{U}$. The vector field is termed $\mathcal{C}^k$, for $k \geq 0$, iff the map has those properties (again, $\mathcal{C}^0$ means continuous, and $\mathcal{C}^k$ that D^kF exists and is continuous, so $D : \mathcal{C}^{k+1} \to \mathcal{C}^k$).

3.21. **Higher order partials.** Given $F : \mathbb{R}^2 \to \mathbb{R}$ then $F_x = \frac{\partial F}{\partial x}$ is a function defined on the plane. Then setting $G(x, y) = F_x(x, y)$ we take its partial derivatives. We write $\frac{\partial}{\partial y}G$ in these equivalent ways:

$$G_y = \frac{\partial}{\partial y}(G) = \frac{\partial}{\partial y}(F_x) = \frac{\partial F_x}{\partial y} = (F_x)_y = F_{yx}.$$

(This notation can be confusing since $F_{yx} = (F_x)_y$!)

Now for $G = F_x$ then $G : \mathbb{R}^2 \to \mathbb{R}$. This has as its gradient

$$\nabla G = (G_x, G_y) = (F_{xx}, F_{yx}).$$

Similarly for $\widetilde{G} = F_y$ then its gradient is

$$\nabla \widetilde{G} = (\widetilde{G}_x, \widetilde{G}_y) = (F_{xy}, F_{yy}).$$

When a function $L : \mathbb{R}^2 \to \mathbb{R}^2$ then writing it in components $L = (L_1, L_2)$ or in matrix form

$$[L] = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix}$$

we know its derivative matrix is

$$[DL] = \begin{bmatrix} \nabla L_1 \\ \nabla L_2 \end{bmatrix} = \begin{bmatrix} (L_1)_x & (L_1)_y \\ (Ł_2)_x & (L_2)_y \end{bmatrix}$$

So for $DF : \mathbb{R}^2 \to \mathbb{R}^2$ we have

$$D^2F = D(DF) = \begin{bmatrix} (F_x)_x & (F_x)_y \\ (F_y)_x & (F_y)_y \end{bmatrix} = \begin{bmatrix} F_{xx} & F_{yx} \\ F_{xy} & F_{yy} \end{bmatrix}.$$

Now in fact it is a bit simpler than this because of the following:

Proposition 3.30. *For $F : \mathbb{R}^2 \to \mathbb{R}$ then $F_{xy} = F_{yx}$.*

3.22. **Equality of mixed partials.** The above important fact in Proposition is often called the *equality of mixed partials.* The next result can be proved using just derivatives, but we like the following "Fubini's Theorem argument", partly because it leads in to Green's Theorem later on. For this proof we need double integrals, see §4.

Lemma 3.31. *If $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentable, then we can change the order in taking two partial derivatives: e.g. for $n = 2$, then*

$$\frac{\partial}{\partial x}\left(\frac{\partial \varphi}{\partial y} f\right) = \frac{\partial \varphi}{\partial y}\left(\frac{\partial \varphi}{\partial x} f\right).$$

Proof. Given two continuous functions $\varphi, \widetilde{\varphi} : \otimes \to \mathbb{R}$ on an open set Ω, then if for every rectangle $R \subseteq \Omega$ we have

$$\int\int_R \varphi \, \mathrm{d}x \, \mathrm{d}y = \int\int_R \widetilde{\varphi} \, \mathrm{d}x \, \mathrm{d}y,$$

then we can conclude that $\varphi = \widetilde{\varphi}$ on Ω. (Because, if they differ at a point, then one is larger than the other on a small rectangle about that point, and the integrals there are different, a contradiction).

We define $\varphi(x, y) = \frac{\partial \varphi}{\partial x}(\frac{\partial \varphi}{\partial y} f(x, y))$ and $\widetilde{\varphi} = \frac{\partial \varphi}{\partial y}(\frac{\partial \varphi}{\partial x} f(x, y))$. Our strategy of proof will be to show that for any $R = [a, b] \times [c, d]$ we have the above equality of integrals, and the result will then follow.

Fubini's Theorem tells us that

$$\int\int_R \varphi(x, y) \, \mathrm{d}x \, \mathrm{d}y = \int_c^d \left(\int_a^b \varphi(x, y) \, \mathrm{d}x\right) \, \mathrm{d}y = \int_c^d \left(\int_a^b \frac{\partial}{\partial x}\left(\frac{\partial}{\partial y} f(x, y)\right) \mathrm{d}x\right) \, \mathrm{d}y$$

Now for any differentiable function $G(x, y)$ we have by the Fundamental Theorem of Calculus that for any fixed y, $\int_a^b \frac{\partial}{\partial x} G(x, y) \mathrm{d}x = G(a, y) - G(b, y)$, so

$$\int_a^b \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right)(x, y) \mathrm{d}x = \frac{\partial f}{\partial y}(b, y) - \frac{\partial f}{\partial y}(a, y)$$

so the iterated integral equals

$$\int_c^d \frac{\partial}{\partial y} f(b,y)\mathrm{d}y - \int_c^d \frac{\partial}{\partial y} f(a,y)\mathrm{d}y =$$
$$\Big(f(b,d) - f(b,c)\Big) - \Big(f(a,d) - f(a,c)\Big).$$

Again, by Fubini's Theorem:

$$\int\int_R \widetilde{\varphi}(x,y)\ \mathrm{d}x\ \mathrm{d}y = \int_a^b \int_c^d \widetilde{\varphi}(x,y)\ \mathrm{d}y\ \mathrm{d}x = \int_a^b \left(\int_c^d \frac{\partial}{\partial y}\left(\frac{\partial \varphi}{\partial x} f(x,y)\right)\mathrm{d}y\right)\ \mathrm{d}x$$

This time,

$$\int_c^d \frac{\partial}{\partial y}\left(\frac{\partial}{\partial x} f(x,y)\right)\mathrm{d}y = \frac{\partial f}{\partial x}(x,d) - \frac{\partial f}{\partial x}(x,c)$$

so the iterated integral equals

$$\int_a^b \frac{\partial}{\partial x} f(x,d)\mathrm{d}x - \int_a^b \frac{\partial}{\partial x} f(x,c)\mathrm{d}x =$$
$$\Big(f(b,d) - f(a,d)\Big) - \Big(f(b,c) - f(a,c)\Big)$$

which equals the previous expression, finishing the proof. □

Corollary 3.32. *The above matrix D^2F is* symmetric*: it has the form* $\begin{bmatrix} a & b \\ b & c \end{bmatrix}$.

In the case of $F : \mathbb{R}^n \to \mathbb{R}$, all of this makes sense: D^2F is a a symmetric $(n \times n)$ matrix. For $F : \mathbb{R}^n \to \mathbb{R}$ the matrix D^2F is sometimes called the *Hessian matrix.* Its determinant is called the *Hessian determinant* or simply the *Hessian.* In Guidorizzi Vol 2 §16.3 this is written $H(x,y)$. See [Gui02], and §3.6 of [HH15].

The meaning of this symmetric matrix becomes clear when discussing Taylor polynomials of order 2, and finding maximums and minimums.

3.23. **Finding maximums and minimums.** We note that:
(1) Given $F : \mathbb{R}^2 \to \mathbb{R}$, if a minimum or maximum value occurs at $\mathbf{p} = (x_0, y_0)$ then the tangent plane must be horizontal. Equivalently, if F is differentiable, then the partial derivatives F_x, F_y at $\mathbf{p}$ are 0.

Definition 3.16. In this case, $\mathbf{p}$ is a *critical point (ponto critico)* of F. Equivalently, $\nabla F(\mathbf{p}) = \mathbf{0}$.

(2) If it is a maximum then it must be a maximum for the function restricted to the line $x = x_0$. We can then consider the second partials and use the second derivative test from Calculus 1: If $F_{xx} > 0$ then F_x is increasing, so it is a minimum along that line. This does not necessarily mean it is a minimum off the line.

However there is a fuller method: see Guidorizzi Vol 2 §16.3 [Gui02], and §3.6 of [HH15].

Theorem 3.33. *Let $F : \mathbb{R}^2 \to \mathbb{R}$ be in $\mathbb{C}^2$. We have:*

1. If F has a local max or min at $\mathbf{p}$ then the tangent plane there is horizontal. Equivalently, $F_x = F_y = 0$ at $\mathbf{p}$.

2. The second partials tell us the following:

(i) If $F_{xx} > 0$ at $\mathbf{p}$ and the Hessian $H(\mathbf{p}) > 0$ then $\mathbf{p}$ is a local minimum.

(i) If $F_{xx} < 0$ at $\mathbf{p}$ and the Hessian $H(\mathbf{p}) > 0$ then $\mathbf{p}$ is a local maximum.

(iii) if $H(\mathbf{p}) < 0$ then $\mathbf{p}$ is a saddle point. Thus it is neither max nor min.

(iv) If $H(\mathbf{p}) = 0$ then we cannot say from this test and have to look more closely.

Exercise 3.18. Compare the above tests for the functions we have encountered, see Figs. 3, 2: $F(x, y) = x^2 + y^2$, $F(x, y) = x^2 - y^2$, $F(x, y) = xy$.

The best way to understand this theorem is via the Taylor polynomials of order two, in two variables. First we examine the case of a single variable, and then return to the above.

3.24. **The Taylor polynomial and Taylor series. Taylor series in one dimension.**

Let us recall that a *polynomial of degree n* is

$$p(x) = a_0 + a_1 x + \cdots + a_k x^k + \cdots + a_n x^n.$$

Here k is a nonnegtive integer and $a_k \in \mathbb{R}$. So since $x^0 = 1$ for any $x \in \mathbb{R}$, this is equal to

$$p(x) = a_0 x^0 + a_1 x^1 + \cdots + a_n x^n = \sum_{k=0}^{n} a_k x^k.$$

(Here we use the definition $0! = 1$.) Thus a polynomial of degree 0 is a constant function $p_0(x) = a_0$, of degree one is an affine function, $p_1(x) = a_0 + a_1 x$, of degree two is quadratic and so on.

Polynomials are great to work with as it is easy to compute their derivatives, integrals and to draw their graphs, and to compute their values at a point you only need to add and multiply. Also, as we describe in this section, more complicated functions (for example, the *transcendental functions* $\sin, \cos, \exp, \log, \tan \ldots$ and the *rational functions* (ratios of polynomials, such as $1/(1-x)$) can often be approximated quite well by polynomials.

Given a map $f : \mathbb{R} \to \mathbb{R}$, the terminology "$k^{\text{th}}$-order approximation" to f at a point $x \in \mathbb{R}$ comes from approximating the function with a certain polynomial of degree k. The polynomial of degree k which best fits the map near that point is termed "the best k^{th}-order approximation at $x \in \mathbb{R}$". This is is called the Taylor polynomial of order k, and is the polynomial which has all the same derivatives at that point, up to order k, as f. The Taylor series is a kind of "polynomial of infinite order" which can, in the nicest cases (for example, $f = \sin, \cos, \exp$) reproduce it exactly as a convergent infinite series.

Thus the best 0^{th}-order approximation of f at $p \in \mathbb{R}$ is the constant map with constant equal to the value of f at that point: the map $p_0(x) = f(p)$ for all x. The $\varepsilon - \delta$ definition of continuity guarantees that this approximates the function fairly well if it is continuous at that point. If f is differentiable, we can do much better:

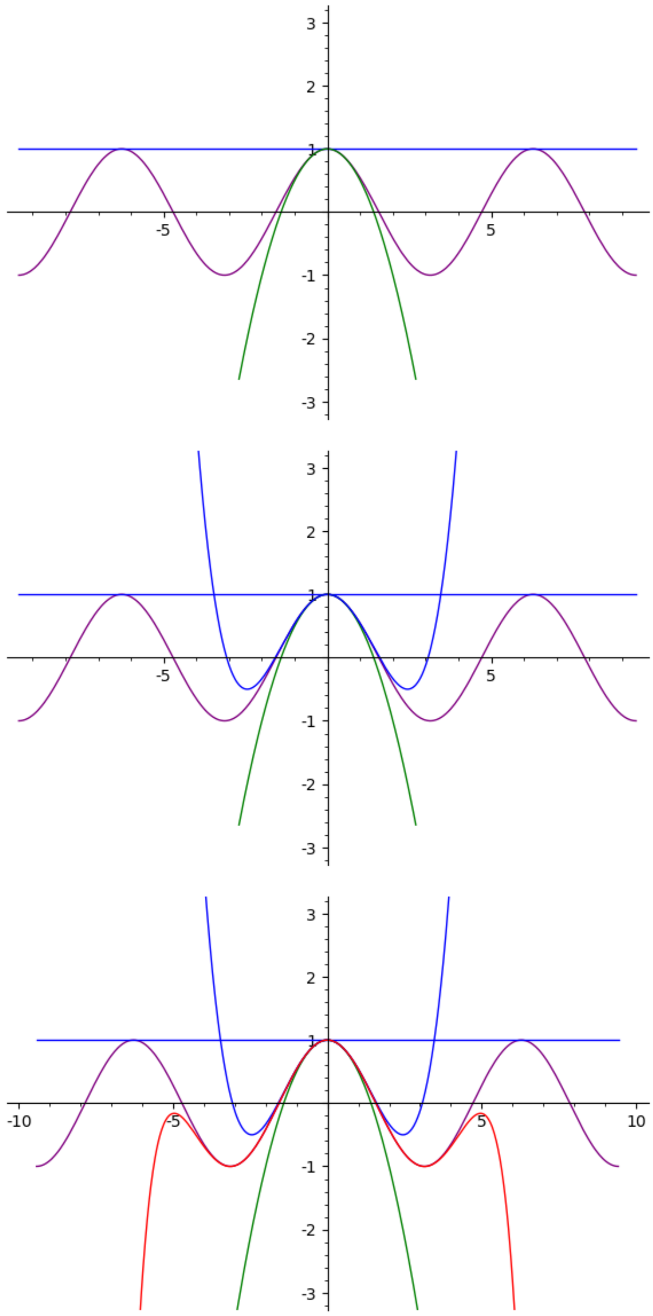


FIGURE 10. $\cos(x)$ and its Taylor polynomials p_n for $n = 0, 2, 4, 10$.

adding on the linear map given by the derivative $f'(p)$ gives the best *first-order approximation*. This is the affine map

$$p_1(x) = f(p) + f'(p)(x - p)$$

whose graph is the tangent line to the graph of f at that point.

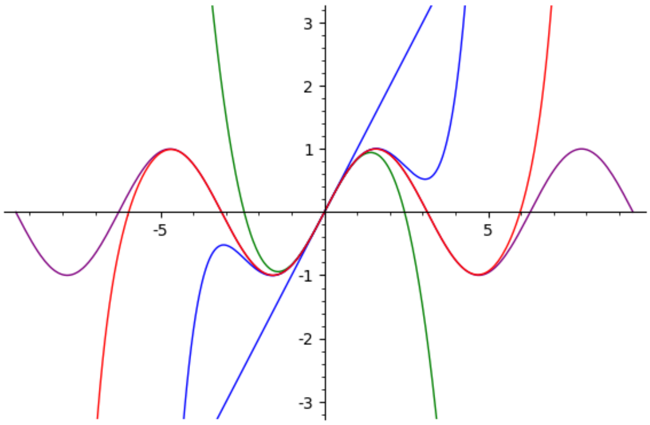


FIGURE 11. $\sin(x)$ and its Taylor polynomials p_n for $n = 1, 3, 5, 13$.

For a function $f : \mathbb{R} \to \mathbb{R}$ in $\mathcal{C}^\infty$, meaning it has derivatives of all orders, we define a sequence of polynomials $p_n(x)$, each of degree n, which approximate f better and better as $n \to \infty$. For this we choose a point about which we make the approximation, and call this the *Taylor polynomial about* x_0. Here for simplicity we work with $x_0 = 0$, and remark that the Taylor polynomials in this case are also called *Maclaurin polynomials*.

We write p_n for the n^{th} Taylor polynomial (about 0). We also say this is the *best n^{th}-order approximation* to f.

In the nicest cases, p_n actually converges to f as $n \to \infty$. This is true, for example, for $f(x) = \sin(x), \cos(x), e^x$.

The *Taylor series* is the infinite series which can be thought of as an *infinite polynomial*. For example, the n^{th} Taylor polynomial for e^x is

$$1 + x + x^2 + x^3/3! + \cdots + x^n/n!$$

and the Taylor series is

$$e^x = 1 + x + x^2 + x^3/3! + \cdots + x^n/n! + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k!} \tag{23}$$

For $f(x) = \sin x$ we have

$$\sin(x) = x - x^3/3! + x^5/5! - \dots$$

with only the odd powers and

$$\cos(x) = 1 - x^2/2! + x^4/4! - \dots$$

with only the even powers, and both with alternating signs.

That the graphs of the polynomials p_n do approach the function can be seen in Figs. 10, 11. The 0^{th} -order approximation is a horizontal line with that value, the 1^{st}-order approximation is the tangent line to the graph at that point. The 2^{nd}-order approximation is the parabola which best fits the curve, and so on. In Fig. 10, 11 we show the functions $f = \cos$ and sin together with some of the Taylor polynomials p_n. Note how close the fit becomes as n increases!

Exercise 3.19. Check that for e^x the derivative of p_{n+1} is p_n, and that for $\sin(x)$ the derivative of p_{n+1} is p_n for $\cos(x)$. This agrees with $(e^x)' = e^x$ and $(\sin)' = \cos, (\cos)' = -\sin$!!

The definition of the Taylor series for a differentiable function (about 0) is

$$\sum_{k=0}^{\infty} a_k x^k$$

where $a_k = f^{(k)}(0)/k!$ (here $f^{(k)}(0)$ is the k^{th} derivative of f at 0)). (Note that we write $f^{(0)}$ for f itself).

Thus the Taylor polynomial of degree n is

$$p_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!} x^k.$$

Exercise 3.20. (1) Check that this general formula does give the above Taylor series for e^x, $\sin(x)$ and $\cos(x)$.
(2) Show that the polynomial p_n has the same derivatives as f at $x = 0$, of order $0, 1, \dots, n$. That is, $p_n(0) = f(0), p_n'(0) = f'(0), p_n''(0) = f''(0)$, and so on.

To define the *Taylor polynomials about* x_0 we simply replace x by $(x - x_0)$ and the derivatives at 0 by those at x_0. Thus the Taylor polynomial of degree n about x_0 is

$$p_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k.$$

The Taylor polynomial in higher dimensions.

We can now understand the role of the Hessian matrix and Hessian determinant in understanding maxes and mins much better, by understanding the Taylor polynomial of a function of 2 variables.

Consider $F : \mathbb{R}^2 \to \mathbb{R}$. Then the best 0^{th} -order approximation about $\mathbf{0} = (0, 0)$ is the constant function with the value $f(\mathbf{0})$. The the 1^{st}-order approximation is the tangent *plane* to the graph at $\mathbf{0}$. The best 2^{nd}-order approximation may for example be a paraboloid but could instead be a parabolic hyperboloid, Fig. 2. This depends on the partial derivatives of order 2 at the point.

The Taylor polynomials p_n will be functions of two variables (x, y), thus $p_n : \mathbb{R}^2 \to \mathbb{R}$. Just as for one dimension, a polynomial of degree n is a linear combination of *basic terms* of degree k for $k = 0$ (a constant) up to $k = n$. Each basic term of degree k will be of the form $x^i y^j$ such that $(i + j) = k$. Thus for example the degree of $x^2 y^3$ is $2 + 3 = 5$, and the basic polynomials of degree 1 are $p(x, y) = x, p(x, y) = y$ and of order 2 are $p(x, y) = x^2, p(x, y) = y^2$ and $p(x, y) = xy$.

Taking a linear combination of terms of degree $\leq n$ gives a polynomial of degree n, for example

$$p(x, y) = 1 + x + 3y + x^2 + y^2 + 5xy$$

has degree 2.

The only polynomials in two variables of degree 0 are the constant functions, $p(x, y) = c$. Those of degree one have the form

$$p(x, y) = c + ax + by.$$

Thus the graph is a plane. For the case of the Taylor polynomial, this will be the tangent plane at that point.

Those with *only* terms of degrre two have a special name:

Definition 3.17. A *quadratic form* on $\mathbb{R}^2$ is a degree-two polynomial with no constant or linear terms, hence of the form

$$Q(x, y) = ax^2 + by^2 + cxy.$$

That is, it is a linear combination of all the possible terms of degree 2.

Consider for example $p(x) = x^2 + y^2$. Its graph is a paraboloid, while the graph of $q(x) = xy$ is a hyperbolic paraboloid. See Figs. 3, 2.

Both of these polynomials have degree 2. Both have a horizontal tangent plane at $\mathbf{0}$. The first has a minimum there while the second is a saddle point hence neither max nor min. For $q(x, y) = xy$, on the line $x = y$ we have $q(x, y) = xy = x^2$, an upward parabola, so a minimum along the line $x = y$. When $x = -y$ we have $q(x, y) = -x^2$, so a maximum. Thus $(0, 0)$ can be neither max nor min. This is the essence of a saddle point. Furthermore the level curves are hyperbolas; see Fig. 3.

In fact the terms of order 2 can be understood with the help of the Hessian, a key point being that this is a symmetric matrix.

Proposition 3.34. *Given a quadratic form Q, there is a symmetric (2×2) matrix A such that for $\mathbf{v} = \begin{bmatrix} x \\ y \end{bmatrix}$, then*

$$Q(\mathbf{v}) = \mathbf{v}^t A \mathbf{v}.$$

That is,

$$Q(\mathbf{v}) = \begin{bmatrix} x & y \end{bmatrix} A \begin{bmatrix} x \\ y \end{bmatrix}.$$

Proof. In fact, for $A = \begin{bmatrix} a & c/2 \\ c/2 & b \end{bmatrix}$ we have

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} a & c/2 \\ c/2 & b \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax^2 + by^2 + cxy \end{bmatrix} = Q(x, y).$$

□

Exercise 3.21. Check that when $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ then $Q(x, y) = 2xy$. What do we get for $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$? For $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$?

To graph a quadratic form we have the following:

Theorem 3.35. *A quadratic form*

$$Q(x,y) = ax^2 + by^2 + cxy = \begin{bmatrix} x & y \end{bmatrix} A \begin{bmatrix} x \\ y \end{bmatrix}$$

has either
(i) a local min or max at $\mathbf{0}$*, if* $\det A > 0$*;*
(ii) a saddle point at $\mathbf{0}$*, if* $\det A < 0$*.*
If $\det A = 0$*, we cannot tell from this test.*

Proof. (Sketch) From Linear Algebra, a symmetric matrix A can be diagonalized: there exists an orthogonal matrix U such that $U^{-1}AU = D$ where D is diagonal. Now an orthogonal matrix is a rotation, a reflection, or a product of these. That does not change whether $\mathbf{0}$ is a saddle point, max or min. Also, $\det D = \det U^{-1}AU = \det A$. This proves that $\det A$ is the product of its eigenvalues, since the eigenvalues of A and D are the same. The graph of the quadratic form defined by D has the two types described, completing the proof. See the above examples. □

We use this to study the Taylor polynomial of order 2 of $F : \mathbb{R}^2 \to \mathbb{R}$. Write A for the Hessian matrix. Then

$$A = D^2F = \begin{bmatrix} F_{xx} & F_{yx} \\ F_{xy} & F_{yy} \end{bmatrix} \tag{24}$$

as explained above in (26); note that A is symmetric!

Then the Taylor polynomials about $\mathbf{p} = (x_0, y_0)$ are, for degree 0:

$$p_0(x,y) = F(\mathbf{p})$$

For degree 1, writing $h = (x - x_0), k = (y - y_0)$ then

$$p_1(x,y) = F(\mathbf{p}) + F_x(\mathbf{p})(x - x_0) + F_y(\mathbf{p})(y - y_0) = F(\mathbf{p}) + F_x(\mathbf{p})h + F_y(\mathbf{p})k$$

(this is the tangent plane);

For degree 2 we have

$$p_2(x,y) = F(\mathbf{p}) + F_x(\mathbf{p})h + F_y(\mathbf{p})k + \frac{1}{2}\begin{bmatrix} h & k \end{bmatrix}\begin{bmatrix} F_{xx} & F_{yx} \\ F_{xy} & F_{yy} \end{bmatrix}\begin{bmatrix} h \\ k \end{bmatrix}$$
$$= F(\mathbf{p}) + F_x(\mathbf{p})h + F_y(\mathbf{p})k + \frac{1}{2}(F_{xx}h^2 + 2F_{xy}hk + F_{yy}k^2)$$

Thus

$$p_2(x,y) = p_1(x,y) + \begin{bmatrix} h & k \end{bmatrix}\frac{D^2F(\mathbf{p})}{2}\begin{bmatrix} h \\ k \end{bmatrix}$$

which reminds us of the formula for dimension 1. Note that the last term is a quadratic form since the Hessian matrix D^2F is symmetric.

Looking for maximum and minimum points, first we see if the tangent plane is horizontal. Then the first-order term is 0 so the Taylor polynomial is simply

$$p_2(x,y) = F(\mathbf{p}) + \begin{bmatrix} h & k \end{bmatrix}\frac{D^2F(\mathbf{p})}{2}\begin{bmatrix} h \\ k \end{bmatrix} = F(\mathbf{p}) + Q(x,y)$$

where Q is a quadratic form.

As shown in Theorem 3.35, the sign of the determinant then tells us whether the surface is a max or min, as for a paraboloid or *elliptic paraboloid* (like a paraboloid but with an ellipse as crossection) or a saddle point, as for $F(x,y) = xy$, discussed above.

For $F : \mathbb{R}^n \to \mathbb{R}$ with $n \geq 2$ a similar formula can be given. The higher terms of the Taylor series also have a nice expression when the derivatives D^kF are viewed as k-linear functions. For a clear treatment see [Mar74]. See also §3.6 of [HH15].

Remark 3.12. The Hessian (matrix) gives a *local form* for critical points, see the Morse Lemma in §1.7 of [GP74], and §I.2 of [Mil16], with a proof in proof Lemma 2.2. See also §7.6 of of [Mar74]. This is related to what we have seen about the Taylor series.

3.25. **Lagrange Multipliers.**

Theorem 3.36. *Given an open set $\mathcal{U} \subseteq \mathbb{R}^n$ and two $\mathcal{C}^1$ functions $F, G : \mathcal{U} \to \mathbb{R}$, let B be a level set for G, so $B = \{\mathbf{x} \in \mathcal{U} : G(\mathbf{x}) = c\}$. Assume that for some point $\mathbf{p} \in \mathcal{U}$, $\nabla G_{\mathbf{p}} \neq \mathbf{0}$. Then if F has a local maximum at $\mathbf{p} \in B$, there exists some $\lambda \in \mathbb{R}$ such that*

$$\nabla F_{\mathbf{p}} = \lambda \nabla G_{\mathbf{p}}.$$

Proof. Since $\nabla G_{\mathbf{p}} \neq \mathbf{0}$, the linear transformation (the matrix with those entries) $DG_{\mathbf{p}}$ is surjective, which allows us to use the Implicit Function Theorem, Theorem 3.27. So the level set B has a parametrization. Calling one of the coordinates t, we have a curve $\gamma(t)$ in B which passes through $\mathbf{p}$ at time 0, and such that given any chosen vector $\mathbf{v}$ tangent to B, $\gamma'(0) = \mathbf{v} \neq \mathbf{0}$.

Then $G(\gamma(t)) = c$ so for all t, $D(G \circ \gamma)(t) = 0$. By the Chain Rule this is

$$D(G \circ \gamma)(t) = DG_{\gamma(t)} D\gamma(t) = \nabla G_{\gamma(t)} \cdot \gamma'(t).$$

In particular fo $t = 0$, $\nabla G_{\mathbf{p}} \cdot \gamma'(0) = 0$.

On the other hand, F has a local maximum at $\mathbf{p}$, so in particular, $F \circ \gamma(t)$ has a local maximum at $t = 0$.

Therefore,

$$D(F \circ \gamma)(0) = DF_{\mathbf{p}} D\gamma(0) = \nabla F_{\mathbf{p}} \cdot \gamma'(0) = 0.$$

Since $\mathbf{v} = \gamma'(0) \neq \mathbf{0}$ was any tangent vector to B, both $\nabla F_{\mathbf{p}}$ and $\nabla G_{\mathbf{p}}$ are orthogonal to any such $\mathbf{v}$. Thus they must be multiples (think for example of a level curve, or a level surface).

□

Remark 3.13. Note that the derivatives are 0 for two completely different reasons: that $G \circ \gamma$ is constant, and that $F \circ \gamma$ has a maximum.

Note that it is possible for λ to be 0, and also possible for $\nabla F_{\mathbf{p}}$ to be $\mathbf{0}$. However for the proof $\nabla G_{\mathbf{p}}$ must be nonzero to be able to apply the Implicit Function Theorem.

4. Double and triple integrals (TO DO!)

4.1. **Review of Riemann integration (TO DO!).**

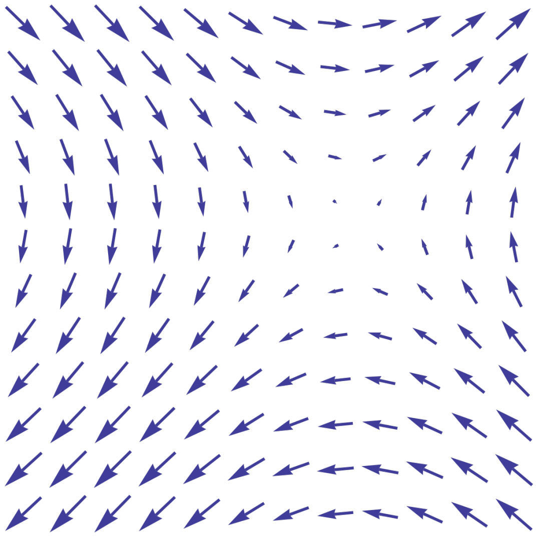


FIGURE 12. A vector field in the plane, from Wikipedia. Compare to the pictures of curves below!

5. Vector Calculus, Part II: the calculus of fields, curves and surfaces

5.1. **Vector Fields.** In Part I we we have already encountered the gradient vector field. Here is the general setting:

Definition 5.1. A *continuous* vector field is a continuous function $V : \mathbb{R}^m \to \mathbb{R}^m$. A *linear* or a *differentiable* vector field on $\mathbb{R}^m$ is simply a linear or differentiable such function.

The reason we call this a *vector field* rather than just a function is because of the special way in which we visualize this. Note that for $m \geq 2$ we cannot draw the graph of a vector field, as we would need too many dimensions! Indeed the graph of V is (by definition) the collection of all ordered pairs $(\mathbf{v}, V(\mathbf{v}))$, a point in $\mathbb{R}^m \times \mathbb{R}^m = \mathbb{R}^{2m}$ so for $\mathbb{R}^2$, to draw the graph of the vector field would require four dimensions.

Instead, we picture the vector field by drawing the vector $\mathbf{w_v} = V(\mathbf{v})$ *based at each point* $\mathbf{v}$. See Fig. 12. We can imagine this field represents the velocity field of a liquid or gas, showing its motion.

Exercise 5.1. Sketch the following linear vector fields $V(x, y) = (ax + by, cx + dy)$ given by the matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ acting on column vectors, that is:

$$A \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax + b \\ cx + d \end{bmatrix},$$

for these matrices:
(i) $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$.

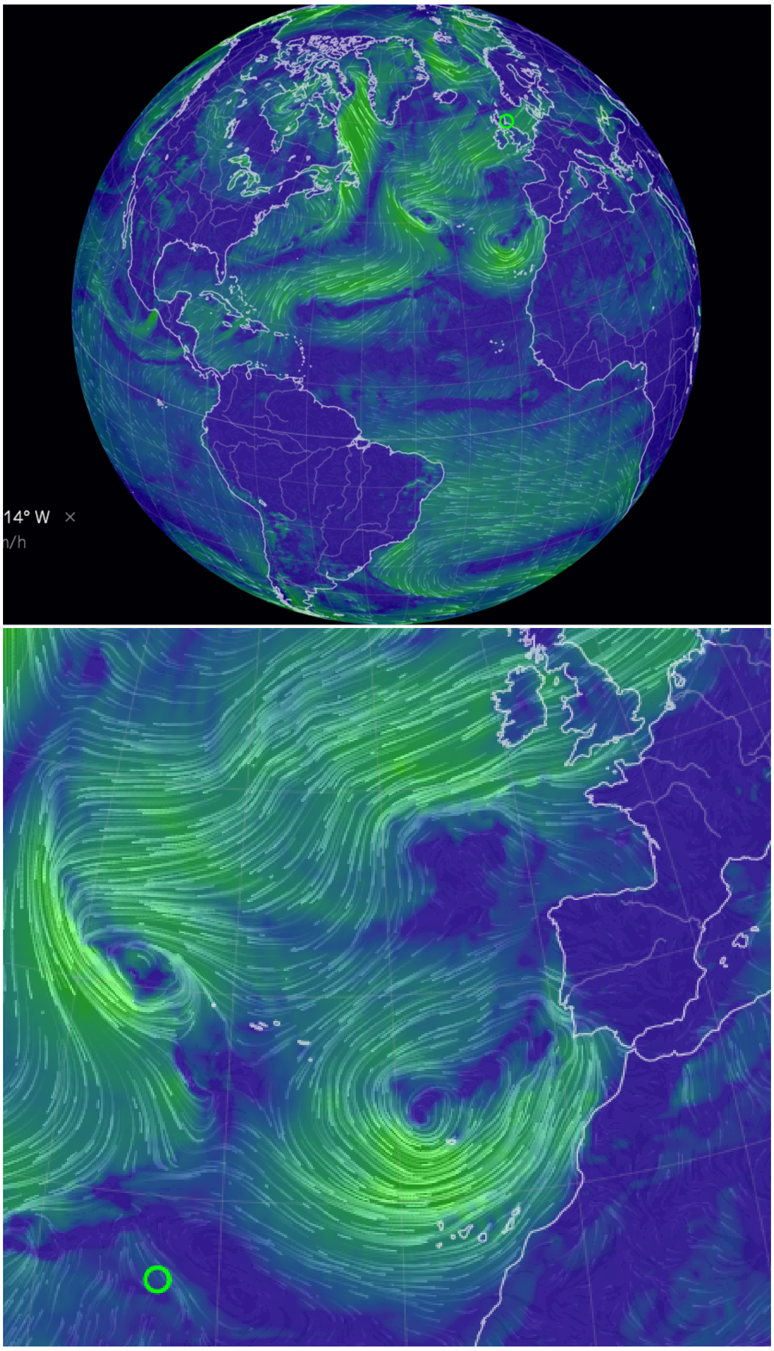


FIGURE 13. A time-varying velocity vector field: the wind at the surface of the Earth, from nullschool.net

(ii) $A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$

(iii) $A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$.

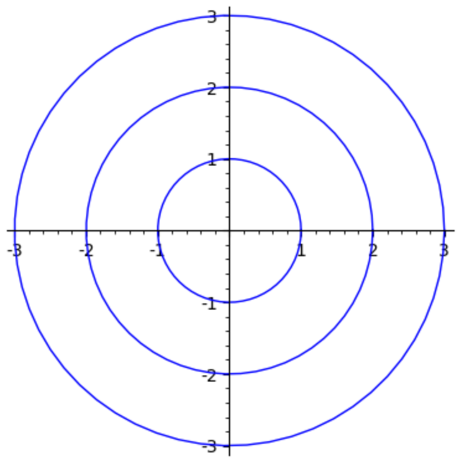


FIGURE 14. Level curves of the function $F(x, y) = x^2 + y^2$, tangent to the velocity vector field of the rotation flow.

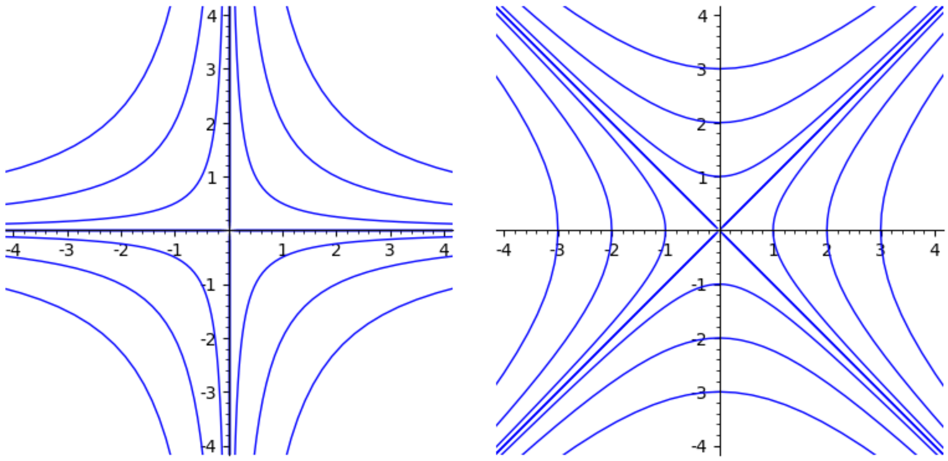


FIGURE 15. Two hyperbolic flows.

(iv) $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.

(v) $A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$.

Compare your sketches to the figures of curves.

Remark 5.1. **Vector fields, flows, and Ordinary Differential Equations** To better understand vector fields, we draw their *integral curves.* meaning that it satisfies $\gamma(\mathbf{0}) = \mathbf{p}$ and for all $t \in \mathbb{R}$, $\gamma'(t) = V(\gamma(t))$. That is, the curve is always *tangent to the vector field.*

Looking at all the curves at once, we see a continuous motion of $\mathbb{R}^n$, called a *flow*: by definition, a *flow* on $\mathbb{R}^n$ is a collection of maps $\tau_t : \mathbb{R}^n \to \mathbb{R}^n$ for each $t \in \mathbb{R}$, which

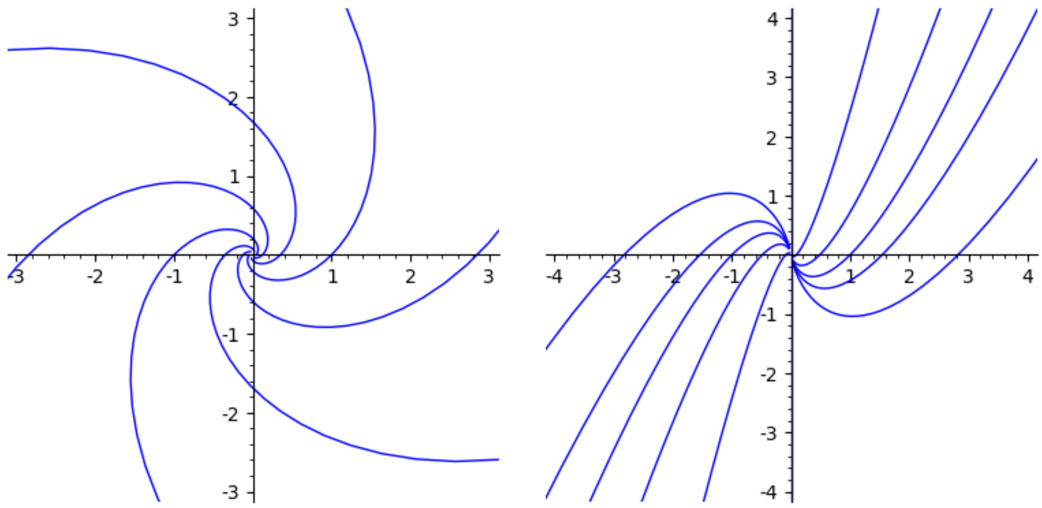


FIGURE 16. Integral curves of some linear vector fields.

satisfy the *flow property* $\tau_{t+s} = \tau_t \circ \tau_s$. Given some initial point $\mathbf{x}$, a flow determines the curve $\gamma(t) = \tau_t(\mathbf{x})$, whose image is called the *orbit* of the point $\mathbf{x}$.

An example is the *rotation flow* of the plane, defined by $R_t(x, y) = (\widehat{x}, \widehat{y})$ where the vector has been rotated counterclockwise by angle t. This is given by the rotation matrix: $R_t(\mathbf{v}) = \begin{bmatrix}\cos t & -\sin t\\ \sin t & \cos t\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix} = \begin{bmatrix}\widehat{x}\\ \widehat{y}\end{bmatrix}$. See Fig. 14. Applying R_t represents flowing for the time t.

In fact, given a $\mathcal{C}^2$ vector field, we can always find the corresponding flow. This is the content of the Fundamental Theorem of Ordinary Differential Equations: in $\mathbb{R}^n$, any $\mathcal{C}^2$ vector field is tangent to a unique family of curves, meaning that there exists a unique curve γ through a point $\mathbf{p}$ tangent to the vector field $V : \mathbb{R}^n \to \mathbb{R}^n$, and furthermore, these can be put together as a flow. Conversely, any such family of curves can be differentiated (by finding the tangent vector at each point) to give the vector field. Finding the curves from the vector field is called *integration*, and the curves are called *integral curves* of the vector field. The equation

$$\gamma'(t) = V(\gamma(t)) \text{ with } \gamma(\mathbf{0}) = \mathbf{p}$$

is called a *(vector) differential equation* with *initial condition* $\mathbf{p}$. When written in coordinates, this gives (equivalenly) a *system of n differential equations*. A curve satisfying this is an integral curve of the vector field, and is also called a *solution* of the differential equation with that initial condition. So to *solve* the differential equation means to find the curve!

Remark 5.2. In the above situation, we call the vector field a *velocity vector field* as its value is the tangent vector (velocity) $\mathbf{v}$ of a solution curve $\gamma(t)$, since $V(\gamma(t)) = \gamma'(t) = \mathbf{v}(t)$.

Now instead of a velocity field, vector fields can also depict a *force field*, such as gravity or an electric field.

By a *force field* we mean the following. The differential equation is now a *second-order* vector ODE as it involves the second derivative: for example, $F(\gamma(t)) = m\gamma''(t)$, which expresses Newton's Law $F = m\mathbf{a}$, where $\gamma(t)$ is the position, $\gamma'(t) = \mathbf{v}(t)$ the velocity, $\mathbf{a}(t) = \gamma''(t)$ the acceleration, and $m \geq 0$ is the mass of the object. Here we need *two* initial conditions: initial position $\gamma(0)$ and initial velocity $\gamma'(0) = \mathbf{v}(0)$; our Fundamental Theorem then guarantees that we will again have a unique solution.

Remark 5.3. (On the interpretation of vectors and of vector fields)

There are various possible interpretations of vectors. The two most important are as *movement* (a translation of position of a particle) and *force* (applied to a particle; it is important to not confuse them as these are completely different! For a simple example, consider the six vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{e}, \mathbf{f}$ in the plane defining the vertices of a regular hexagon, so d $= -\mathbf{a}$ and so on. We prove that $\mathbf{a}+\mathbf{b}+\mathbf{c}+\mathbf{d}+\mathbf{e}+\mathbf{f} = \mathbf{0}$ in two different ways, using these interpretations. First, movement: we consider the path $\mathbf{0}, \mathbf{a}, \mathbf{a}+\mathbf{b}, \dots, \mathbf{a}+\mathbf{b}+\mathbf{c}+\mathbf{d}+\mathbf{e}+\mathbf{f}$; this walks along a translated hexagon and returns us to $\mathbf{0}$, so that is the total motion. Second, each vector represents a force applied to a particle located at $\mathbf{0}$. Now they cancel pairwise, giving $\mathbf{0}$ as the resultant force.

The same two interpretations arise for a vector field F. Now the movement interpretation is that the field is tangent to the flow of a fluid, and a particle (perhaps an ant on a leaf!) is being carried along the flow lines. The second interpretation is that a particle is moving according to Newton's law $F = M\mathbf{a}$ in this force field.

Further possible interpretations are for example that F represents a magnetic field, or an area element of a surface as the covector for a two-form. But the first two are certainly the most common and important for our intuition.

It is possible that the force on the object also depends on its velocity; in that case, this is given by a vector field F where $m\gamma''(t) = F(\gamma(t), \gamma'(t))$. This is the case for a charged particle moving in a magnetic field.

The definition of a second-order vector DE in $\mathbb{R}^n$ is just that: we are given $F : \mathcal{U} \to \mathbb{R}^n$ which is $\mathcal{C}^1$ then

$$\gamma''(t) = F(\gamma(t), \gamma'(t)). \tag{25}$$

In the time-varying case it would be

$$\gamma''(t) = F(t, \gamma(t), \gamma'(t)). \tag{26}$$

In the first case, F only depends on posiition so is a vector field on $\mathcal{U} \subseteq \mathbb{R}^n$. In the last case it is a vector field on R^{1+2n}, however with values in $\mathbb{R}^n \subseteq R^{1+2n}$.

In fact, all higher-order vector DEs can be converted into *first*-order vector DEs; if the order is k, we need k times the dimension. Thus for a second-order vector DE in $\mathbb{R}^n$, so with dimension n, to write it as a first-order system we simply include the vector γ' as a new variable, giving a new solution curve $\eta = (\gamma, \gamma')$ in dimension $2n$. Furthermore, *time-dependent* vector fields, so-called *nonstationary* or *nonautonomous* DEs, can be seen in this context by adding one more parameter (time). See Fig. 13. Thus all DEs can be interpreted geometrically, as finding integral curves (and flows) to a velocity vector field.

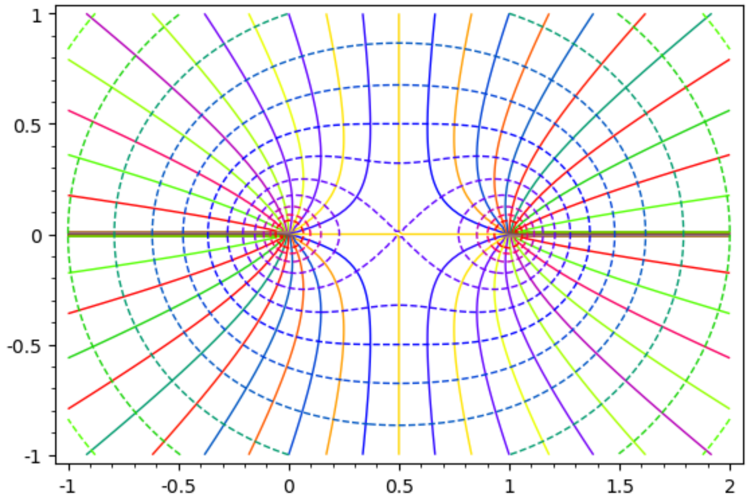


FIGURE 17. Equipotential curves and lines of force for the electrostatic field of two like charges in the plane. For a gravitational potential this would be a topographic map showing either two mountains or two valleys. Note thaat there is a saddle (hyperbolic) point between the two.

The electrostatic fields in Figs. 25, 31, 26, are *gradient* vector fields: depicted are two families of curves, orthogonal to each other; the electrostatic field is tangent to the *lines of flux* between the charges. (Even though they are actually force, not velocity fields, it is useful to picture them as velocity fields). The curves going around the charges are the level curves of the *electrostatic potential* function Φ. Thus $F = \nabla\Phi$ is the electrostatic field. Not all vector fields are gradient fields for some potential; below we find conditions such that this important property holds.

5.2. **The line integral.** Given a vector field F on $\mathbb{R}^n$, the *line integral* of F along γ is

$$\int_\gamma F \cdot \mathrm{d}\gamma \equiv \int_a^b F(\gamma(t)) \cdot \gamma'(t)\mathrm{d}t.$$

A line integral gives a weight at each point of the curve which depends not only on the location $\gamma(t)$ but also on the direction, $\gamma'(t)$ with respect to $F(\gamma(t))$: if these two vectors are aligned it gets a positive weight, if opposed it is negative, and if perpendicular it is zero. If for example F gives a *force field*, then the dot product measures the amount of work needed to move in that direction. Thus an ice skater glides on the ice doing no work, because the plane of the frozen lake is perpendicular to the direction of gravity.

The line integral can also be interpreted as is the integral of the curve with respect to a one-form, the one-form dual to the vector field, just as the dual space F^* is dual to F. We return to this below.

Given a curve $\gamma_1 : [c, d] \to \mathbb{R}^n$, by a *reparametrization* γ_2 of the curve γ_1 we mean $\gamma_2 : [a, b] \to \mathbb{R}^n$ satisfying the following: we have an invertible differentiable

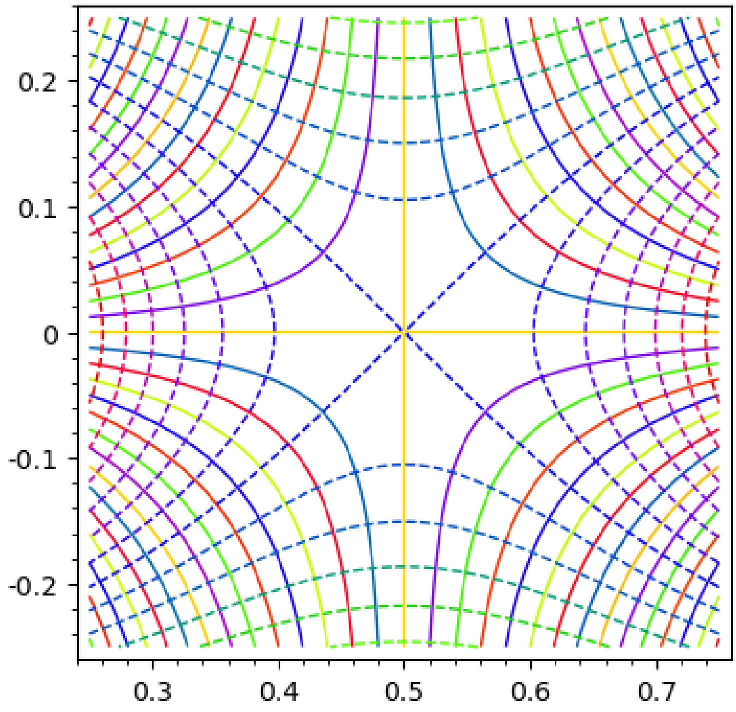


FIGURE 18. Equipotential curves and lines of force for the electrostatic field of two like charges in the plane, showing a closeup view of the saddle point in the center.Note that this approximates the dual hyperbolas of Fig. 21.

function $h : [a, b] \to [c, d]$ with $h'(t) \neq 0$, such that $\gamma_2 = \gamma_1 \circ h$. Note that the curves γ_1 and γ_2 have the same image, and that the tangent vectors are multiples: $\gamma_2'(t) = \gamma_1 \circ h'(t) = \gamma_1'(h(t))h'(t)$. We call this a *positive* or *orientation–preserving parameter change* if $h'(t) > 0$, *negative* or *orientation–reversing* if < 0.

Proposition 5.1.
(i) The line integral is unchanged for an orientation–preserving parametrization. That is,

$$\int_{\gamma_1} F \cdot d\gamma_1 = \int_{\gamma_2} F \cdot d\gamma_2.$$

(ii) For an orientation–reversing parametrization, we change the sign.

Proof. (*i*) Writing $u = h(t)$, we have $\gamma_2(t) = \gamma_1(h(t)) = \gamma_1(u)$. Since $\mathrm{d}u = h'(t)\mathrm{d}t$ then using the Chain Rule:

$$\begin{aligned} \int_{\gamma_2} F \cdot \mathrm{d}\gamma_2 \equiv \int_{t=a}^{t=b} F(\gamma_2(t)) \cdot \gamma_2'(t)\mathrm{d}t = \int_{t=a}^{t=b} F(\gamma_1(h(t))) \cdot (\gamma_1 \circ h)'(t)\mathrm{d}t = \\ \int_{t=a}^{t=b} F(\gamma_1(h(t))) \cdot \gamma_1'(h(t))h'(t)\mathrm{d}t = \int_{u=c}^{u=d} F(\gamma_1(u)) \cdot \gamma_1'(u)\mathrm{d}u = \int_{\gamma_1} F \cdot \mathrm{d}\gamma_1. \end{aligned} \tag{27}$$

(*ii*) For $h' < 0$, then $h(a) = d, h(b) = c$. The calculation is the same, with that change of the limits of integration:

$$\begin{aligned}\int_{\gamma_2} F \cdot \mathrm{d}\gamma_2 \equiv \int_{t=a}^{t=b} F(\gamma_2(t)) \cdot \gamma_2'(t)\mathrm{d}t = \\ \int_{u=d}^{u=c} F(\gamma_1(u)) \cdot \gamma_1'(u)\mathrm{d}u = -\int_{u=c}^{u=d} F(\gamma_1(u)) \cdot \gamma_1'(u)\mathrm{d}u = -\int_{\gamma_1} F \cdot \mathrm{d}\gamma_1.\end{aligned} \tag{28}$$

□

Corollary 5.2. *If* $\gamma : [a, b] \to \mathbb{R}^n$ *is a path, then writing* $\widetilde{\gamma}$ *for the orientation–reversed path, we have* $\int_{\widetilde{\gamma}} F \cdot d\widetilde{\gamma} = -\int_\gamma F \cdot d\gamma$.

Proof. Define $h : [a, b] \to [a, b]$ by $h(b) = a, h(a) = b$, interpolated linearly. Thus,

$$h(t) = -t + (a + b).$$

Then $\widetilde{\gamma}(t) \equiv \gamma \circ h(t)$. The claim follows from the Proposition.

□

There is a second notion of integral along a curve $\gamma : [a, b] \to \mathbb{R}^n$, but where we integrate a function rather than a vector field, so there is no dot product:

Definition 5.2. Given $f : \mathbb{R}^n \to \mathbb{R}$, the *line integral of second type* of f along γ is

$$\int_\gamma f(\mathbf{v})\mathrm{d}s \equiv \int_a^b f(\gamma(t))||\gamma'(t)||\mathrm{d}t.$$

Taking the special case of $f \equiv 1$, we define the *arc length* of γ to be:

$$\int_\gamma \mathrm{d}s = \int_a^b ||\gamma'(t)||\mathrm{d}t.$$

We have already seen this special case above in Part I. (So there is some overlap here with our earlier discussion!)

For an example we already know from first semester Calculus, consider a function $g : [a, b] \to \mathbb{R}$, we consider its graph $\{(x, g(x)) : a \le x \le b\}$. We know from Calculus the arc length of this graph is

$$\int_a^b \sqrt{1 + (g'(t))^2}\mathrm{d}x.$$

We claim that the new formula includes this one: parametrizing the graph as a curve in the plane $\gamma(t) = (t, g(t))$. Then $\gamma'(t) = (1, g'(t))$ so $||\gamma'(t)|| = \sqrt{1 + (g'(t))^2}$, whence indeed $\int_\gamma \mathrm{d}s = \int_a^b \sqrt{1 + (g'(t))^2}\mathrm{d}x$ as claimed.

Proposition 5.3.
(i) The line integral of second type of a function along a curve gives the same value for any change of parametrization, independent of orientation. That is,

$$\int_{\gamma_1} f(\mathbf{v})\, ds = \int_{\gamma_2} f(\mathbf{v})\, ds.$$

Proof. (*i*) Writing $u = h(t)$, then for $\gamma_2 = \gamma_1 \circ h$, we have $\gamma_2(t) = \gamma_1(h(t)) = \gamma_1(u)$. By the Chain Rule, $\gamma_2'(t) = (\gamma_1 \circ h)'(t) = \gamma_1'(h(t))\, h'(t)$. Now $\mathrm{d}u = h'(t)\mathrm{d}t$, therefore we have, assuming first that $h' > 0$,

$$\int_{\gamma_2} f(\mathbf{v})\,\mathrm{d}s \equiv \int_{t=a}^{t=b} f(\gamma_2(t))\,||\gamma_2'(t)||\,\mathrm{d}t = \int_{t=a}^{t=b} f(\gamma_1(h(t)))\,||\gamma_1'(h(t))\,h'(t)||\,\mathrm{d}t$$

Recall the property of norms that $||\alpha\mathbf{v}|| = |\alpha|||v||$.

So assuming first that $h' > 0$, $|h'| = h'$ and this equals

$$= \int_{t=a}^{t=b} f(\gamma_1(u))\,||\gamma_1'(u)||\,h'(t)\,\mathrm{d}t = \int_{u=c}^{u=d} f(\gamma_1(u))\,||\gamma_1'(u)||\,\mathrm{d}u = \int_{\gamma_1} f(\mathbf{v})\,\mathrm{d}s. \tag{29}$$

If instead $h' < 0$, then we have as before

$$\int_{\gamma_2} f(\mathbf{v})\,\mathrm{d}s \equiv \int_{t=a}^{t=b} f(\gamma_2(t))\,||\gamma_2'(t)||\,\mathrm{d}t = \int_{t=a}^{t=b} f(\gamma_1(h(t)))\,||\gamma_1'(h(t))\,h'(t)||\,\mathrm{d}t$$

Since now $h' < 0$, $|h'| = -h'$ and this equals

$$= -\int_{t=a}^{t=b} f(\gamma_1(u))\,||\gamma_1'(u)||\,h'(t)\,\mathrm{d}t$$

But since $h' < 0$, also the limits of integration are switched.

That is, $h(b) = c, h(a) = d$, and this cancels with the sign change, and this equals

$$= -\int_{u=d}^{u=c} f(\gamma_1(u))\,||\gamma_1'(u)||\,\mathrm{d}u = \int_{u=c}^{u=d} f(\gamma_1(u))\,||\gamma_1'(u)||\,\mathrm{d}u \equiv \int_{\gamma_1} f(\mathbf{v})\,\mathrm{d}s. \tag{30}$$

□

For the simplest example of a line integral of second type take the constant function $f(\mathbf{v}) = 1$ for all $\mathbf{v} \in \mathbb{R}^n$. Then

$$\int_\gamma f(\mathbf{v})\mathrm{d}s = \int_\gamma \mathrm{d}s = \int_a^b ||\gamma'(t)||\mathrm{d}t.$$

This is the *arc length* of the curve, see Def. 3.7 above.

We next see how this can be used to give a *unit speed parametrization* of a curve $\gamma : [a, b] \to \mathbb{R}^n$. Set $l(t) = \int_a^t ||\gamma'(r)||\,\mathrm{d}r$, so $l(t)$ is the arclength of γ from time a to time t. Note that $l'(t) = ||\gamma'(t)||$. Therefore, if $||\gamma'(t)|| > 0$ for all t, this is invertible. Our parameter change will be given by $h(t) = l^{-1}(t)$, the inverse function.

Proposition 5.4. *Assume that* $||\gamma'(t)|| > 0$ *for all* t*. Then the reparametrized curve* $\widehat{\gamma} = \gamma \circ h$ *has speed one.*

Proof. Now $1 = (l \circ h)'(t) = l'(h(t))h'(t)$ so $||\widehat{\gamma}'(t)|| = ||(\gamma \circ h)'(t)|| = ||(\gamma'(h(t))h'(t)|| = 1$. □

The function l maps $[a, b]$ to $[0, l(\gamma)]]$ whence the parameter-change function h maps $[0, l(\gamma)]$ to $[a, b]$. We keep t for the variable in $[a, b]$ and define $s = l(t)$, the arc length up to time t, so now s is the variable in $[0, l(\gamma)]$ and $h(s) = t$.

The change of parameter gives $\widehat{\gamma}(s) = (\gamma \circ h)(s) = \gamma(h(s) = \gamma(t)$. This indeed parametrizes the curve $\widehat{\gamma}$ is by arc length s.

Note further that

$$\int_\gamma f(\mathbf{v})\mathrm{d}s \equiv \int_a^b f(\gamma(t))||\gamma'(t)||\,\mathrm{d}t = \int_0^{l(b)} f(\widehat{\gamma}(s))||\widehat{\gamma}'(s)||\mathrm{d}s \equiv \int_{\widehat{\gamma}} f(\mathbf{v})\mathrm{d}s$$

From $s = l(t)$ we have $\mathrm{d}s = l'(t)\mathrm{d}t = ||\gamma'(t)||\mathrm{d}t$. Now we understand rigorously what is $\mathrm{d}s$: it represents the infinitesimal arc length; this helps explain the notation for this type of integral.

Level curves and parametrized curves.

There are two very distinct types of curves we encounter in Vector Calculus: the curves of this section, and the level curves of a function. Next we describe a link between the two:

Proposition 5.5. *Let $G : \mathbb{R}^2 \to \mathbb{R}$ be differentiable and suppose $\gamma : [a, b] \to \mathbb{R}^2$ is a curve which stays in a level curve of G of level c. Then $\gamma'(t)$ is perpendicular to the gradient of G.*

Proof. We have that $G(\gamma(t)) = c$ for alll t. Then by the chain rule, $D(G \circ \gamma)(t) = DG(\gamma(t)D\gamma(t)$. The derivatives here are matrices, with DG a (1×2) matrix (a row vector) and $D\gamma$ a column vector; in vector notation, these are the gradient and tangent vector, so this reads $0 = \frac{\mathrm{d}}{\mathrm{d}t}c = (G \circ \gamma)'(t) = (\nabla G)(\gamma(t)) \cdot \gamma'(t)$. □

Corollary 5.6. *If γ is a curve with $||\gamma(t)|| = c$, then $\gamma' \perp \gamma''$.*

Here is a second, direct proof; see also Corollary 3.6 above:

Proposition 5.7. *For a unit-speed curve γ, then always $\gamma' \perp \gamma''$.*

Proof. $1 = \gamma' \cdot \gamma'$ whence by Leibnitz' Rule,

$$(\gamma' \cdot \gamma')' = 2(\gamma' \cdot \gamma'') = 0.$$

□

This fact allows us to make the following

Definition 5.3. The *curvature* of a twice differentiable curve γ in $\mathbb{R}^n$ at time t is the following. For its unit-speed parametrization $\widehat{\gamma}(s)$ we define the curvature at time s to be $\widehat{\kappa}(s) = ||\widehat{\mathbf{a}}(s)||$; for γ the curvature at time t is $\kappa(t) = (\widehat{\kappa} \circ l)(t) = \kappa(t)$

For example, the curve $\gamma_r(t) = r(\cos t/r. \sin t/r)$ has velocity $\gamma_r'(t) = (-\sin t/r, \cos t/r)$ which has norm one; the acceleration is $\gamma_r''(t) = \frac{1}{r}(\cos(t/r). \sin(t/r)) = -\frac{1}{r^2}\gamma_r(t)$, with norm $\frac{1}{r}$. The curvature is therefore $\frac{1}{r}$. So if the radius of the next curve on the race track is half as much, you will feel twice the force, since by Newton's law, $F = m\mathbf{a}$! This is the physical (and geometric) meaning of the curvature. In differential geometry see p. 59 of [O'N06], For how curvature can be defined for surfaces and manifolds, see e.g. [DC16].

We have seen how a level curve $F = c$ can (sometimes) be filled in by a parametrized curve $\gamma(t)$.

This is for $f : \mathbb{R}^2 \to \mathbb{R}$. For functions on $\mathbb{R}^3$ the notion of level curve is replaced by *level surfaces*. When these can also be parametrized; the exact conditions which permit this are given by the *Implicit Function Theorem*, see §**??** and vector calculus texts.

5.3. Conservative vector fields.

Definition 5.4. Given $A \subseteq \mathbb{R}^n$, a *relatively open* subset of A is $\mathcal{U}_A = \mathcal{U} \cap A$ where $\mathcal{U}$ is open in $\mathbb{R}^n$. (We make a similar definition for any metric space, indeed any topological space). A set A is *connected* iff it is not the union of two disjoint nonempty relatively open subsets. That is, A is not $(A \cap \mathcal{U}) \cup (B \cap \mathcal{V})$ for $\mathcal{U}, \mathcal{V}$ open in $\mathbb{R}^n$.

A related definition is this: a subset $\mathcal{V}$ of $\mathbb{R}^n$ is *pathwise connected* iff given two points A, B in $\mathcal{V}$, there exists a continuous path $\gamma : [a, b] \to \mathcal{V}$ such that $\gamma(a) = A, \gamma(b) = B$.

Lemma 5.8. *A continuous real-valued function which takes two values on a connected set is constant.*

Proof. Suppose the values of the function f on the domain Ω are $c < d$. Then $\mathcal{U} = f^{-1}((-\infty, d))$ and $\mathcal{V} = f^{-1}((c, \infty))$ are both open sets whose union is Ω, so by connectedness of Ω one of the sets must be empty, whence the function is constant. □

Lemma 5.9. *The closed interval $[a, b]$ is connected.*

Proof. Suppose $[a, b] = \mathcal{U} \cup \mathcal{V}$ where these are disjoint open subsets of $\mathbb{R}$. Now one of them contains b. Let $c = \sup\{x \in \mathcal{U} : x < b\}$. Then $\mathcal{U} \subseteq [a, c)$ and $\mathcal{V} \subseteq (c, b]$. But then the union is not the whole interval as it does not contain the point c.

□

Lemma 5.10. *A pathwise connected set is connected.*

Proof. Suppose E is not connected, so it is the union of two disjoint relatively open sets: $E = (F \cap \mathcal{U}) \cup (G \cap \mathcal{V})$ for $\mathcal{U}, \mathcal{V}$ open in $\mathbb{R}^n$. Let $A \in (F \cap \mathcal{U}), B \in (G \cap \mathcal{V})$. By path-connectedness there exits a continuous curve $\gamma : [a, b] \to E$ with $\gamma(a) = A, \gamma(b) = B$. Now since γ is continuous, $\gamma^{-1}(\mathcal{U})$ and $\gamma^{-1}(\mathcal{V})$ are open and disjoint subsets of $[a, b]$. However the interval $[a, b]$ is connected, giving a contradiction. □

Example 14. Here is an example of a connected subset of $\mathbb{R}^2$ which is not pathwise connected:

Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} \sin(1/x) \text{ for } x \neq 0 \\ 0 \text{ for } x = 0 \end{cases} \tag{31}$$

This function is continuous at every point $\neq 0$ but not at 0, because (draw a graph!) the limit at 0 from the right is $\lim_{x \to 0^+} f(x) = [-1, 1]$ (and also from the left the limit is this whole interval), whereas to be continuous we need $\lim_{x \to a} f(x) = f(a)$ for any a.

If we take A to be a point on the left half of the graph, and B on the right half, say $A = f(-2/\pi) = -1, B = f(2/\pi) = 1$, we cannot connect them by a continuous curve γ, as we must then have $\gamma(t) = (t, f(t))$ but then $\lim_{t\to 0^+} \gamma(t) = \{0\} \times [-1, 1] \neq \gamma(0) = (0, 0)$.

On the other hand A is connected: (Sketch of proof!) If there are open sets $\mathcal{U}, \mathcal{V} \subseteq \mathbb{R}^2$ with $A = (\mathcal{U} \cap A) \cup (\mathcal{V} \cap A)$, then one of the sets must contain $A \cap [0, b]$ for some $b > 0$, and one must contain $A \cap [a, 0]$ for some $a > 0$. But these sets are not disjoint, as each contains $\{0\} \times [-1, 1]$.

Definition 5.5. By a *region* Ω in $\mathbb{R}^n$ we shall mean a pathwise connected open set. A vector field F on a region $\Omega \subseteq \mathbb{R}^n$ is *conservative* iff there exists $\varphi : \Omega \to \mathbb{R}$ such that the gradient $\nabla\varphi = F$. Such a function is called a *potential* for F.

Lemma 5.11. *If Ω is connected and φ, ψ are two potentials for F then they differ by a constant.*

Proof.
$$\frac{\partial\varphi}{\partial x} = \frac{\partial\psi}{\partial x} \implies \varphi(x, y) = \psi(x, y) + c(y); \frac{\partial\varphi}{\partial y} = \frac{\partial\psi}{\partial y} \implies \varphi(x, y) = \psi(x, y) + d(x).$$

Subtracting, $c(y) = d(x)$ so this is locally a constant (fixing y and changing x shows that $d(x) = c(y)$ for all x, similarly for $c(y)$) hence by connectedness is constant. □

Proposition 5.12. *If F is conservative and $\gamma : [a, b] \to \Omega$ with $A = \gamma(a), B = \gamma(b)$ then*
$$\int_\gamma F \cdot d\gamma = \varphi(B) - \varphi(A).$$

Proof.
$$\int_\gamma F \cdot \mathrm{d}\gamma \equiv \int_a^b F(\gamma(t)) \cdot \gamma'(t)\mathrm{d}t.$$
And $F(\gamma(t)) = \nabla\varphi(\gamma(t))$ so
$$F(\gamma(t)) \cdot \gamma'(t) = \nabla\varphi(\gamma(t)) \cdot \gamma'(t) = (\varphi \circ \gamma)'(t)$$
thus
$$\int_\gamma F \cdot \mathrm{d}\gamma = \int_a^b (\varphi \circ \gamma)'(t)\mathrm{d}t = \varphi \circ \gamma(t)|_a^b = (\varphi(\gamma(b) - \varphi(\gamma(a)) = \varphi(B) - \varphi(A).$$
□

Remark 5.4. This says that for conservative vector fields, we can find a potential and then evaluate a line integral in a very simple way, just as in one dimension with the Fundamental Theorem of Calculus. Both of these are special cases of Stokes' Theorem; see §5.12 below.

Next we review some equivalent conditions for F to be conservative.

Proposition 5.13. *The following are equivalent, for a vector field on a pathwise connected domain $\Omega \subseteq \mathbb{R}^n$:*
(i) F is conservative, i.e. there exists a potential function for F, that is, $\varphi : \Omega \to \mathbb{R}$ such that $\nabla\varphi = F$.
(ii) The line integral is path-independent.

(iii)For γ a piecewise $\mathcal{C}^1$ path which is closed *i.e. $\gamma(a) = \gamma(b)$, the line integral is* 0.

Proof. $(i) \implies (ii)$: From Proposition 5.12,

$$\int_\gamma F \cdot \mathrm{d}\gamma = \varphi(B) - \varphi(A);$$

thus this value only depends on $\varphi(A)$ and $\varphi(B)$, not on the path taken to get there. Hence if there are two paths γ_1, γ_2 with the same initial and final points A, B, then $\int_{\gamma_1} F \cdot \mathrm{d}\gamma_1 = \int_{\gamma_2} F \cdot \mathrm{d}\gamma_2$.

$(ii) \implies (iii)$: If γ is a closed path, then $\gamma(a) = A = \gamma(b) = B$. Define a second path η with the same initial and final points $A = B$ but with $\eta(t) = A$ for all t. Then $\eta'(t) = 0$ so $\int_\eta F \cdot \mathrm{d}\eta = 0$, whence by (ii) also $\int_\gamma F \cdot \mathrm{d}\gamma = 0$.

Another proof is the following: Given a closed path γ, we choose some $c \in [a, b]$ and define $C = \gamma(c)$. Write γ_1 for the path γ restricted to $[a, c]$ and γ_2 for γ restricted to $[c, b]$. Then by (ii) γ_1 and the time-reversed path $\widetilde{\gamma}_2$ have the same initial and final points, so

$$\int_{\gamma_1} F \cdot \mathrm{d}\gamma_1 = \int_{\widetilde{\gamma}_2} F \cdot \mathrm{d}\widetilde{\gamma}_2.$$

Therefore

$$\int_\gamma F \cdot \mathrm{d}\gamma = \int_a^b F(\gamma(t)) \cdot \gamma'(t)\mathrm{d}t =$$
$$\int_a^c F(\gamma(t)) \cdot \gamma'(t)\mathrm{d}t + \int_c^b F(\gamma(t)) \cdot \gamma'(t)\mathrm{d}t = \int_{\gamma_1} F \cdot \mathrm{d}\gamma_1 + \int_{\gamma_2} F \cdot \mathrm{d}\gamma_2 =$$
$$\int_{\gamma_1} F \cdot \mathrm{d}\gamma_1 - \int_{\widetilde{\gamma}_2} F \cdot \mathrm{d}\widetilde{\gamma}_2 = 0.$$

$(iii) \implies (ii)$: We essentially reverse this last argument. We are given that the integral over a closed path is 0. If there are two paths γ_1, γ_2 with the same initial and final points A, B we are to show that $\int_{\gamma_1} F \cdot \mathrm{d}\gamma_1 = \int_{\gamma_2} F \cdot \mathrm{d}\gamma_2$.

As above, we write $\widetilde{\gamma}_2$ for the time-reversed path. Then $\gamma = \gamma_1 + \widetilde{\gamma}_2$ is a closed loop, so

$$0 = \int_\gamma F \cdot \mathrm{d}\gamma = \int_{\gamma_1} F \cdot \mathrm{d}\gamma_1 + \int_{\widetilde{\gamma}_2} F \cdot \mathrm{d}\widetilde{\gamma}_2 = \int_{\gamma_1} F \cdot \mathrm{d}\gamma_1 - \int_{\gamma_2} F \cdot \mathrm{d}\gamma_2 = 0.$$

$(ii) \implies (i)$: We define a function φ by fixing some point A and choosing $\varphi(A)$ arbitrarily; for example we can take $\varphi(A) = 0$. Then we define the other values as follows. Letting $B \in \Omega$, since the region is path connected there exists a piecewise $\mathcal{C}^1$ path $\gamma : [a, b] \to \Omega$ with $A = \gamma(a), B = \gamma(b)$. We set

$$\varphi(B) = \int_\gamma F \cdot \mathrm{d}\gamma.$$

By (ii), this is well-defined as it does not depend on the path.

We claim that $\nabla\varphi = (\varphi_x, \varphi_y) = F = (F_1, F_2)$, showing the calculation for the case of $F : \mathbb{R}^2 \to \mathbb{R}$. We compute $\frac{\partial\varphi}{\partial x}$ at the point $B = (B_0, B_1)$ and shall show that $\frac{\partial\varphi}{\partial x}|_B = F_1(B)$.

We extend the path γ for $t \geq b$ by defining $\gamma(t) = B + s\mathbf{e}_1$ for $s = t - b$. That is, defining a path η by $\eta(s) = B + s\mathbf{e}_1$, γ is extended by sticking η on the end.

By path-independence of (ii), we still have for any $c \geq b$, with for $C = \gamma(c)$,

$$\varphi(C) = \int_a^c F(\gamma(t)) \cdot \gamma'(t)\mathrm{d}t = \int_a^b F(\gamma(t)) \cdot \gamma'(t)\mathrm{d}t + \int_b^c F(\gamma(t)) \cdot \gamma'(t)\mathrm{d}t.$$

Now since with the substitution $s = t - b$ we have $\mathrm{d}s = \mathrm{d}t$, this equals

$$\varphi(B) + \int_0^{c-b} F(\eta(s)) \cdot \eta'(s)\mathrm{d}s.$$

By definition of the partial derivative,

$$\frac{\partial\varphi}{\partial x}|_B = \lim_{c\to b} \frac{1}{c-b}(\varphi(\gamma(c)) - \varphi(\gamma(b))).$$

Writing $h = c-b$, this is $\lim_{h\to 0}(\frac{1}{h}(\varphi(B+h\mathbf{e}_1) - \varphi(B)) = \lim_{h\to 0}(\frac{1}{h}(\varphi(\eta(h)) - \varphi(\eta(0))$. Now for $t > b$, equivalently $s > 0$, $\gamma'(t) = \eta'(s) = \mathbf{e}_1$. So

$$\lim_{h\to 0}(\frac{1}{h}(\varphi(B + h\mathbf{e}_1) - \varphi(B)) = \lim_{h\to 0}\frac{1}{h}\int_b^{b+h} F(\gamma(t)) \cdot \gamma'(t)\mathrm{d}t = \lim_{h\to 0}\frac{1}{h}\int_b^{b+h} F(\gamma(t)) \cdot \mathbf{e}_1\mathrm{d}t$$

$$= \lim_{h\to 0}\frac{1}{h}\int_b^{b+h} F(\gamma(t)) \cdot (1,0)\mathrm{d}t = \lim_{h\to 0}\frac{1}{h}\int_b^{b+h} F(B_0 + t, B_1) \cdot (1,0)\mathrm{d}t$$

$$= \lim_{h\to 0}\frac{1}{h}\int_0^{h} F_1(B_0 + t, B_1)\mathrm{d}t = F_1(B_0, B_1) = F_1(B)$$

since the partial derivative F_1 is assumed to be continuous. Here we are making use of one version of the Fundamental Theorem of Calculus, (b) below:

Lemma 5.14.
(a) If $f : \mathbb{R} \to \mathbb{R}$ is Riemann integrable, then $G(x) = \int_a^x f(t)dt$ is continuous.
(b) If f is continuous, then G is differentiable, indeed is an antiderivative for f: $G'(x) = f(x)$.

That is,

$$\frac{d}{dx}\int_a^x f(t)dt = \lim_{h\to 0}\frac{G(x+h) - G(x)}{h} = \lim_{h\to 0}\frac{1}{h}\int_x^{x+h} f(t)dt = f(x).$$

Hence every continuous function has an antiderivative.
(c)Any two antiderivatives differ by a constant. For example, this happens in the above case if we replace a by a different number $\widetilde{a}$, as $G(x) = \int_a^x f(t)dt = \int_a^{\widetilde{a}} f(t)dt + \int_{\widetilde{a}}^x f(t)dt = c + \widetilde{G}(x)$.

□

This shows that $\frac{\partial\varphi}{\partial x}|_B = F_1(B)$. In the same way, $\frac{\partial\varphi}{\partial y}|_B = F_2(B)$. So $\nabla\varphi = F$. The same argument works for any dimension, proving the theorem.

□

Next we explain where the term "conservative" comes from: from the conservation of energy in mechanics!

Suppose we have an object (a point mass) and a vector field F of forces acting on this object. This will move according to Newton's law $F = m\mathbf{a}$; here F and also the acceleration $\mathbf{a}$ are vector quantities, while the mass m is a positive scalar. If the position of the object in time is given by the curve $\gamma(t)$, then we write $\mathbf{v}(t) = \gamma(t)$ for the velocity and $\mathbf{a}(t) = \mathbf{v}'(t) = \gamma''(t)$ for the acceleration. So Newton's law states

$$F(\gamma(t)) = m\mathbf{a}(t) = m\gamma''(t).$$

Definition 5.6. *Work* is defined in mechanics to be (force) $\cdot$ (distance). This means that the work done by moving a particle against a force is given by that expression. The continuous-time version of this is given by a line integral.

Precisely, we define the *work done* by moving a particle along a path (a curve) γ in a force field F to be $\int_\gamma F \cdot \mathrm{d}\gamma$.

The *kinetic energy* of the particle is $\frac{1}{2}m||\mathbf{v}||^2$.

Proposition 5.15. *The work done by moving along the path γ in a force field F from time a to time b is the difference in kinetic energies, $E_{kin}(b) - E_{kin}(a)$.*

Proof. The work done by moving along the path γ from time a to time b is

$$\int_\gamma F \cdot \mathrm{d}\gamma = \int_a^b F(\gamma(t)) \cdot \gamma'(t)\mathrm{d}t = m\int_a^b \gamma''(t)) \cdot \gamma'(t)\mathrm{d}t$$

Now by Leibnitz' Rule,

$$\gamma''(t)) \cdot \gamma'(t) = \frac{1}{2}(\gamma'(t) \cdot \gamma'(t))' = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}||\mathbf{v}||^2(t)$$

so our integral is

$$\frac{1}{2}m\int_a^b \frac{\mathrm{d}}{\mathrm{d}t}||\mathbf{v}(t)||^2\mathrm{d}t = \frac{1}{2}m||\mathbf{v}(t)||^2|_a^b = \frac{1}{2}m||\mathbf{v}(b)||^2 - \frac{1}{2}m||\mathbf{v}(a))||^2 = E_{\mathrm{kin}}(b) - E_{\mathrm{kin}}(a).$$

□

This is valid for any force field, conservative or not.

Definition 5.7. Given a conservative vector field F, so with potential function φ, we define the *potential energy* of F to be $E_{\mathrm{pot}} = -\varphi$.

Note that the potential energy function of physics has the opposite sign from the potential function used in mathematics, whose gradient gives the field.

The *total energy* of a particle moving in a force field is the sum of the potential and kinetic energies, $E_{\mathrm{tot}} = E_{\mathrm{pot}} + E_{\mathrm{kin}}$. Note that the potential energy at time a depends only on the position $A = \gamma(a)$, so we write this as $E_{\mathrm{pot}}(A)$, while the kinetic energy depends on time and the path, so we write this as $E_{\mathrm{kin}}(a)$, as for the total energy $E_{\mathrm{tot}}(a)$.

Proposition 5.16. *In a conservative force field F, the work done by moving along the path γ from time a to time b is $\varphi(B) - \varphi(A) = E_{pot}(A) - E_{pot}(B)$.*

Proof. This is just Proposition 5.12 restated in the context of mechanics. □

Theorem 5.17. *If a particle moves according to Newton's law $F = m\mathbf{a}$ in a conservative force field, then the total energy is preserved: $E_{tot}(a) = E_{tot}(b)$.*

Proof. We have shown in Proposition 5.15 that the work done (in any field) is

$$\int_\gamma F \cdot \mathrm{d}\gamma = E_{\mathrm{kin}}(b) - E_{\mathrm{kin}}(a).$$

But in a conservative field, we also have a second expression for this: the work done is

$$\int_\gamma F \cdot \mathrm{d}\gamma = \varphi(B) - \varphi(A) = E_{\mathrm{pot}}(A) - E_{\mathrm{pot}}(B).$$

Thus

$$E_{\mathrm{kin}}(b) - E_{\mathrm{kin}}(a). = E_{\mathrm{pot}}(A) - E_{\mathrm{pot}}(B)$$

so

$$E_{\mathrm{tot}}(a) = E_{\mathrm{kin}}(a) + E_{\mathrm{pot}}(A) = E_{\mathrm{kin}}(b) + E_{\mathrm{pot}}(B) = E_{\mathrm{tot}}(b).$$

□

Remark 5.5. Note that we calculated the line integral $\int_a^b F(\gamma(t)) \cdot \gamma'(t)\mathrm{d}t$ in two different ways, in Proposition 5.12 and Proposition 5.15. For the first we used the existence of a potential to rewrite $F(\gamma(t))$ as $\nabla\varphi(\gamma(t))$ and use the Chain Rule; for the second we used Newton's Law to rewrite F as $m\mathbf{a} = m\gamma''$ and apply Leibnitz' Rule.

It is interesting that this are the same two very different techniques applied to give two different proofs of Corollary 5.6 above.

The curl of a vector field; conservative vector fields.

Definition 5.8. The *curl* of a vector field $F = (P, Q)$ on $\mathbb{R}^2$ is $\mathrm{curl}(F) = (\frac{\partial}{\partial x}Q - \frac{\partial}{\partial y}P)\mathbf{k}$; this is in $\mathbb{R}^3$ not $\mathbb{R}^2$, but we will soon see the reason for this convention. The curl of a vector field $F = (P, Q, R)$ on $\mathbb{R}^3$ is

$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ P & Q & R \end{vmatrix} = \begin{vmatrix} \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ Q & R \end{vmatrix}\mathbf{i} - \begin{vmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial z} \\ P & R \end{vmatrix}\mathbf{j} + \begin{vmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} \\ P & Q \end{vmatrix}\mathbf{k} = \big(R_y - Q_z, P_z - R_x, Q_x - P_y\big).$$

This can also be written as a vector product, since

$$\mathbf{v} \wedge \mathbf{w} = \mathbf{v} \times \mathbf{w} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix},$$

see Part I of these Notes. So one writes

$$\mathrm{curl}(F) = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right) \wedge (P, Q, R),$$

which is often abbreviated as

$$\mathrm{curl}(F) = \nabla \wedge F = \nabla \times F.$$

Note that to define the curl of a vector field in $\mathbb{R}^2$, we have to understand that $\mathbb{R}^2$ is identified with the $x-y$ plane embedded in $\mathbb{R}^3$, with the the curl a vector in $\mathbb{R}^3$ which is perpendicular to this embedded plane.

Remark 5.6. Note that these formulas represent the determinant of a matrix of symbols rather than numbers, so only make sense as formulas. Nevertheless some of the properties carry over from the usual situation of a matrix of numbers. For example, multilinearity of the determinant or linearity of the vector product is reflected in linearity of the curl: given two vector fields on $\mathbb{R}^3$, F, G then $\mathrm{curl}(\alpha F + \beta G) = \alpha\,\mathrm{curl}(F) + \beta\,\mathrm{curl}(G)$.

The formulas for $\mathbb{R}^2$ and $\mathbb{R}^3$ are connected. To understand this, take the fields $F = (P, Q)$ and $\widehat{F} = (\widehat{P}, \widehat{Q}, \widehat{R})$ with $\widehat{R} \equiv 0$ and with $\widehat{P}(x, y, z) = P(x, y)$ and $\widehat{Q}(x, y, z) = Q(x, y)$, whence $\widehat{Q}_z = \widehat{P}_z = 0$ so then $\mathrm{curl}(\widehat{F}) = \big(\widehat{R}_y - \widehat{Q}_z, \widehat{P}_z - \widehat{R}_x, \widehat{Q}_x - \widehat{P}_y\big) = \big(-\widehat{Q}_z, \widehat{P}_z, \widehat{Q}_x - \widehat{P}_y\big) = \big(0, 0, \widehat{Q}_x - \widehat{P}_y\big) = (\widehat{Q}_x - \widehat{P}_y)\mathbf{k}$.

In other words, $\mathrm{curl}(\widehat{F}) = \mathrm{curl}(F)$ in this case.

Proposition 5.18. *If a field F on $\mathbb{R}^2$ is conservative, then the curl is* $\mathbf{0}$.

Proof. This follows immediately from the equality of mixed partials, Lemma 3.31. □

Remark 5.7. The proposition says: $\mathrm{curl}(\mathrm{grad}\varphi) = \mathbf{0}$, that is,

$$\nabla \wedge (\nabla\varphi) = \nabla \times (\nabla\varphi) = \mathbf{0}.$$

In fact, the curl in $\mathbb{R}^3$ can be understood with the help of that in $\mathbb{R}^2$: if $\widehat{F}$ is $\mathbf{0}$ in some other direction $\mathbf{v}$ (replacing the direction $\mathbf{k}$), then the curl is a multiple of $\mathbf{v}$, and is equal to the curl on the plane perpendicular to $\mathbf{v}$.

This will always be the case for a linear vector field, because we can rotate the field so that $\mathbf{v}$ now lines up with $\mathbf{k}$ and we are in the previous situation. If $\widehat{F}$ is not linear, we define:

Definition 5.9. The *linearization* of F at $\mathbf{p}$ is the linear vector field defined by the derivative matrix $F^* = DF_{\mathbf{p}}$.

As we next show, the curl of F at $\mathbf{p}$ is equal to that for its linearization: $\mathrm{curl}(\widehat{F})|_{\mathbf{p}} = \mathrm{curl}(\widehat{F^*})|_{\mathbf{0}}$:

Theorem 5.19. *Let $F = (P, Q, R)$ be a differentiable vector field on $\mathbb{R}^3$, with derivative $DF_{\mathbf{p}}$ at the point $\mathbf{p}$. Let F^* denote the linear vector field defined by the matrix $DF_{\mathbf{p}}$.*

Then curl$(F)|_{\mathbf{p}}$ = curl$(F^*)|_{\mathbf{0}}$, which is constant.

The same holds for $\mathbb{R}^2$.

Proof. For the case of $\mathbb{R}^2$, so $F = (P, Q)$, the derivative matrix is $DF = \begin{bmatrix} P_x & P_y \\ Q_x & Q_y \end{bmatrix}$. The curl is calculated from the off-diagonal entries. So $\mathrm{curl}(F)$ and $\mathrm{curl}(F^*)$ are the same, as they are determined by these entries. More precisely, $DF_{\mathbf{p}}(x, y) = (xP_x + yP_y, xQ_x + yQ_y) = (\widetilde{P}, \widetilde{Q})$ which has curl $((\widetilde{Q})_x - (\widetilde{P})_y)\mathbf{k} = (Q_x - P_y)\mathbf{k}$.

For the (3×3) case, the derivative of a linear map is constant, so for all $\mathbf{x}$, $D(F^*)(\mathbf{x}) = D(F^*)(\mathbf{0}) = DF_{\mathbf{p}} = F^*$.

$$F^* \equiv DF_{\mathbf{p}} = \begin{bmatrix} P_x & P_y & P_z \\ Q_x & Q_y & Q_z \\ R_x & R_y & R_z \end{bmatrix} |_{\mathbf{p}}$$

Write the rows as $\widetilde{P}, \widetilde{Q}, \widetilde{R}$. Then the curl of the linear vector field defined by F^* is

$$\big(\widetilde{R}_y - \widetilde{Q}_z, \widetilde{P}_z - \widetilde{R}_x, \widetilde{Q}_x - \widetilde{P}_y\big) = \big(R_y - Q_z, P_z - R_x, Q_x - P_y\big) = \text{curl}(F)_{\mathbf{p}},$$

proving the claim.

We note that since for any chosen $\mathbf{p}$, $DF_{\mathbf{p}}$ is a linear map, its derivative is constant, equal to that linear map at any point. Thus $\text{curl}(F^*)|_{\mathbf{q}} = \text{curl}(F^*)|_{\mathbf{0}}$, for any $\mathbf{q}$. Another way to say this is that for any linear vector field the curl is the same at all points.

□

The curl is a type of derivative, so it makes sense that it can be calculated from the derivative matrix. The geometrical meaning of curl is an *infinitesimal rotation*: a sphere in $\mathbb{R}^3$ rotates about an axis. (To prove this, in Linear Algebra the Spectral Theorem tells us that a rotation –given by an orentation-preserving orthogonal matrix–has an eigenvector; this gives the axis). The curl measures the infinitesimal rotation of the vector field, and its vector points along that axis, using the right-hand rule to indicate the direction of the vector. Why this is an infinitesimal rotation is explained by the notion of the exponential of a matrix, illustrated in the next example.

See the online text https://activecalculus.org/vector/ for some nice illustrations.

5.4. **Rotations and exponentials; angle as a potential.** First we consider the linear vector field V on $\mathbb{R}^2$ defined by $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. We shall explain how this is tangent to the rotation flow

$$R_t = \begin{bmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{bmatrix},$$

see Fig. 19.

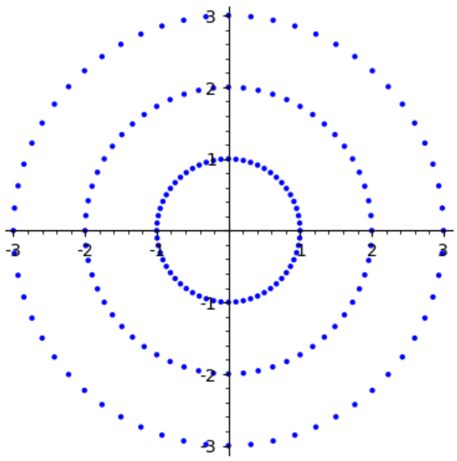
The relationship between the matrices A and R_t is simple, beautiful and profound. We extend the definition of e^x to a square matrix M via the Taylor series

$$\exp(M) = I + M + M^2/2 + \cdots + M^k/k! + \dots$$

It is not hard to show (using comparison and the matrix norm) that this always converges. In particular, for $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, then

$$e^{tA} = \begin{bmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{bmatrix} = R_t$$

gives the rotation flow. To see this, write out the first few terms of the matrix series and use the Taylor series for $\sin, \cos$:



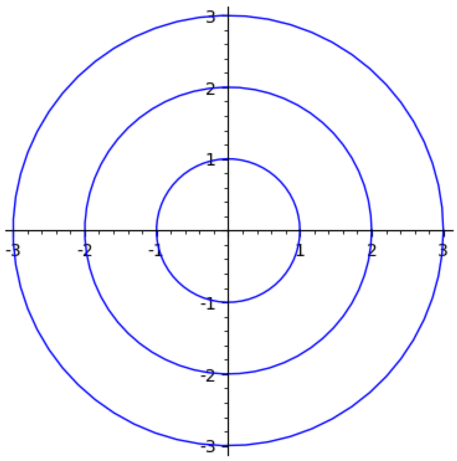


FIGURE 19. Orbits of the rotation flow.

$$\sin(t) = t - t^3/3! + t^5/5! - \dots$$

$$\cos(t) = 1 - t^2/2! + t^4/4! - \dots$$

Conversely, A is the infinitesimal version of this flow, since $\frac{d}{dt}e^{tA} = Ae^{tA}$ exactly as for real functions, hence at $t = 0$ this equals A. Thus, $\frac{d}{dt}|_0 R_t = A$ so A does give the infinitesimal rotation.

A similar equation holds in $\mathbb{R}^3$, which explains why the curl of a vector field does measure the infinitesimal rotation.

This is related to the most basic and most important differential equation: that for exponential growth,

$$f'(t) = f(t)$$

which has as its solution

$$f(t) = Ke^t.$$

The same holds for the vector differential equation $\gamma'(t) = A\gamma(t)$ where $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ and $\gamma = (x, y)$, that is in matrix form,

$$\begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}$$

with initial condition (x_0, y_0) has solution

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$$

The derivative of the linear map $V : \mathbb{R}^2 \to \mathbb{R}^2$ at a point $\mathbf{p}$ is $DV_{\mathbf{p}} = A$ for all $\mathbf{p}$, since the derivative of a linear map is constant, with value equal to the matrix itself. We claim the field V is not conservative. Now writing $V = (P, Q)$, $DV =$

$\begin{bmatrix} P_x & P_y \\ Q_x & Q_y \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, so the curl is $Q_x - P_y = 1 + 1 = 2$. Thus by Proposition 5.18, V is not conservative.

For a second proof, we calculate the line integral $\int_\gamma V \cdot \mathrm{d}\gamma$ for the curve $\gamma(t) = (\cos t, \sin t)$, $t \in [0, 2\pi]$. This is

$$\int_0^{2\pi} V(\gamma(t)) \cdot \gamma'(t)\mathrm{d}t = \int_0^{2\pi} (-\sin t, \cos t) \cdot (-\sin t, \cos t)\mathrm{d}t = 2\pi.$$

But this is a closed loop, hence by (*iii*) of Proposition 5.13 is not conservative.

Next we modify V to a nonlinear vector field F, defined everywhere on the plane except at $\mathbf{0}$.

Thus on $\mathcal{U}$ the open set $\mathbb{R}^2 \setminus (0,0)$ we define

$$F = (P, Q) = \left(\frac{-y}{x^2 + y^2}, \frac{x}{x^2 + y^2} \right)$$

Exercise 5.2. What is $||F(\mathbf{v})||$ for $\mathbf{v} = (x, y)$ in terms of $r = ||\mathbf{v}||$? Calculate the derivative, DF, and use that to verify that $\mathrm{curl}(F) = \mathbf{0}$.

Lemma 5.20. *Verify:*
(i) For $\theta \in [0, +\infty)$ and for $\gamma : [0, \theta] \to \mathcal{U}$ with $\gamma(t) = (\cos t, \sin t)$ then $\int_\gamma F \cdot d\gamma = \theta$.
(ii) For $\theta \in (-\infty, 0]$ then also then $\int_\gamma F \cdot d\gamma = \theta$.

We can use line integrals to measure (more precisely, to *define*!) the number of times a curve in the plane "winds about" a certain point. Here is the definition for the point $\mathbf{0}$:

Definition 5.10. Given a closed curve γ in $\mathbb{R} \setminus \mathbf{0}$, the *winding number* or *index of γ about $\mathbf{0}$* of $I(\gamma; \mathbf{0}) \equiv 1/2\pi \int_\gamma F \cdot \mathrm{d}\gamma$.

Corollary 5.21. *For $\gamma(t) = (\cos(2\pi n t), \sin(2\pi n t))$ with $t \in [0, 1]$, $n \in \mathbb{Z}$ then the winding number of γ about $\mathbf{0}$ is n.*

Exercise 5.3. Let $A = (1, 0)$ and $B = (1, 1)$ and suppose $\gamma : [a, b] \to \mathbb{R} \setminus \mathbf{0}$ with $\gamma(a) = A, \gamma(b) = B$. What are the possible values of $\int_\gamma F \cdot \mathrm{d}\gamma$? Why, precisely?

To define this for a different point $\mathbf{x} \in \mathbb{R}^2$, we would translate F to $F_\mathbf{x} = F - \mathbf{x}$ and set $I(\gamma; \mathbf{x}) \equiv 1/2\pi \int_\gamma F_\mathbf{x} \cdot \mathrm{d}\gamma$.

Remark 5.8. This provides one way of defining the inside and outside of a curve: $\mathbf{x}$ is on the outside iff $I(\gamma; \mathbf{x}) = 0$, otherwise on the inside. (For $\mathbf{x} \in Im(\gamma)$ it is not defined).

Conclusion: Despite the fact that we have $\mathrm{curl}(F) = \mathbf{0}$, this field F cannot be conservative because the integral around the closed loop γ with $\theta = 2\pi$ is $\int_\gamma F \cdot \mathrm{d}\gamma = 2\pi$.

We set $\Omega = \mathbb{R}^2 \setminus \{(x, y) : y = 0, x \geq 0\}$, the plane with the positive part of the x-axis removed. We define the *angle function* $\Theta : \Omega \to (0, 2\pi)$ to be the angle of the point (x, y) measured in the counterclockwise direction from this halfline.

A formula for Θ is given in 34 below. See Fig. 20.

Definition 5.11. Two curves $\gamma, \eta : [a, b] \to \mathbb{R}^m$ are *homotopic* iff there is a continuous function $\Phi : [0, t] \times [a, b] \to \mathbb{R}^m$ such that $\Phi(0, t) = \gamma(t)$ and $\Phi(1, t) = \eta(t)$. If you draw a picture of this you will see that it says that the first curve can be continuously deformed into the second. A curve γ is said to be *homotopic to a point* iff it is homotopic to a constant curve $\eta(t) = \mathbf{p}$ for all t. For an example, the curve $\gamma(t) = (\cos t, \sin t)$ in $\mathbb{R}^2$ is homotopic to a point; however in the domain $\mathcal{U} = \mathbb{R}^2 \setminus \{\mathbf{0}\}$ it is *not*.

A region $\Omega \subseteq \mathbb{R}^2$ is *simply connected* iff it is pathwise connected and has no "holes", meaning every closed curve is homotopic to a point. In the above example, $\mathcal{U} = \mathbb{R}^2 \setminus \{\mathbf{0}\}$ has a "hole" at $\mathbf{0}$.

The basic result is:

Theorem 5.22. *If a region Ω is simply connected, and if curl$(F) = \mathbf{0}$ on Ω, then there exists a potential function φ for F defined on Ω.*

Proof. We proved in Proposition 5.13 that if the domain Ω is pathwise-connected, and the line integral over a closed loop is 0 (or equivalently, path-independent) then there exists a potential function φ. The method of proof was to define φ by integration; the path-independence means this function is well-defined.

We claim that if Ω is simply connected,and the curl is $\mathbf{0}$, then path-independence holds. The reason is that the path integral changes continuously over a continuous homotopy. But...TO DO □

Example 15. We analyze the important specific example of the angle function Θ. This is a potential function for the field F, but only on the restricted, simply connected domain $\mathbb{R}^2$ minus the positive real axis.

What happens at the limit as the angle goes to 2π is quite interesting, explained geometrically by the graph of Θ.

For the angle function Θ example we carry this out directly. The domain of definition of $\cot(\theta) = \cos(\theta)/\sin(\theta)$ is $(0, \pi)$. So:

$$\Theta(x, y) = \begin{cases} \text{arccot}(x/y) & \text{for } y > 0;\ \text{taking values } \theta \in (0, \pi) \\ \text{arccot}(x/y) + \pi & \text{for } y < 0;\ \text{taking values } \theta \in (\pi, 2\pi) \end{cases} \tag{32}$$

We choose the initial point $A = (-1, 0)$ and connect it to $B \in \Omega$ by a path γ in Ω, defining φ by $\varphi(A) = 0$, and

$$\varphi(B) = \int_\gamma F \cdot \mathrm{d}\gamma.$$

This is well-defined since Ω is pathwise connected, so by (*ii*) of Proposition 5.13 it is path-independent.

Lemma 5.23. *We claim that $\varphi(x, y) + \pi = \Theta(x, y)$ for all $(x, y) \in \Omega$.*

Proof. We use the following path to connect A and $B = (x, y)$. We define $\gamma_1(t) = (-1, t)$ for $t \in [0, y]$ and $\gamma_2(t) = (t, y)$ for $t \in [-1, x]$. Note that $\gamma_1' = (0, 1)$, $\gamma_2' = (1, 0)$. We define $\gamma = \gamma_1 + \gamma_2$. This goes vertically up from A to the point $(-1, y)$ and then horizontally over to B.

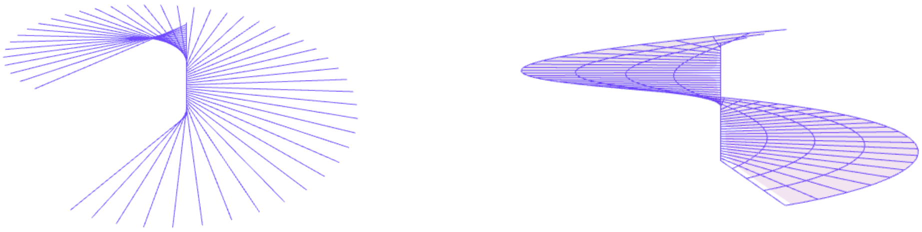


FIGURE 20. The angle function Θ.

We have

$$\varphi(x,y)=\int_{\gamma_1} F\cdot \mathrm{d}\gamma_1+\int_{\gamma_2} F\cdot \mathrm{d}\gamma_2. \tag{33}$$

To evaluate this we need to recall some facts about inverse trigonometric functions. We have $\cot(\theta)=\cos(\theta)/\sin(\theta)=x/y$ so $\operatorname{arccot}(x/y)=\theta$. The domain of definition of cot is $(0,\pi)$.

So we have these formulas for the angle function Θ:

$$\Theta(x,y)=\operatorname{arccot}(x/y),\ \text{taking values } \theta\in(0,\pi),\ \text{so for } y>0,$$

and

$$\Theta(x,y)=\operatorname{arccot}(x/y)+\pi,\ \text{taking values } \theta\in(\pi,2\pi),\ \text{for } y<0$$

Summarizing,

$$\Theta(x,y)=\begin{cases}\operatorname{arccot}(x/y) \text{ for } y>0\\ \operatorname{arccot}(x/y)+\pi \text{ for } \Theta\in(\pi,2\pi), \text{ so for } y<0\\ \pi \text{ for } \Theta=\pi, \text{ for } y=0, x<0\end{cases} \tag{34}$$

Next we evaluate φ from (33):

$$\int_{\gamma_1} F\cdot \mathrm{d}\gamma_1=\int_0^y F(-1,t)\cdot(0,1)\mathrm{d}t=\int_0^y \frac{x}{x^2+y^2}\circ(-1,t)\mathrm{d}t=$$
$$\int_0^y \frac{-1}{1+t^2}\mathrm{d}t=\operatorname{arccot}(y)-\operatorname{arccot}(0)=\operatorname{arccot}(y)-\pi/2.$$

And:

$$\int_{\gamma_2} F\cdot \mathrm{d}\gamma_2=\int_{-1}^x \frac{-y}{x^2+y^2}\circ(t,y)\mathrm{d}t=\int_{-1}^x \frac{-y}{t^2+y^2}\mathrm{d}t$$

Here we use the substitution $u=t/y$, so $\mathrm{d}u=1/y\mathrm{d}t$, $t=uy$, and

$$\frac{-y}{t^2+y^2}=\frac{-y}{(uy)^2+y^2}=\frac{-1}{u^2+1}.$$

Thus we have

$$\int_{-1}^x \frac{-y}{t^2+y^2}\mathrm{d}t=\int_{u=-1/y}^{u=x/y}\frac{-1}{u^2+1}\mathrm{d}u=\operatorname{arccot}(x/y)-\operatorname{arccot}(-1/y).$$

$$\varphi(x,y) = \int_{\gamma_1} F\cdot d\gamma_1 + \int_{\gamma_2} F\cdot d\gamma_2 = \operatorname{arccot}(y) - \operatorname{arccot}(-1/y) + \operatorname{arccot}(x/y) - \pi/2. \quad (35)$$

We claim that the first part of this is locally constant, in fact:

$$\operatorname{arccot}(y) - \operatorname{arccot}(-1/y) = \begin{cases} -\pi/2 \text{ for } y > 0 \\ \pi/2 \text{ for } y < 0 \end{cases} \quad (36)$$

To prove this, we calculate that the following derivative is 0:

$$\frac{d}{dy}\Big(\operatorname{arccot}(y) - \operatorname{arccot}(-1/y)\Big) = \frac{-1}{1+y^2} - -\frac{-1}{1+(\frac{-1}{y})^2}\cdot y^{-2} =$$
$$\frac{-1}{1+y^2} + \frac{1}{y^2+1} = 0$$

To find the constant we evaluate at a single point (actually at two points), where it is easy: at 1 and -1: Now $\cot(\pi/4) = 1, \operatorname{arccot}(1) = \pi/4$; this is for the case $y > 0$, and $\cot(-\pi/4) = -1, \operatorname{arccot}(-1) = 3\pi/4$, for the case $y < 0$, of (36).

So as claimed in (36),

$$\operatorname{arccot}(y) - \operatorname{arccot}(-1/y) = \begin{cases} -\pi/4 - 3\pi/4 = -\pi/2, \text{ for } y = 1 \\ 3\pi/4 - \pi/4 = \pi/2, \text{ for } y = -1 \end{cases}$$

Combining this with (35),

$$\varphi(x,y) = \operatorname{arccot}(x/y) + \begin{cases} -\pi \text{ for } y > 0 \\ 0 \text{ for } y < 0 \end{cases}$$

$$\varphi(x,y) + \pi = \operatorname{arccot}(x/y) + \begin{cases} 0 \text{ for } y > 0 \\ \pi \text{ for } y < 0 \end{cases}$$

From (34),

$$\Theta(x,y) = \operatorname{arccot}(x/y) + \begin{cases} 0 \text{ for } y > 0 \\ \pi \text{ for } y < 0 \end{cases} \quad (37)$$

Lastly, for $y = 0, x < 0$ we have $\varphi + \pi = 0 + \pi = \pi$ and also $\Theta = \pi.$, since $\varphi(-1,0) = 0$ while $\Theta(-1,0) = \pi$.

This proves the Claim for all cases, that

$$\Theta = \varphi + \pi.$$

□

Remark 5.9. To better understand the potential function Θ, draw its level curves; they are rays from the origin, climbing up like a spiral staircase.

Note that for $\gamma(t) = (\cos t, \sin t)$ then

$$\int_0^{2\pi} F(\gamma(t)) \cdot \gamma'(t)\mathrm{d}t = \int_0^{2\pi} (\cos t, \sin t) \cdot (-\sin t, \cos t)\mathrm{d}t = 2\pi$$

and also

$$\lim_{t\to 2\pi} \Theta(\gamma(t)) - \Theta(1) = \lim_{B\to\mathbf{0}} \Theta(B) - \Theta(\mathbf{0}) = 2\pi - 0 = 2\pi$$

so the formula $\int_\gamma F \cdot d\gamma = \varphi(B) - \varphi(A)$ is still valid in the limit; it is also valid if we can somehow allow for a "multi-valued function" as a potential!

See §5.15, Fig 24 below for a different view of this potential: it is related to the electrostatic field of a single charge at the origin.

5.5. **Line integral with respect to a differential form.** We have been studying line integrals,

$$\int_\gamma F \cdot \mathrm{d}\gamma,$$

with this expression defined to equal

$$\int_a^b F(\gamma(t)) \cdot \gamma'(t)\mathrm{d}t.$$

We now introduce a different notation for the line integral. If $F = (P, Q)$ is a vector field on $\mathbb{R}^2$, we write $\int_\gamma P\mathrm{d}x + Q\mathrm{d}y$ and define this to be simply equal to the line integral with respect to F. But exactly what is the meaning of the expression $P\mathrm{d}x + Q\mathrm{d}y$? (Do *not* mistake this for a formula from first-semester Calculus!)

To explain this we recall that given a vector space V, its *dual space* V^* is the set of all *linear functionals* on V, that is all $\lambda : V \to \mathbb{R}$ linear. Note that V^* is itself a vector space; the operations on V^* are defined pointwise as for any collection of functions taking values in a vector space, that is $(\lambda_1 + \lambda_2)(\mathbf{v}) \equiv \lambda_1(\mathbf{v}) + \lambda_2(\mathbf{v})$, and similarly, $(a\lambda)(\mathbf{v}) \equiv a(\lambda(\mathbf{v}))$. We call λ a *dual vector* or a *co-vector*. If we have an inner product $\langle \mathbf{v}, \mathbf{w} \rangle$ on V, then we define an explicit vector $\mathbf{v}^* = \lambda_\mathbf{v} \in V^*$ dual to $\mathbf{v} \in V$ by $\lambda_\mathbf{v}(\mathbf{w}) = \langle \mathbf{v}, \mathbf{w} \rangle$. We write this function as $\langle \mathbf{v}, \cdot \rangle = \lambda_\mathbf{v}(\cdot)$. By bilinearity, V defines linear functionals on V^*, denoted $\langle \cdot, \mathbf{v} \rangle$.

The map $\mathbf{v} \mapsto \langle \mathbf{v}, \cdot \rangle = \mathbf{v}^*$ from V to V^* depends on the choice of the inner product, or equivalently on the choice of a basis, which we define to be orthonormal. Given this choice we can think of a co-vector as simply a vector.

The term *duality* in math refers to any situation where you can switch back and forth; in this case, $\mathbf{v} \mapsto \mathbf{v}^* \mapsto (\mathbf{v}^*)^* = \mathbf{v}$. Thus the *double dual* map is just the identity on V, and $V^{**} = V$.

A *differential one-form* η is a *field of dual vectors*, elements of the dual vector space V^* to V. That is, $\eta : V \to V^*$.

Given a function $\varphi : \mathbb{R}^n \to \mathbb{R}$, we define $\eta = \mathrm{d}\varphi$ to be the one-form dual to the gradient, $\nabla\varphi$. In particular, taking $\varphi(x, y) = x$, then the one-form $\mathrm{d}\varphi = \mathrm{d}x$ is dual to the constant vector field $F(\mathbf{x}) = \mathbf{e}_1$ where $\mathbf{e}_1 = (1, 0)$. So any one-form η can be written as a linear combination:

$$\eta = P\mathrm{d}x + Q\mathrm{d}y.$$

Note that the coefficients depend on the location: they are *functions* $P(x, y)$, $Q(x, y)$.

This one-form is *dual* to the vector field $F = (P, Q)$, and conversely, F is dual to η.

Similarly in $\mathbb{R}^3$ we can express a one-form as

$$\eta = P\mathrm{d}x + Q\mathrm{d}y + R\mathrm{d}z.$$

Again, we then define the line integral with respect to a one-form as equal to its line integral over the associated vector field.

Given a one-form η, we define line integral of a curve γ over η to be simply the line integral of the corresponding vector field F, so

$$\int_\gamma \eta = \int_\gamma P\mathrm{d}x + Q\mathrm{d}y = \int_\gamma F \cdot \mathrm{d}\gamma = \int_a^b F(\gamma(t)) \cdot \gamma'(t)\mathrm{d}t.$$

Thus to calculate a line integral over a one-form, the first step is to write it out as a standard line integral with respect to the dual field $F = (P, Q)$.

A key fact about line integrals is that the *orientation* of γ is important, since for $\gamma : [a, b] \mapsto \mathbb{R}^2$ with opposite curve $\widetilde{\gamma} = -\gamma$, then as we know, $\int_{\widetilde{\gamma}} F \cdot \mathrm{d}\widetilde{\gamma} = -\int_\gamma F \cdot \mathrm{d}\gamma$.

Thus γ is an oriented curve, and not just the point set $\mathrm{Im}(\gamma)$, the image of the curve.

This is the same as the difference between the Riemann integral $\int_{[a,b]} f(x)\mathrm{d}x$ and the integral $\int_a^b f(x)\mathrm{d}x = F(b) - F(a)$ defined from a primitive F, since in the second case $A = [a, b]$ is treated as an *oriented interval* and we have $\int_b^a f(x)\mathrm{d}x = -\int_a^b f(x)\mathrm{d}x$.

These matters become more subtle for double and triple and k-tuple integrals, where V^* is replaced by the set of *alternating* k*-tensors* on $\mathbb{R}^k$, as we explain below.

So far we have only treated one-tensors:

Definition 5.12. Given a vector space V, a *one-tensor* is an element of the dual space V^*. A *differential one-form* η on a vector space V is a function taking values in the one-tensors, so equivalently, $\eta : V \to V^*$. Choice of an inner product associates V to V^*, by sending $\mathbf{v} \mapsto \lambda_\mathbf{v} \in V^*$ with $\lambda_\mathbf{v}(\mathbf{w}) = \langle \mathbf{v}, \mathbf{w} \rangle$. This is an isomorphism, which depends on the choice of inner product.

5.6. **Green's Theorem: Stokes' Theorem in the Plane.** Here we follow the outlines of Guidorizzi's Calculus 3 text: [Gui02]. In my opinion this is (for those who know Portuguese) a good text to teach from, as it is well organized, with correct proofs and good worked-out examples and exercises of a consistent level, but it's not so easy to study from as it is too dry and also because it lacks the beauty of a more advanced and abstract approach. The latter is given in spades in Spivak's beautiful [Spi65] and Guillemin and Pollack's transcendent [GP74]; the approach in these notes is to bridge the way to this very beautiful and powerful more abstract approach while keeping our feet firmly on the ground of simplicity.

Definition 5.13. Given a simple closed C^1 curve γ in $\mathbb{R}^2$, so $\gamma : [a,b] \to \mathbb{R}^2$ with $\gamma(a) = \gamma(b)$, we define a curve on the circle by
$wh\gamma(t) = \gamma(t)/||\gamma(t)||$. This is just the normalized tangent vector, so to see how the tangent vector turns, we look at how $\widehat{\gamma}$ moves along the unit circle.

One can prove (and it makes sense intuitively) that:

Lemma 5.24. *$\widehat{\gamma}$ either goes around once in the clockwise direction or once in the counterclockwise direction.*

We say γ is *oriented positively* if it is a counterclockwise motion, otherwise we say it is *oriented negatively.*

Given a simple closed curve γ in the plane, to state Green's Theorem we need tp be able to talk about its *inside* and *outside*. This enables us to define its orientation as well.

These ideas are made precise by the famous *Jordan Curve Theorem*:

Theorem 5.25. *(Jordan) A continuous simple closed curve γ in $\mathbb{R}^2$ partitions the plane into three connected sets:*
–the interior *of the curve, an open set we call K;*
–the image of γ, a closed set, which is the topological boundary of K, so we call it $\partial K =$ Im(γ), the boundary *of K;*
–the exterior *of γ, the open set which is the complement of $K \cup \partial K$.*

Definition 5.14. Given such a curve, we say it *has positve orientation* iff it goes in the counterclockwise direction as seen from the inside.

Proposition 5.26. *If γ is oriented positively and piecewise $\mathcal{C}^1$, then the interior region K is to the left of the tangent vector $\gamma'(t)$ for all t where $\gamma'(t)$ exists and is nonzero.*

Unfortunately, we will not prove any of these beautiful results here, as good proofs require a more advanced perspective, bringing in ideas from algebraic or differential topology; see [Arm83], and as they are clear intuitively by sketching a few pictures. These ideas also are needed in Complex Analysis. There is a nice treatment relating this to line integrals in the third edition of Marsden-Hoffman: [MH98].

Theorem 5.27. *(Green's Theorem) Let γ be a simple closed positively oriented curve in $\mathbb{R}^2$, piecewise $\mathcal{C}^1$, with non-empty interior. Write K for the closure of the interior of γ. Let $F = (P,Q)$ be a $\mathcal{C}^1$ vector field defined on some open set $\mathcal{U} \supseteq K$.*

Then
$$\int_\gamma F \cdot d\gamma = \int\int_K curl(F) \cdot \mathbf{k}\, dx\, dy.$$
equivalently,
$$\int_\gamma P\,dx + Q\,dy = \int\int_K (Q_x - P_y)\, dx\, dy.$$

The proof of Green's Theorem will be given in stages:

Proof. **Proof for rectangle:** Let $K = [a,b] \times [c,d]$. Write $A = (a,c), B = (b,c), C = [b,d), D = (a,d)$. Let $\gamma = \gamma_1 + \cdots + \gamma_4$ be unit-speed boundary curves traversing the segments in a counterclockwise direction, γ_1 from A to B and so on. Thus $\gamma_1(t) = A + t(1,0) = (t,c)$ for $t \in [a,b]$, so $\gamma_1' = (1,0)$. We have

$$\int_{\gamma_1} P\mathrm{d}x + Q\mathrm{d}y = \int_a^b P(t,c)\mathrm{d}t$$

and similarly for the other cases, so

$$\int_\gamma P\mathrm{d}x + Q\mathrm{d}y = \int_a^b P(t,c)\mathrm{d}t + -\int_a^b P(t,d)\mathrm{d}t + \int_c^d Q(b,t)\mathrm{d}t - \int_c^d Q(a,t)\mathrm{d}t =$$
$$\int_a^b P(t,c) - P(t,d)\mathrm{d}t + \int_c^d Q(b,t) - Q(a,t)\mathrm{d}t.$$

On the other hand,

$$\int_K (Q_x - P_y)\mathrm{d}x\mathrm{d}y = \int_c^d \left(\int_a^b \frac{\partial Q}{\partial x}\mathrm{d}x\right)\mathrm{d}y - \int_a^b \left(\int_c^d \frac{\partial P}{\partial y}\mathrm{d}y\right)\mathrm{d}x =$$
$$\int_c^d Q(b,y) - Q(a,y)\mathrm{d}y - \int_a^b P(x,d) - P(x,c)\mathrm{d}x =$$
$$\int_c^d Q(b,t) - Q(a,t)\mathrm{d}t - \int_a^b P(t,d) - P(t,c)\mathrm{d}t$$

which is exactly what we had before!

Proof for right triangle: We take for K the triangle with corners $A = (0,0)$, $B = (1,0)$, $C = (1,1)$ and with boundary curve $\gamma_1 + \gamma_2 - \gamma_3$ with $\gamma_1(t) = (t,0)$, $\gamma_2(t) = (1,t)$ and $\gamma_3(t) = (t,t)$, all for $t \in [0,1]$.. (Here $-\gamma_3$ means the opposite, i.e. orientation–reversed, curve.) Thus $\gamma_1' = (1,0), \gamma_2' = (0,1)$ and $-\gamma_3' = (1,1)$.

We have for $F = (P,Q)$,

$$\int_\gamma F \cdot \mathrm{d}\gamma = \int_\gamma P\mathrm{d}x + Q\mathrm{d}y = \int_0^1 P(t,0) + Q(1,t)\mathrm{d}t - \int_0^1 P(t,t) + Q(t,t)\mathrm{d}t.$$

On the other hand,

$$\int\int_K Q_x - P_y\mathrm{d}x\mathrm{d}y = \int_{y=0}^{y=1} \left(\int_{x=y}^{x=1} \frac{\partial Q}{\partial x}\mathrm{d}x\right)\mathrm{d}y - \int_{x=0}^{x=1} \left(\int_{y=0}^{y=x} \frac{\partial P}{\partial y}\mathrm{d}y\right)\mathrm{d}x =$$
$$\int_{y=0}^{y=1} Q(x,y)|_{x=y}^{x=1}\mathrm{d}y - \int_{x=0}^{x=1} P(x,y)|_{y=0}^{y=x}\mathrm{d}x =$$
$$\int_{y=0}^{y=1} Q(1,y) - Q(y,y)\mathrm{d}y - \int_{x=0}^{x=1} P(x,x) - P(x,0)\mathrm{d}x$$

which equals the line integral! □

Proof for right triangle with one curvy side:

Next we consider a topological triangle with vertices at $A = (a,c)$, $B = (b,c)$, $C = (b,d)$ and with boundary curve $\gamma_1 + \gamma_2 - \gamma_3$ with $\gamma_1(t) = (t,c)$ for $t \in [a,b]$; $\gamma_2(t) = (b,t)$ for $t \in [c,d]$, and $-\gamma_3$ where $\gamma_3(t) = (t, f(t))$ for $t \in [a,b]$, $f(a) = c$ and $f(b) = d$.

We assume that f is invertible, with inverse g.

We have for $F = (P, Q)$:

$$\int_\gamma F \cdot \mathrm{d}\gamma = \int_\gamma P\mathrm{d}x + Q\mathrm{d}y = \int_a^b P(t, c)\mathrm{d}t + \int_c^d Q(b, t)\mathrm{d}t - \int_a^b (P, Q)(\gamma_3(t)) \cdot (1, f'(t))\mathrm{d}t$$

Here

$$\int_a^b (P, Q)(\gamma_3(t)) \cdot (1, f'(t))\mathrm{d}t = \int_a^b P(t, f(t)) + Q(t, f(t))f'(t)\mathrm{d}t$$

so the total is

$$\int_a^b P(t, c)\mathrm{d}t + \int_c^d Q(b, t)\mathrm{d}t - \int_a^b P(t, f(t)) - \int_a^b Q(t, f(t))f'(t)\mathrm{d}t.$$

On the other hand,

$$\int\int_K Q_x - P_y \,\mathrm{d}x\mathrm{d}y = \int_{y=c}^{y=d} \left(\int_{x=g(y)}^{x=b} \frac{\partial Q}{\partial x}\mathrm{d}x \right)\mathrm{d}y - \int_{x=a}^{x=b} \left(\int_{y=c}^{y=f(x)} \frac{\partial P}{\partial y}\mathrm{d}y \right)\mathrm{d}x =$$

$$\int_{y=c}^{y=d} Q(b, y) - Q(g(y), y)\mathrm{d}y - \int_{x=a}^{x=b} P(x, f(x)) - P(x, c)\mathrm{d}x =$$

$$\int_{x=a}^{x=b} P(x, c)\mathrm{d}x + \int_{y=c}^{y=d} Q(b, y)\mathrm{d}y - \int_{x=a}^{x=b} P(x, f(x))\mathrm{d}x - \int_{y=c}^{y=d} Q(g(y), y)\mathrm{d}y$$

We are almost done. Note that each expression has four terms, and the first three of them agree, just changing the variable of integration from time t to the spatial coordinates x and y. It remains to check the last term. This is a substitution, making use of the inverse function: writing $s = f(t)$, so $t = g(s)$, then $\mathrm{d}s = f'(t)\mathrm{d}t$ whence indeed

$$\int_a^b Q(t, f(t))f'(t)\mathrm{d}t = \int_{s=c}^{s=d} Q(g(s), s)\mathrm{d}s = \int_{y=c}^{y=d} Q(g(y), y)\mathrm{d}y$$

completing the proof.

□

Proof for more complicated regions.

Once we have these special cases we can build up to the general statement of Green's Theorem as follows. First we consider other cases of an open region K with a simple closed piecewise-$\mathcal{C}^1$ boundary curve γ. Using straigt lines, we cut K into pieces of the above forms and add up the results. The key point is that the pieces have nonintersecting interiors, and meet on their boundaries in curves with opposite orientation. The double integrals add as this boundary has content zero so those add; on the line integral side of the equation, the question is why do the boundary intersections always meet in curves with opposite orientation? But this is easy to justify: we prove this by induction on the number of pieces, reducing to two regions. Their boundaries meet in curves with opposite orientations because each is counter-clockwise as seen from its own interior, hence opposite as seen from the other region.

The next step is to consider two disjoint simple closed piecewise-$\mathcal{C}^1$ boundary curves γ_1, γ_2 with regions K_1, K_2. If these regions are disjoint, we simply define the boundary of the union $k_1 \cup K_2$ to be γ_1 together with γ_2, which we write as $\gamma_1 + \gamma_2$. The result clearly hols for this case also. Next consider the case where γ_2 is inside of K_1. Then we consider the region $K = K_1 \setminus (K_2 \cup \text{Im}(\gamma_2)$. For example, if the curves are concentric circles, then K is called an *annulus*: a disk with a hole removed from it. We defined the boundary curve to be $\gamma = \gamma_1 - \gamma_2$. That is, the outer curve γ_1 is oriented positively, while the inner curve is oriented negatively.

Note that the resulting boundary curve γ has the property that as we traverse the curve, the region K always occurs on the *left-hand side.*

It is then easy to show by subtracting the two results for γ_1, γ_2 that Green's Theorem still holds.

Note that such a region is now not simply connected.

We do similarly for a disk with k holes removed.

A more formal proof uses the notion of *chains* as developed in [Spi65] or [GP74].

Exercise 5.4. Consider the field

$$F = (P, Q) = \left(\frac{-y}{x^2 + y^2}, \frac{x}{x^2 + y^2} \right)$$

of Exercise 5.2, for the region with two boundary circles of radius 1 and 2. What does Green's Theorem say in this case?

Remark 5.10. The proof of Green's Theorem for rectangular regions given above may remind the reader of the proof we gave above for the equality of mixed partials Lemma 3.31. We next see the exact connection between the two arguments, by showing how the equality of mixed partials follows as a corollary of Green's Theorem.

Given a vector field $F = (P, Q)$ and the rectangle from the proof above, $R = [a, b] \times [c, d]$, we parametrize the boundary of R by a counterclockwise curve γ and calculate the line integral $\int_\gamma F \cdot \mathrm{d}\gamma$. The corners of R are A, B, C, D with $A = (a, c), B = (b, c), C = (b, d)$ and $D = (a, d)$. We have $\gamma = \gamma_{AB} + \gamma_{BC} + \gamma_{CD} + \gamma_{DA}$ where these are the unit-speed paths; we use the inverse paths for the last two. Thus

$$\begin{aligned}
\gamma_{AB}(t) &= (t, c) \text{ for } t \in [a, b]; \gamma'_{AB} = (1, 0) \\
\gamma_{BC}(t) &= (b, t) \text{ for } t \in [c, d]; \gamma'_{BC} = (0, 1) \\
-\gamma_{CD}(t) &= (t, d) \text{ for } t \in [a, b]; -\gamma'_{CD} = (1, 0) \\
-\gamma_{DA}(t) &= (a, t) \text{ for } t \in [c, d]; -\gamma'_{DA} = (0, 1)
\end{aligned}$$

Then

$$\begin{aligned}
\int_{\gamma_{AB}} F \cdot \mathrm{d}\gamma_{AB} &= \int_a^b (P, Q)(\gamma_{AB}) \cdot (1, 0)\mathrm{d}t = \int_a^b P(t, c)\mathrm{d}t \\
\int_{\gamma_{BC}} F \cdot \mathrm{d}\gamma_{BC} &= \int_c^d (P, Q)(\gamma_{BC}) \cdot (0, 1)\mathrm{d}t = \int_c^d Q(b.t)\mathrm{d}t \\
\int_{\gamma_{CD}} F \cdot \mathrm{d}\gamma_{CD} &= \int_a^b (P, Q)(\gamma_{CD}) \cdot (1, 0)\mathrm{d}t = \int_a^b P(t, d)\mathrm{d}t \\
\int_{\gamma_{DA}} F \cdot \mathrm{d}\gamma_{DA} &= \int_c^d (P, Q)(\gamma_{DA}) \cdot (0, 1)\mathrm{d}t = \int_c^d Q(a, t)\mathrm{d}t
\end{aligned}$$

So far this is true for any vector field. We now assume F is conservative, so there exists φ with $F = \nabla\varphi$, so $F = (P, Q)$ where $P(x, y) = \frac{\partial}{\partial x}\varphi(x, y)$ and $Q(x, y) = \frac{\partial}{\partial y}\varphi(x, y)$. So

$$\int_{\gamma_{AB}} F \cdot \mathrm{d}\gamma_{AB} = \int_a^b P(t, c)\mathrm{d}t = \int_a^b \frac{\partial}{\partial x}\varphi(t, c)\mathrm{d}t = \varphi(t, c)|_a^b$$

and we have:

$$\begin{aligned} \int_{\gamma_{AB}} F \cdot \mathrm{d}\gamma_{AB} &= \varphi(t, c)|_a^b = \varphi(b, c) - \varphi(a, c) \\ \int_{\gamma_{BC}} F \cdot \mathrm{d}\gamma_{BC} &= \varphi(b, t)|_c^d = \varphi(b, d) - \varphi(b, c) \\ -\int_{\gamma_{CD}} F \cdot \mathrm{d}\gamma_{CD} &= \varphi(t, c)|_a^b = \varphi(b, d) - \varphi(a, d) \\ -\int_{\gamma_{CD}} F \cdot \mathrm{d}\gamma_{CD} &= \varphi(t, c)|_c^d = \varphi(a, d) - \varphi(a, c) \end{aligned}$$

Thus

$$\begin{aligned} \int_\gamma F \cdot \mathrm{d}\gamma = & \quad \big(\varphi(b, c) - \varphi(a, c)\big) + \big(\varphi(b, d) - \varphi(b, c)\big) \\ & + \big(\varphi(a, d) - \varphi(b, d)\big) + \big(\varphi(a, c) - \varphi(a, d)\big) = 0 \end{aligned}$$

Note that this statement

$$\int_\gamma F \cdot \mathrm{d}\gamma = 0.$$

is equivalent to that

$$\int_{\gamma_{AB}+\gamma_{CD}} F\mathrm{d}\gamma_1 = -\int_{\gamma_{BC}+\gamma_{DA}} F\mathrm{d}\gamma_2,$$

as we traverse the sides of R in a different order. And this was exactly the concluding step in the proof of Lemma 3.31.

We have proved that if F is conservative, then the integral around any rectangular loop is 0. This prrof has *not* used the equality of mixed partials.

But we can prove the equality of mixed partials from this fact as a corollary of Green's Theorem. Green's Theorem states that

$$\int\int_R \mathrm{curl}(F)\,\mathrm{d}x\mathrm{d}y = \int_\gamma F \cdot \mathrm{d}\gamma$$

for $\gamma = \partial R$ as above.

Since we have proved that the line integral around ∂B is 0 for each rectangle, Green's Theorm tells us that $\mathrm{curl}(F) = \mathbf{0}$.

But $\mathrm{curl}(F) = (\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y})\mathbf{k}$ and hence the mixed partials are equal.

5.7. **The Divergence Theorem in the plane.**

Definition 5.15. Let $F = (P, Q)$ be a $\mathcal{C}^1$ vector field in $\mathbb{R}^2$. The *divergence* of F is defined to be:

$$\mathrm{div}(F) = P_x + Q_y.$$

We shall use the notation: given a vector $\mathbf{v} = (a, b) \in \mathbb{R}^2$, then $\mathbf{v}^* = (b, -a)$ and $\widetilde{\mathbf{v}}^* = (-b, a)$.

For the particular case of $F = (P, Q)$ we write G for $\widetilde{F}^* = (-Q, P)$.

Theorem 5.28. *Let $F = (P, Q)$ be a $\mathcal{C}^1$ vector field in the plane, and let γ be a piecewise $\mathcal{C}^1$, positively oriented simple closed curve, with interior region K. We define $\mathbf{n} = \gamma'^*/||\gamma'^*|| = \gamma'^*/||\gamma'||$; this is the* outward normal vector *of γ.*

Then

$$\int_\gamma F \cdot \mathbf{n} ds = \int\int_K div(F)\, dx\, dy.$$

The same holds more generally for a finite collection of disjoint such regions $K_1, \dots K_n$ with boundaries $\gamma_1, \dots, \gamma_n$ and then writing $K = \cup K_n$ and $\gamma = \gamma_1 + \cdots + \gamma_n$.

Proof. We place the two statements side-by-side, for γ the boundary curve of K, one for the field F and the other for $G = \widetilde{F}^*$:

Green's Theorem:

$$\int_\gamma G \cdot \mathrm{d}\gamma = \int\int_K \mathrm{curl}(G) \cdot \mathbf{k}\, \mathrm{d}x\mathrm{d}y$$

Divergence Theorem:

$$\int_\gamma F \cdot \mathbf{n}\, \mathrm{d}s = \int\int_K \mathrm{div}(F)\, \mathrm{d}x\mathrm{d}y.$$

Note here that $\mathrm{curl}(G) \cdot \mathbf{k} = \mathrm{div}(F)$, so once the prove the two different types of line integrals are equal, the theorem is proved!

For $\gamma(t) = (x(t), y(t))$, then $\gamma'(t) = (x'(t), y'(t))$, and $\mathbf{n} = \gamma'^*/||\gamma'|| = (y', -x')/||\gamma'|| = (y', -x')/||(y', -x')||$.

Recall (Def. 5.2) that the line integral of second type of a function $f : \mathbb{R}^2 \to \mathbb{R}$ over $\gamma : [a, b] \to \mathbb{R}^2$ is defined to be

$$\int_\gamma f(\mathbf{v})\mathrm{d}s \equiv \int_a^b f(\gamma(t))\, ||\gamma'(t)||\mathrm{d}t$$

where $\mathrm{d}s$ is the element of arclength, $\mathrm{d}s = ||\gamma'(t)||\mathrm{d}t$. As we showed in Proposition 5.3, this value is independent of prametrization.z Now for this to make sense, it is enough for the function f to be defined on the image of γ, not necessarily on all of $\mathbb{R}^2$. So when we write the formula

$$\int_\gamma F \cdot \mathbf{n}\, \mathrm{d}s$$

what we mean by this is the line integral of second type of the function f over γ, where f is defined on the image of γ by

$$f(\gamma(t)) = F(\gamma(t)) \cdot \mathbf{n}(t).$$

Thus
$$\int_\gamma F\cdot \mathbf{n}\,\mathrm{d}s \equiv \int_\gamma f(\mathbf{v})\mathrm{d}s \equiv \int_a^b f(\gamma(t))||\gamma'(t)||\mathrm{d}t.$$

Now writing in components $F=(P,Q)$, we have

$$\int_\gamma F\cdot \mathbf{n}\,\mathrm{d}s = \int_a^b F(\gamma(t))\cdot \mathbf{n}(t)\;||\gamma'(t)||\mathrm{d}t = \int_a^b F(\gamma(t)))\cdot(y',-x')/||\gamma'(t)||\;\;||\gamma'(t)||\mathbb{R}dt$$
$$= \int_a^b (P,Q)(\gamma(t))\cdot(y',-x')\,\mathrm{d}t = \int_a^b (-Q,P)(\gamma(t))\cdot(x',y')\,\mathrm{d}t = \int_a^b G(\gamma(t))\cdot\gamma'(t)\,\mathrm{d}t$$
$$= \int_\gamma G\cdot \mathrm{d}\gamma = \int\int_K \mathrm{curl}(G)\cdot\mathbf{k}\,\mathrm{d}x\mathrm{d}y = \int\int_K \mathrm{div}(F)\,\mathrm{d}x\mathrm{d}y$$

proving the Theorem.

□

Remark 5.11. An explanation is that F is lined up with $\mathbf{n}$, thus producing positive divergence, iff $\widetilde{F}^*$ is lined up with γ', thus producing positive curl. The reason for using $\widetilde{F}^*$ rather than F^* is so the sign matches; the key point is that for $\mathbf{v}=(a,b)$ and $\mathbf{w}=(c,d)$, then $\mathbf{v}^*=(b,-a)$ and $\widetilde{\mathbf{w}}^*=(-c,d)$, and $\mathbf{v}\cdot\mathbf{w}^*=\mathbf{w}\cdot\widetilde{\mathbf{v}}^*$. So $\alpha(\mathbf{v},\mathbf{w})\equiv \mathbf{v}\cdot\mathbf{w}^*$ defines an alternating form; indeed, it equals $\det(\mathbf{v},\mathbf{w})$! See Proposition **??** ff. regarding two-tensors.

Using this notation, the last part of the proof can be summarized as:
$$\int_\gamma F\cdot\mathbf{n}\,\mathrm{d}s = \int_a^b F(\gamma(t))\cdot\gamma'^*(t)\mathrm{d}t = \int_a^b \widetilde{F}^*(\gamma(t))\cdot\gamma'(t)\mathrm{d}t =$$
$$\int\int_K \mathrm{curl}(\widetilde{F}^*)\cdot\mathbf{k}\,\mathrm{d}x\mathrm{d}y = \int\int_K \mathrm{div}(F)\,\mathrm{d}x\mathrm{d}y.$$

See p. 79 of [War71] regarding the star operator.

5.8. **Surface area and the "determinant" of a rectangular matrix.** To define the surface area, and more generally, k-dimensional volume in $\mathbb{R}^n$ for $k\le n$, we need an interesting bit of linear algebra.

We recall from above the equivalent definitions of the determinant, proved in Theorem 3.21.
These were first the standard
Algebraic Definition; Note that this is only defined for square matrices.
We then gave this
Geometric Definition: Let M be an $(n\times n)$ real matrix. Then
$$\det M = (\pm 1)(\text{factor of change of volume})$$
where we take $+1$ if M preserves orientation, -1 if that is reversed. (Here this is *n-dimensional volume* and so is length, area in dimensions $1,2$).

We next discussed the vector product in $\mathbb{R}^3$, presenting three definitions. The first two were:
Algebraic definition of $\mathbf{v}\wedge\mathbf{w}$*, via the symbolic "determinant" formula*;

Geometric definition of $\mathbf{v} \wedge \mathbf{w}$.

Letting $P(\mathbf{v}, \mathbf{w})$ denote the parallelogram spanned by $\mathbf{v}, \mathbf{w} \in \mathbb{R}^3$, that is, $P(\mathbf{v}, \mathbf{w}) \equiv \{(a\mathbf{v} + b\mathbf{w} : a, b \in [0,1]\}$, then from the geometric definition, the norm of the vector product $||\mathbf{v} \wedge \mathbf{w}||$ equals the area of $P(\mathbf{v}, \mathbf{w})$.

Now this parallelogram $P(\mathbf{v}, \mathbf{w})$ is the image of the unit square $I \times I = [0,1] \times [0,1]$ in $\mathbb{R}^2$ by the linear map

$$A = \begin{bmatrix} v_1 & w_1 \\ v_2 & w_2 \\ v_3 & w_3 \end{bmatrix}$$

This suggests that we can turn the above definition around and give an analogue of the geometric definition of the determinant of the *rectangular* matrix A: it is to be the area of this image parallelogram.

In other words, although the algebraic definition of determinant does *not* extend to rectangular matrices, the geometric definition *does*, and more generally for a k-parallelopiped P in $\mathbb{R}^n$, analogous to this simplest case of 2-parallelograms in $\mathbb{R}^3$. The tantalizing task is then to find an algebraic formula for this geometric definition in general, which must of course include the usual $(n \times n)$ case.

An answer comes from the following formula for k-dimensional volume in $\mathbb{R}^n$, see [HH15] p. 526. Noting that AA^t is a square matrix, Hubbard proves the volume of a k-parallelopiped P in $\mathbb{R}^n$ equals:

$$\text{vol}(P) = \sqrt{\det(A^t A)}.$$

The *Gram matrix* is $A^t A$; this is useful in Linear Algebra. The name comes from Jorgen Pedersen Gram, famous for many things including the Gram-Schmidt orthogonalization procedure; see Wikipedia. The *Gram determinant* is the determinant of the Gram matrix, so Hubbard's formula is the square root of this. See [**?**] p. 191. We prefer Hubbard's presentation to Courant-John, in part because we find the matrix form both easier to work with and to understand.

We present three proofs of the volume formula, first for the simplest cases:

Lemma 5.29.
(i)For A a (2×1) matrix $\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ then the length of the image of the unit line segment $I = \{(0,t) : t \in [0,1]\}$ is $\sqrt{det(A^t A)}$.
(ii)For A the (3×2) matrix with columns $\mathbf{v}, \mathbf{w}$ as above, then

$$||\mathbf{v} \wedge \mathbf{w}|| = \sqrt{det(A^t A)}.$$

Proof. For (i) we have $A^t A = \begin{bmatrix} v_1 & v_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} v_1^2 + v_2^2 \end{bmatrix}$.

For (ii),

$$A^t A = \begin{bmatrix} \longrightarrow & \mathbf{v} & \longrightarrow \\ \longrightarrow & \mathbf{w} & \longrightarrow \end{bmatrix} \begin{bmatrix} \downarrow & \downarrow \\ \mathbf{v} & \mathbf{w} \\ \downarrow & \downarrow \end{bmatrix} = \begin{bmatrix} v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{bmatrix} \begin{bmatrix} v_1 & w_1 \\ v_2 & w_2 \\ v_3 & w_3 \end{bmatrix} = \begin{bmatrix} \mathbf{v} \cdot \mathbf{v} & \mathbf{v} \cdot \mathbf{w} \\ \mathbf{w} \cdot \mathbf{v} & \mathbf{w} \cdot \mathbf{w} \end{bmatrix}$$

so $\det A^tA = ||\mathbf{v}||^2||\mathbf{w}||^2 - 2\mathbf{v}\cdot\mathbf{w}$ while we know from Corollary 3.22 that the area of the parallelogram $P(\mathbf{v},\mathbf{w}) \subset \mathbb{R}^3$ satisfies

$$(\text{area})^2 = (\mathbf{v}\cdot\mathbf{v})(\mathbf{w}\cdot\mathbf{w}) - (\mathbf{v}\cdot\mathbf{w})^2.$$

Note that the formula

$$\text{vol}(P)^2 = \det(A^tA)$$

is not only completely general but is much easier to remember! □

Following Hubbard, we shall extend the above ideas to $(n\times k)$ matrices. For this we introduce the follwing notation.

Definition 5.16. As we have explained, the word determinant is reserved for square matrices, so we suggest using this notation for the general case of an $(n\times k)$ matrix:

$$\text{Det}(A) \equiv \sqrt{\det(A^tA)}.$$

Note that (unlike for $\det A$), $\text{Det}A$ is always ≥ 0.

Lemma 5.30. *Let A be a $(n\times k)$ matrix. Then $Det(A) = Det(A^t)$. That is, $det(A^tA) = det(AA^t)$.*

Proof. Note that so far we know this only for square matrices! Also, the two matrices A^tA and AA^t are quite different, as the first is (2×2), while the second is (3×3).

We give the proof first for the case of a (3×2) matrix

$$A = \begin{bmatrix} v_1 & w_1 \\ v_2 & w_2 \\ v_3 & w_3 \end{bmatrix}.$$

We define $\mathbf{n}$ to be the unit vector perpendicular to $\mathbf{v},\mathbf{w}$ such that $(\mathbf{v},\mathbf{w},\mathbf{n})$ has positve orientation, so $\mathbf{n} = \mathbf{v}\wedge\mathbf{w}/||\mathbf{v}\wedge\mathbf{w}||$. Write $\widehat{A}$ for the (3×3) matrix:

Thus

$$\widehat{A}^t\widehat{A} = \begin{bmatrix} \longrightarrow & \mathbf{n} & \longrightarrow \\ \longrightarrow & \mathbf{v} & \longrightarrow \\ \longrightarrow & \mathbf{w} & \longrightarrow \end{bmatrix} \begin{bmatrix} \downarrow & \downarrow & \downarrow \\ \mathbf{n} & \mathbf{v} & \mathbf{w} \\ \downarrow & \downarrow & \downarrow \end{bmatrix} = \begin{bmatrix} n_1 & n_2 & n_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{bmatrix} \cdot \begin{bmatrix} n_1 & v_1 & w_1 \\ n_2 & v_2 & w_2 \\ n_3 & v_3 & w_3 \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{n}\cdot\mathbf{n} & \mathbf{n}\cdot\mathbf{v} & \mathbf{n}\cdot\mathbf{w} \\ \mathbf{v}\cdot\mathbf{n} & \mathbf{v}\cdot\mathbf{v} & \mathbf{v}\cdot\mathbf{w} \\ \mathbf{w}\cdot\mathbf{n} & \mathbf{w}\cdot\mathbf{v} & \mathbf{w}\cdot\mathbf{w} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \mathbf{v}\cdot\mathbf{v} & \mathbf{v}\cdot\mathbf{w} \\ 0 & \mathbf{w}\cdot\mathbf{v} & \mathbf{w}\cdot\mathbf{w} \end{bmatrix}$$

Now we know that $\det(\widehat{A^t}) = \det(\widehat{A})$, since these are square matrices. Thus

$$\text{Det}A = \det(\widehat{A}) = \det(\widehat{A^t}) = \text{Det}(A^t).$$

This is not a priori obvious since as noted above, $A^tA \neq AA^t$ (the first being (2×2), the second (3×3); also the second is more complicated!)

For the general case, with A an $(n\times k)$ matrix, first for $k<n$, instead of adding the single row $\mathbf{n}$ at the top to form $\widehat{A}$, we add $n-k$ vectors, as follows. We consider the subspace V generated by $\mathbf{v}_1,\dots\mathbf{v}_k$. and find an orthonormal basis $(\mathbf{u}_1,\dots\mathbf{u}_k)$ for V by the Gram-Schmidt procedure. We then complete this to an orthonormal basis

$(\mathbf{u}_1, \dots \mathbf{u}_k, \widehat{\mathbf{u}}_{k+1}, \dots, \widehat{\mathbf{u}}_n)$ of $\mathbb{R}^n$. Thus these last vectors are perpendicular to V. The resulting calculations are identical to the (3×3) case.

A fortiori this also proves the case $k > n$.

□

Theorem 5.31. *Given an $(n \times k)$ matrix A with $k \leq n$ and denoting by P the parallelopiped in $\mathbb{R}^n$ which is the image by A of $I_1 \times \cdots \times I_k$ where each $I_j = I$, then the k-dimensional volume of P is $DetA \equiv \sqrt{det(A^t A)}$.*

Proof. $\mathrm{vol}(P) = |\det(\widehat{A})| = \mathrm{Det}A.$ □

We would like to show this is independent of parameter change; the first proof is geometric and is simply that volume does not depend on parameterization. The second, algebraic proof shows how nice the formulas are:

Lemma 5.32. *(linear change–of–variables theorem) Let B be a $(k \times k)$ matrix and A a $(n \times k)$ matrix. Then for $\widetilde{A} = AB$, $Det(\widetilde{A}) = Det(A)|det(B)|$.*

Proof.

$$\begin{aligned}\mathrm{Det}(AB) &= \sqrt{\det((AB)^t AB)} = \sqrt{\det((B^t A^t) AB)} = \sqrt{\det(B^t (A^t A) B}\\ &= \sqrt{\det(B^t)\det(A^t A)\det(B)} = |\det B|\sqrt{\det(A^t A)}\end{aligned}$$

since as we know, $\det B = \det B^t$.

□

Corollary 5.33. *If $\mathbf{v}_1, \cdots, \mathbf{v}_k$ are the columns of A, and $\widetilde{\mathbf{v}}_1, \widetilde{\mathbf{v}}_k$ are the columns of AB, and if $detB = 1$, then the parallelopiped spanned by $\widetilde{\mathbf{v}}_1, \cdots, \widetilde{\mathbf{v}}_k$ and that spanned by $\mathbf{v}_1, \cdots, \mathbf{v}_k$ have the same volume.*

We give a second proof of Theorem 5.31, using only the geometric definition of determinant. This will use the same "sliding" idea we used in our geometric definition of the determinant, in Theorem 3.19.

Given the parallelogram $P(\mathbf{v}, \mathbf{w}) \subset \mathbb{R}^n$ spanned by the linearly independent vectors $\mathbf{v}, \mathbf{w}$, then a

(TO DO)

....

Given a (2×2) matrix $B = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, define $\widehat{B}$ to be the following (3×3) matrix:

$$\widehat{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & a & b \\ 0 & c & d \end{bmatrix}$$

then $\det \widehat{B} = \det B$.

....

Noting that:

$$\det(\widehat{A}) = (\mathrm{Det}A)$$

and so the area of $P(\mathbf{v}, \mathbf{w})$, which we write as area$(\mathbf{v}, \mathbf{w})$, equals the volume of the image of the unit cube by $\widehat{A}$, which we write as vol$(\widehat{A})$, we conclude that

$$\text{area}(\mathbf{v}, \mathbf{w}) = \text{vol}(\widehat{A}) = \det\widehat{A} = \text{Det}A$$

proving Hubbard's formula for this case.

5.9. **Surface area and surface integrals.** Given a domain (a connected open set) $B \subseteq \mathbb{R}^2$ and a $\mathcal{C}^1$ map $\sigma : B \to S \subseteq \mathbb{R}^3$, such that for every $(u.v) \in B$ is a *regular point*, i.e. such that $||\sigma_u \wedge \sigma_v|| \neq 0$, then as above, σ is a *parametrized surface.* (Recall that the regularity condition guarantees that the tangent plane exists at that point).

Definition 5.17. We define the *surface area* of σ to be

$$\text{area}(\sigma) = \int\int_B ||\sigma_u \wedge \sigma_v||\, \mathrm{d}u\mathrm{d}v.$$

This makes sense because area$(P(\mathbf{v}.\mathbf{w})) = ||\mathbf{v} \wedge \mathbf{w}||$ is the area of the parallelogram spanned by the vectors $\mathbf{v}, \mathbf{w}$, so $||\sigma_u \wedge \sigma_v||$ is the infinitesimal area, and the integral adds this up. The intuition is that for a $\mathbb{C}^1$ map, the surface can be well-approximated by a polygonal surface made up of parallelograms.

Given a parametrized surface σ as above and an invertible $\mathcal{C}^1$ map $H : A \to B$ then $\widetilde{\sigma} = \sigma \circ H$ is a *reparametrization* of σ via the change of parameter H.

The next result is the analogue of Proposition 3.7 for curves:

Theorem 5.34. *(area is invariant of parametrization). Suppose $A \subseteq \mathbb{R}^2$ is a domain and $H : A \to B$ is $\mathbb{C}^1$ and invertible. Then for $\widetilde{\sigma} = \sigma \circ H$,*

$$\textit{area}(\widetilde{\sigma}) = \textit{area}(\sigma).$$

Proof. We are to show that

$$\int\int_A ||\widetilde{\sigma}_s \wedge \widetilde{\sigma}_t||\, \mathrm{d}s\mathrm{d}t. = \int\int_B ||\sigma_u \wedge \sigma_v||\, \mathrm{d}u\mathrm{d}v.$$

We know from the change-of-variables formula for double integrals that for $F : B \to \mathbb{R}$, defining $\widetilde{F} = F \circ H$, then

$$\int\int_A \widetilde{F}(s,t)|\det DH(s,t)|\mathrm{d}s\mathrm{d}t = \int\int_A F\circ H(s,t)|\det DH(s,t)|\mathrm{d}s\mathrm{d}t = \int\int_B F(u,v)\, \mathrm{d}u\mathrm{d}v.$$

We define F to be $F(u,v) = ||\sigma_u \wedge \sigma_v||(u,v) = \text{Det}D\sigma(u,v)$ and as above $\widetilde{F} = F \circ H$. Since by the Chain Rule,

$$D(\sigma \circ H)(s,t) = D\sigma|_{(u,v)} DH|_{(s,t)}$$

we have at the point (s,t), using Lemma 5.32:

$$\begin{aligned} ||\widetilde{\sigma}_s \wedge \widetilde{\sigma}_t|| &= ||(\sigma \circ H)_s \wedge (\sigma \circ H)_t|| \\ &= \text{Det}(D(\sigma \circ H))|_{(s,t)} = \text{Det}(D(\sigma \circ H)|_{(s,t)}) DH|_{(s,t)}) \\ &= \text{Det}(D(\sigma \circ H)|_{(s,t)}) \det|DH|_{(s,t)}|. \end{aligned}$$

So

$$\int\int_A ||\widetilde{\sigma}_s \wedge \widetilde{\sigma}_t|| \det(DH|_{(s,t)})\, \mathrm{d}s\mathrm{d}t = \int\int_A \mathrm{Det}(D(\sigma\circ H)|_{(s,t)})\det(DH|_{(s,t)})\, \mathrm{d}s\mathrm{d}t$$

$$= \int\int_A \widetilde{F}(s,t)\det|DH|_{(s,t)}|\, \mathrm{d}s\mathrm{d}t = \int\int_B F(u,v)\, \mathrm{d}u\mathrm{d}v = \int\int_B ||\sigma_u \wedge \sigma_v||\, \mathrm{d}u\mathrm{d}v.$$

□

Theorem 5.35. *(surface integral is invariant of parametrization). Suppose $A \subseteq \mathbb{R}^2$ is a domain and $H : A \to B$ is $\mathbb{C}^1$ and invertible, and with $F : B \to \mathbb{R}$ continuous. Then for $\widetilde{\sigma} = \sigma \circ H$,*

$$\int\int_A F\circ H(s,t)||\widetilde{\sigma}_s \wedge \widetilde{\sigma}_t||\, ds dt = \int\int_B F(u,v)||\sigma_u \wedge \sigma_v||\, du dv.$$

Proof. We follow the above proof, again using the change-of-variables theorem for double integrals. □

Given a parametrized surface σ as above, with its image the parametrized surface $S = \sigma(B)$, and a function $G : S \to \mathbb{R}$, we define the *surface integral* of G over σ to be:

$$\int\int_S G(\mathbf{v})\mathrm{d}A = \int\int_B G(u,v)||\sigma_u \wedge \sigma_v||\, \mathrm{d}u\mathrm{d}v.$$

Theorem 5.35 shows this is indeed well-defined as it is invariant with respect to change of parameteriization.

This notation is analagous to the line integral of second type in Def. 5.2:

$$\int_\gamma f(\mathbf{v})\mathrm{d}s \equiv \int_a^b f(\gamma(t))\, ||\gamma'(t)||\mathrm{d}t$$

where we called $\mathrm{d}s = ||\gamma'(t)||\mathrm{d}t$ the infinitesimal arc length, and this integral is integration with respect to arc length. Here we call $\mathrm{d}A = ||\sigma_u \wedge \sigma_v||\, \mathrm{d}u\mathrm{d}v$ the *area form*, and this is an integral with repect to area.

5.10. **Integrals over parametrized submanifolds.** Let us note that the above formula for surface area of a parametrized surface σ can be written in a different way, given our definition of $\mathrm{Det}(A)$ for a rectangular matrix:

$$\mathrm{area}(\sigma) = \int\int_B ||\sigma_u \wedge \sigma_v||\, \mathrm{d}u\mathrm{d}v \int\int_B \mathrm{Det}D\sigma\, \mathrm{d}u\mathrm{d}v.$$

In this integral, we are not keeping track of orientation of the surface, since surface area is always positive. ($\mathrm{Det}M = \sqrt{M^tM} \geq 0$) and the integral of a positive function with respect to $\mathrm{d}A$ is always positive. This is just like a line integral of second type.

When we wish to include orientation, we use instead the notion of a *two-form* on $\mathbb{R}^3$ (or a *one-form* for line integrals).

The above formula, and the proof of invariance for change of parameter, extends immediately to the situation of a $\mathbf{k}$-dimensional space inside of $\mathbb{R}^n$:

Definition 5.18. Given a domain (a connected open set) $B \subseteq \mathbb{R}^k$ and a $\mathcal{C}^1$ map $\varphi : B \to S \subseteq \mathbb{R}^n$, such that for every $\mathbf{u} \in B$ is a *regular point*, i.e. such that $\mathrm{Det}D\varphi \neq 0$, equivalently $D\varphi$ has maximal rank ($= k$) then as above, φ is a *parametrized submanifold of dimension* k. The regularity condition guarantees that the tangent *space* to φ exists at tha point.

We define the *k-dimensional volume* of σ to be

$$\mathrm{vol}(\varphi) = \int\int_B \mathrm{Det}D\varphi \, \mathrm{d}\mathbf{u}$$

where $\mathrm{d}\mathbf{u} = \mathrm{d}x_1 \mathrm{d}x_2 \dots \mathrm{d}x_k$.

Given a parametrized submanifold φ as above and an nvertible $\mathbb{C}^1$ map $H : A \to B$ then $\widetilde{\varphi} = \varphi \circ H$ is a *reparametrization* of φ via the change of parameter H.

Theorem 5.36. *(k-dimensional volume is invariant of parametrization). Suppose $A \subseteq \mathbb{R}^k$ is a domain and $H : A \to B$ is $\mathbb{C}^1$ and invertible. Then*

$$vol(\widetilde{\varphi}) = vol(\varphi).$$

Proof. We exactly follow the proof for surfaces. □

Similar to the case for surfacees, we say for $M = \varphi(B)$: given a function $G : M \to \mathbb{R}$, we define the *volume integral* of G to be:

$$\int\int_B G(\mathbf{u}) \mathrm{Det}D\varphi \, \mathrm{d}\mathbf{u}.$$

Theorem 5.37. *(Change-of-variables theorem)....(TO DO)*

Proof. □

Next we prove Hubbard's formula in a different way. We consider a k-parallelopiped in $\mathbb{R}^n$.

First we give the proof for $k = 2$ and n arbitrary.

(TO DO)...

5.11. **The Divergence Theorem in space.** (TO DO)

5.12. **Stokes' Theorem.** Green's Theorem and the Divergence Theorem both turn out to be a special case of the fundamental result of vector calculus: Stokes' Theorem, where the points A, B, C, D are the boundary of the curve γ and get replaced by the boundary of any domain.

$$\int_{\partial\Omega} \omega = \int_{\Omega} \mathrm{d}\omega$$

or, in a different notation,

$$\langle \partial\Omega, \omega \rangle = \langle \Omega, \mathrm{d}\omega \rangle.$$

In this notation, which can be called *functional notation*, $\langle \cdot, \cdot \rangle$ is a *pairing* . A pairing is a bilinear operator, but on the right we have a vector space (of d-forms) and on the left an additive group (of d-chains, generated by d dimensional submanifolds). Here $d = k - 1$ on the left and $d = k$ on the right. The analogous assumption to the field being conservative is hidden here, in that we begin with a $k-1$-form on the left, like the potential, and take its derivative on the right, like its gradient.

(TO DO)

5.13. **Poincaré's Lemma: Existence of the vector potential.** A key idea of Vector Calculus is to extend the Fundamental Theorem of Calculus in a variety of ways. The first is that if for a vector field F on $\mathbb{R}^n$, we have a function $\phi : \mathbb{R}^n \to \mathbb{R}$ such that $\nabla\varphi = F$, then for a $\mathcal{C}^1$ path $\gamma : [a, b] \to \mathbb{R}^n$ with endpoints $A = \gamma(a), B = \gamma(b)$ then

$$\int_\gamma F \cdot \mathrm{d}\gamma = \int_a^b F(\gamma(t)) \cdot \gamma'(t)\mathrm{d}t = \varphi(B) - \varphi(A).$$

This is just like the case in one dimension where given $f : [a, b] \to \mathbb{R}$ and a function F satisfying $F' = f$ then

$$\int f(x)\mathrm{d}x = F(b) - F(a).$$

There is however one important difference: for f Riemann integrable there always exists such a *primitive* or *antiderivative* F, while for higher dimensions this only works if the field F is *conservative*, is which case φ is called a *potential function.*

Equivalently, in differental form notation, say for $F = (P, Q)$ then the form $\eta = P\mathrm{d}x + Q\mathrm{d}y$ has a *primitive* φ such that $\mathrm{d}\varphi = \eta$. This leads to the nice formula

$$\int_\gamma \eta = \int_\gamma P\mathrm{d}x + Q\mathrm{d}y = \int_{\partial\gamma} \mathrm{d}\eta = \int_{B=A} \varphi = \varphi(B) - \varphi(A).$$

The terminology thus that F has a potential φ iff η has a primitive φ.

In one dimension the potential is simply called the primitive, and can be defined from the integral by: $F(x) = \int_{x_0}^x f(r)\mathrm{d}r$. This is defined up to a constant, choice of which corresponds to changing the initial point x_0; thus we have made the choice $F(x_0) = 0$.

Exactly the same thing works for the line integrals, where we can attempt to define a potential function in the same way; we did this in Theorem 5.22. Recalling the proof, the potential is defined up to an initial poin A, setting $\varphi(A) = 0$ and $\varphi(B) = \int_a^b F(\gamma(t)) \cdot \gamma'(t)\mathrm{d}t$. This will be defined if there exists a path γ connecting A and B (by definition, iff the region is pathwise-connected) and will be *well*-defined iff this definition is independent of the path chosen. That is one of the equivalent characterizations of conservative field. (To prove that this definition indeed gives a potential, we calculated the partials and showed one indeed recovers the field; this is the "hardest" step in proving the equivalence of the conditions).

This result, which extends the Fundamental Theorem of Vector Calculus, itself extends much further, to Green's Theorem, the Divergence Theorem, and Stokes' Theorem in $\mathbb{R}^2$ and $\mathbb{R}^3$. All of this becomes simultaneously much more complicated and much simpler in its most natural setting: the generalized Stokes Theorem on manifolds with boundary.

Why we say "much simpler" is shown by the statement:

$$\int_{\partial B} \eta = \int_B \mathrm{d}\eta$$

or equivalently

$$\langle \partial B, \eta \rangle = \langle B, \mathrm{d}\eta \rangle.$$

The second notation exhibits the integral as a bilinear form, like an inner product. However here the elements on the right-hand side are differential forms, which form a vector space, while on the left-hand side these are *chains*, parametrized manifolds which can be added, subtracted or multiplied by integers, thus belonging to a *module* (over the ring $\mathbb{Z}$) rather than a vector space.

This second equation says that the *boundary operator* ∂ on chains is *dual* to the *exterior derivative* operator d on forms. This relationship can be summarized by saying that these operators are *adjoints*. (Note that this is indeed analogous to the definition of the transpose, or adjoint, of a linear operator!)

The first difficulty hidden by this simple notation is all in the definitions, which are equally abstract and deep. The secondary difficult comes in bridging the abstraction to the concrete versions of Vector Calculus in $\mathbb{R}^2$ and $\mathbb{R}^3$.

We mention two auxilliary points which come up in all these settings. The basic theorem is Stokes, which can be thought of as (and indeed can be called) the *Fundamental Theorem of Vector Calculus.*

We shall need:

Definition 5.19. A differential k- form η is *closed* iff $\mathrm{d}\eta = \mathbf{0}$.

It is *exact* iff there exists a $(k-1)$-form α such that $\mathrm{d}\alpha = \eta$.

Lemma 5.38. *If $d\alpha = \eta$ then $d\eta = 0$. Thus, $d(d\alpha) = \mathbf{0}$. That is, an exact form is closed.*

In fact, we have seen a special case of this in Proposition 5.18, that $\nabla \times (\nabla \varphi) = \mathbf{0}$.

The two other results are these:

Theorem 5.39. *(Poincaré Lemma) On a simply connected domain, a closed form is exact.*

Thus the Poincaré Lemma says that for topologically nice doman (simply connected), a primitive always exists; specifically, for one-forms in $\mathbb{R}^n$, we know this, since Again, we have seen this in the special case: in $\mathbb{R}^3$, for a simpky connected domain, for the dual vector field F, if $\mathrm{curl}(F) = \mathbf{0}$ then there exists φ such that $\nabla\varphi = F$, thus F has a potential. And $\nabla\varphi = F$ iff $\mathrm{d}\varphi = \eta = \sum P_i \mathrm{d}x_i$.

The second related result is:

Theorem 5.40. *(Hodge Decomposition) On a simply connected domain, every differential form can be uniquely written as the sum of a closed form and an exact form.*

For vector fields in $\mathbb{R}^n$, we say:

Definition 5.20. A vector field F is *divergence-free* or *incompressible* iff $\mathrm{div}(F) = 0$. It is *curl-free* or *conservative* or *irrotational* iff $\mathrm{curl}(F) = 0$.

The Hodge decomposition then gives:

Theorem 5.41. *(Helmholtz Decomposition) On a simply connected domain, every vector field which vanishes fast enough at ∞ can be uniquely written as the sum of a two vector fields, one divergence-free and one curl-free.*

Corollary 5.42. *A vector field on a simply connected domain, which vanishes fast enough at ∞, is determined by its divergence and its curl.*

Proof. By the Helmholtz Decomposition, our field$F = F_d + F_c$ where F_d is curl-free and F_c is divergence-free. Then $\text{curl}(F) = \text{curl}(F_c) + \text{curl}(F_d) = \text{curl}(F_c)$ and $\text{div}(F) = \text{div}(F_c) + \text{div}(F_d) = \text{div}(F_d)$. Hence $F = \text{curl}(F) + \text{div}(F)$.??? □

Both curl-free and divergence-free fields are of special interest. A curl-free field is locally conservative so is related to mechanics; a divergence-free field represents for example water flow. It can have non-0 curl, as water can exhibit vortex lines and vortex tubes. Water is incompressible hence divergence-free, while a gas like air is in between: both curl andd divergence can be non-0, and we do see rotation and vortex lines and tubes like hurricanes, tornadoes, and smoke rings.

For vector fields on a simply connected domain in $\mathbb{R}^n$, there are two versions of Poincaré's Lemma. The first says that a curl-free vector field has a potential, hence is conservative: if $\text{curl}F = 0$ then there exists φ such that $\nabla\varphi = F$.

The second statement is:

Theorem 5.43. *If div$(F) = 0$, then there exists a field A such that curl$(A) = F$.*

For the proof we need:

Lemma 5.44. *(Derivative under the Integral) Suppose for $\mathcal{U} \subseteq \mathbb{R}^2$ open that $f : \mathcal{U} \to \mathbb{R}$ is continuous, and that $\partial f/\partial y$ exists and is continuous. Define $\varphi(y) = \int_a^b f(x,y)\,dx$. Then*

$$\varphi'(y) = \frac{d}{dy}\int_a^b f(x,y)\,dx = \int_a^b \frac{\partial f}{\partial y}(x,y)\,dx.$$

Example 16. Before the proof, we consider some examples.

Remark 5.12. For these examples, recall that the Gaussian function e^{-x^2} is a well-known example for which the antiderivative cannot be found "in closed form". Roughly this means as a finite formula involving other elementary functions (polynomials, trigonometric functions, log and exp); for a precise statement, which makes use of the notion of a *differential field,* see [Ros68]. (Note that one can however easily give an infinite formula, using Taylor's series.)

Problem: Find $\partial\varphi/\partial x$ and $\partial\varphi/\partial y$ for

$$\varphi(x,y) = \int_0^x e^{-yt^2}\mathrm{d}t.$$

The first is easy, as by the Fundamental Theorem of Calculus we have: $\frac{\partial\varphi}{\partial x} = e^{-yx^2}$.

For the second, we apply the Lemma, giving:

$$\frac{\partial\varphi}{\partial y}\int_0^x e^{-yt^2}\mathrm{d}t = \int_0^x \frac{\partial}{\partial y}e^{-yt^2}\mathrm{d}t = \int_0^x -t^2e^{-yt^2}\mathrm{d}t.$$

Problem: For

$$h(t) = \int_0^{t^2} e^{-tu^2}\,\mathrm{d}u,$$

calculate $h'(t)$.

Well, this is indeed pretty confusing, as the variable t occurs in two different spots! The trick is to first define a function of two variables

$$\varphi(x,y) = \int_0^x e^{-yu^2}\,\mathrm{d}u$$

and then compose it with a curve. Now as above $\frac{\partial\varphi}{\partial x} = e^{-yx^2}$, while

$$\frac{\partial\varphi}{\partial y} = \int_0^x -u^2 e^{-yu^2}\,\mathrm{d}u.$$

Defining the curve $\gamma(t) = (t^2, t)$ then $h(t) = \varphi(\gamma(t))$. We then apply the Chain Rule:

$$h'(t) = \nabla\varphi|_{\gamma(t)} \cdot \gamma'(t) = \frac{\partial\varphi}{\partial x}|_{\gamma(t)}x'(t) + \frac{\partial\varphi}{\partial y}|_{\gamma(t)}y'(t) =$$

$$e^{-tt^4}2t + \int_0^{t^2} -u^2e^{-tu^2}\mathrm{d}u = 2te^{-t^5} - \int_0^{t^2} u^2e^{-tu^2}\mathrm{d}u.$$

Proof. (of Lemma) Our proof follows Apostol p. 448 [**?**].

We are given that $\varphi(y) = \int_a^b f(x,y)\mathrm{d}x$ and want to find $\varphi'(y)$. We have:

$$\frac{\varphi(y+h)-\varphi(y)}{h} = \frac{1}{h}\left(\int_a^b f(x,y+h)\mathrm{d}x - \int_a^b f(x,y)\mathrm{d}x\right) = \frac{1}{h}\left(\int_a^b f(x,y+h) - f(x,y)\mathrm{d}x\right)$$

Now by the Mean Value Theorem, for each fixed y there exists $c_y \in [a,b]$ such that

$$f(x,y+h) - f(x,y) = \frac{\partial f}{\partial y}(c_y,y)\cdot h.$$

So

$$\frac{\varphi(y+h)-\varphi(y)}{h} = \frac{1}{h}\left(\int_a^b \frac{\partial f}{\partial y}(c_y,y)\cdot h\ \mathrm{d}x\right) = \int_a^b \frac{\partial f}{\partial y}(c_y,y)\ \mathrm{d}x$$

But by continuity of the partial derivative, $\frac{\partial f}{\partial y}(c_y,y) \to \frac{\partial f}{\partial y}(x,y)$ as $h \to 0$. This gives

$$\frac{\varphi(y+h)-\varphi(y)}{h} \to \int_a^b \frac{\partial f}{\partial y}(x,y)\ \mathrm{d}x$$

as claimed.

□

Proof. (of Theorem)

We give the proof for the easier case of a star-shaped domain.

Let $F = (P,Q,R)$, with $0 = \mathrm{div}F = P_{xx} + Q_{yy} + R_{zz}$.

We want to find a field $A = (L,M,N)$ such that $\mathrm{curl}(A) = F$. Now

$$\mathrm{curl}A = (N_y - M_z, L_z - N_x, M_x - L_y).$$

We consider the simpler case where $L = 0$. Then we have

$$L_z - N_x = -N_x = Q, M_x - L_y = M_x = R$$

Now $\partial N(x,y,z)/\partial x = \mathrm{d}/\mathrm{d}t|_{t=x}(Q(t,y,z)$ so for any inital point $x_0,y,z)$ we have

$$N(x,y,z) = \int_{t=x_0}^x Q(t,y,z)\mathrm{d}t + c(y,z)$$

where $c(y,z)$ is constant in x. Similarly

$$M(x,y,z) = \int_{t=x_0}^{x} R(t,y,z)\mathrm{d}t + d(y,z)$$

where $c(y,z)$ is constant in x.

We look for a solution with $c(y,z) = 0$. We know that $P = N_y - M_z$ so subtracting the previous two equations gives

$$P = N_y - M_z = \partial/\partial y \int_{t=x_0}^{x} Q(t,y,z)\mathrm{d}t - \partial/\partial z \int_{t=x_0}^{x} R(t,y,z)\mathrm{d}t + \partial/\partial z d(y,z)$$

Now from the Lemma, taking the derivative inside the integral, this gives

$$\int_{t=x_0}^{x} \partial/\partial y Q(t,y,z)\mathrm{d}t - \int_{t=x_0}^{x} \partial/\partial z R(t,y,z)\mathrm{d}t + \partial/\partial z d(y,z) =$$
$$\int_{t=x_0}^{x} -\partial/\partial y Q(t,y,z) - \partial/\partial z R(t,y,z)\mathrm{d}t + \partial/\partial z d(y,z)$$

Using the fact that $\mathrm{div}F = 0$, we know that $-Q_y - R_z = P_x$ so this is

$$\int_{t=x_0}^{x} -\partial/\partial x P(t,y,z)\mathrm{d}t + \partial/\partial z d(y,z) = P(x,y,z) - P(x_0,y,z) + \partial/\partial z d(y,z)$$

We now have the equation

$$P(x,y,z) = P(x,y,z) - P(x_0,y,z) + \partial/\partial z d(y,z)$$

So we will be done if we can find a function $d(y,z)$ satisfying

$$\partial/\partial z d(y,z) = -P(x_0,y,z)$$

(check all signs!)

So we simply define

$$d(y,z) = \int_{t=z_0}^{z} P(x_0,y,r)\mathrm{d}r$$

giving the first part of the solution, defined up to a constant.

So far we have shown that for $F = (P,Q,R)$ then the field $A = (L,M,N)$ with $L = 0$.

$$N(x,y,z) = \int_{t=x_0}^{x} Q(t,y,z)\mathrm{d}t$$

$$M(x,y,z) = \int_{t=x_0}^{x} R(t,y,z)\mathrm{d}t + d(y,z)$$

where

$$d(y,z) = \int_{t=z_0}^{z} P(x_0,y,r)\mathrm{d}r$$

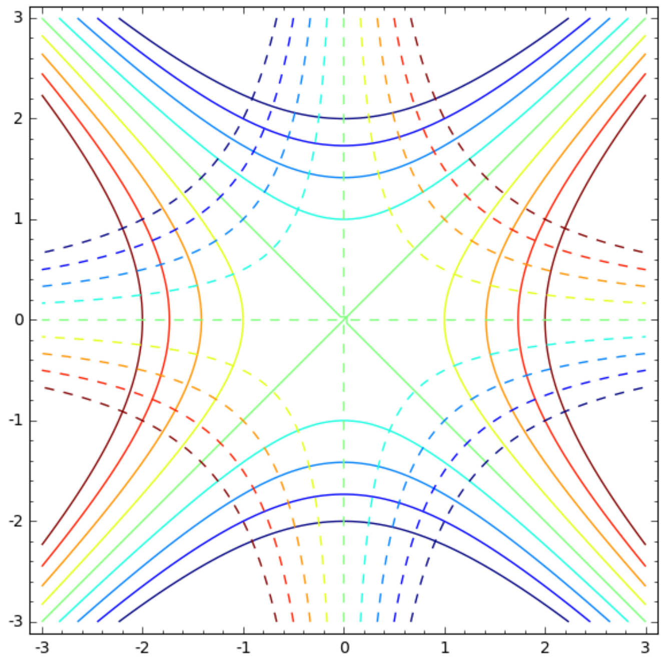


FIGURE 21. Dual families of hyperbolas: Level curves (equipotential curves) for the real and imaginary parts of $f(z) = z^2 = (x+iy)(x-iy) = (x^2 - y^2) + 2(xy)i$

Putting these together we have shown that given $F = (P, Q, R)$, then for $A = (L, M, N)$ defined by

$$L = 0$$
$$N(x, y, z) = \int_{t=x_0}^{x} Q(t, y, z)\mathrm{d}t$$
$$M(x, y, z) = \int_{t=x_0}^{x} R(t, y, z)\mathrm{d}t + \int_{t=z_0}^{z} P(x_0, y, r)\mathrm{d}r.$$

then we have

$$\mathrm{curl}(A) = F$$

.

□

Remark 5.13. There is a strangeness in the above proof as we arbitrarily chose $L = 0$ and yet somehow found a solution.

This is explained by noting that any solution A above is defined up to addition of a field B with $\mathrm{curl}(B) = 0$. Call the particular solution above A_L. Then if we carry out the above construction assuming instead that $M = 0$ we get solution A_M and if we assume instead $N = 0$ we get solution A_N. But then indeed $\mathrm{curl}(A_L - A_M) = F - F = 0$ and similarly $\mathrm{curl}(A_L - A_N) = 0$, $\mathrm{curl}(A_M - A_N) = 0$.

5.14. **Analytic functions and harmonic conjugates.**

Definition 5.21. A remarkable fact in Complex Analysis is that we have these three equivalent definitions:
(i) A function $f : \mathcal{U} \subseteq \mathbb{C} \to \mathbb{C}$ is *holomorphic* iff it is complex differentiable, i.e. its derivative, given by the usual limit, exists and is a complex number. Thus, f is holomorphic at $z \in \mathcal{U}$, with derivative $f'(z)$, iff $f'(z) = \lim_{h\to 0}(f(z+h) - f(z))/h$ exists.

If this number is $z \in \mathbb{C}$, then writing $z = re^{i\theta}$ for $r \geq 0$, since by *Euler's formula* $e^{i\theta} = \cos\theta + i\sin\theta = c + is$, we see that the multiplication map $w \mapsto z \cdot w$ is in real coordinates

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \mapsto r \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

In other words the function has a very special type of $\mathbb{R}^2$-derivative: a dilation composed with a rotation.
(ii) This implies the map is *conformal*: angles and orientation are preserved infinitesimally. By contrast, an *anti*conformal map preserves angles but *reverses* orientation; the simplest example is $z \mapsto \overline{z}$ where for $z = a+ib$, its *complex conjugate* is $\overline{z} = a-ib$. A general antiholomorpic map is given by a holomorphic map preceded or followed by complex conjugation, so the $\mathbb{R}^2$-derivative is a rotation composed with a reflection in a line through $(0,0)$. Note that for both conformal and anticonformal maps, infinitesimal circles are taken to infinitesimal circles (not ellipses, which is the general case).
(iii) A function is *(complex) analytic* iff it has a power series expansion near a point.

In particular, knowing a function has one continuous complex derivative, i.e. in $\mathcal{C}^1$, implies, very differently from the real case, it is not only infinitely continuously differentiable ($\mathcal{C}^\infty$) but has a power series (is $\mathcal{C}^\omega$).

Remark 5.14. Recalling Definition 5.21, if $f : \mathbb{C} \to \mathbb{C}$ is a complex analytic function, with $f = u + iv$, then this defines a vector field $F = (u, v)$ on $\mathbb{R}^2$. We note that in this case the field F has a special form:

$$DF = \begin{bmatrix} u_x & u_y \\ v_x & v_y \end{bmatrix} = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$$

since f is analytic iff it is complex differentiable, meaning that $f'(z)$ is a complex number $w = a + ib = re^{i\theta}$, giving a dilation times a rotation. This proves the *Cauchy-Riemann equations* $u_x = v_y$, $u_y = -v_x$.

Now the line integral $\int_\gamma F d\gamma$ is closely related to the *contour integral* of f over γ, written $\int_\gamma f$. The beginnings of the theory are developed in parallel; see e.g. [MH87] p.95 ff. In particular, the winding mumber can be defined using a contour integral. Of course this is only a starting point for the deep and beautiful subject of Complex Analysis.

Definition 5.22. A function $u : \mathcal{U} \to \mathbb{R}$ is *harmonic* iff u is $\mathcal{C}^2$ and $u_{xx} + u_{yy} = 0$.

We define a linear operator Δ, also written as ∇^2 and called the *Laplacian*, on the vector space $\mathcal{C}^2(\mathcal{U}, \mathbb{R})$ by $\Delta(u) = u_{xx} + u_{yy}$. So u is harmonic iff $\Delta(u) = 0$, iff u is in the kernel of the operator.

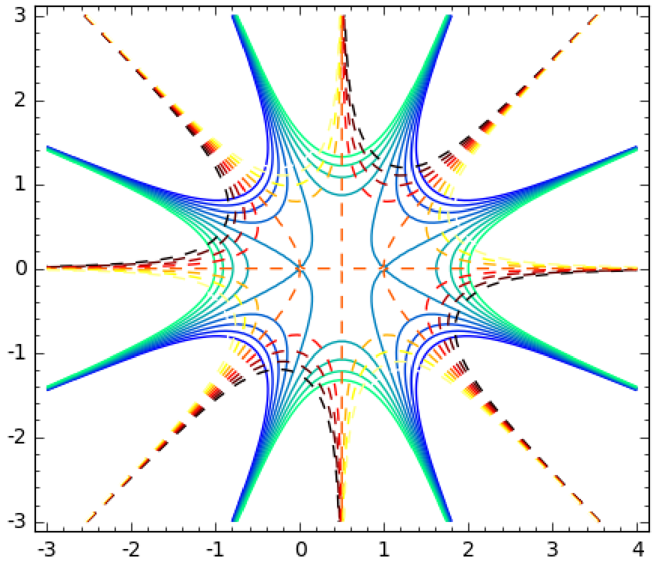


FIGURE 22. Level curves for the real and imaginary parts of $f(z) = z^2(z-1)^2$.

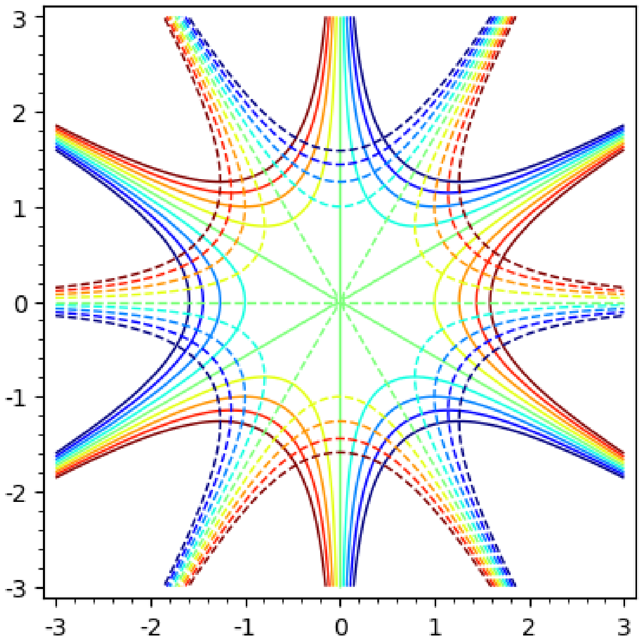


FIGURE 23. Level curves for the real and imaginary parts of $f(z) = z^3$.

The reason for the notation ∇^2 isbeacuse it is notationally suggestive, as we can think of it as the dot product: $\nabla^2\varphi = (\nabla\cdot\nabla)(\varphi) = \nabla\cdot(\nabla(\varphi) = (\delta/\delta x+\delta/\delta y)\cdot(\varphi_x+\varphi_y)$.

Theorem 5.45. *For a complex analytic function $f : \mathcal{U} \to \mathbb{C}$, where $\mathcal{U} \subseteq \mathbb{C}$ is open, with real and imaginary parts $u = \Re(f), v = \ Im(f)$ so $f = u + iv$, then thought of as real functions on $\mathcal{U} \subset \mathbb{R}^2$,*
(i) these satisfy the Cauchy-Riemann equations $u_x = v_y$, $u_y = -v_x$;
(ii) u, v are both harmonic functions;
(iii) their gradient vector fields are orthogonal;
(iv) their families of level curves are orthogonal

Proof. If $f : \mathbb{C} \to \mathbb{C}$ is a complex analytic function, then by definition the derivative $f'(z) = \lim_{h\to 0}(f(z+h) - f(z))/h$ is a complex number $w = (a + ib) = re^{i\theta}$. Now multiplication by a complex number defines a linear transformation of $\mathbb{C}$ hence of $\mathbb{R}^2$; since this is a rotation followed by a dilation, this matrix has a special form. Writing $f = u + iv$, then thought of as a map F of $\mathbb{R}^2$, this is the vector field $F = (u, v)$, the derivative of which is the matrix

$$DF = \begin{bmatrix} u_x & u_y \\ v_x & v_y \end{bmatrix}.$$

Because we know this is a rotation by θ followed by a dilation by $r \geq 0$, this equals

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix} = r \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

This proves the Cauchy-Riemann equations $u_x = v_y$, $u_y = -v_x$.

Now $u_x = v_y$ whence $u_{xx} = v_{xy}$ and
$u_y = -v_x$ whence $u_{yy} = -v_{yx}$, giving that

$$u_{xx} + u_{yy} = v_{xy} - v_{yx} = 0$$

by the equality of mixed partials, Lemma 3.31.

Similarly, from $u_x = v_y$ we have that $u_{yx} = v_{yy}$ and from $u_y = -v_x$ that $u_{xy} = -v_{xx}$ whence

$$v_{xx} + v_{yy} = u_{xy} - u_{yx} = 0.$$

So both u and v are harmonic.

Recalling the notation that for $F = (P, Q)$ then $F^* = (Q, -P)$, we have

$$\nabla u = (u_x, u_y) = (v_y, -v_x)$$

so

$$\nabla v = F^*$$

gives an orthogonal field.

Lastly the level curves are perpendicular to the gradient fields, F and F^*, so since these are orthogonal so are those families of curves.

□

In fact the converse also holds:

Proposition 5.46. *If u, v are $\mathcal{C}^2$ functions which are harmonic, such that the pair (u, v) satisfies the Cauchy-Riemann equations, then $f = u + iv$ is analytic.*

Proof. As above, the derivative of $F : \mathbb{R}^2 \to \mathbb{R}^2$ with $F = (u, v)$ is the matrix

$$DF = \begin{bmatrix} u_x & u_y \\ v_x & v_y \end{bmatrix}.$$

The Cauchy-Riemann equations imply that this equals

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$$

and so the map is given by multiplication by the complex number $w = re^{i\theta}$ as above. That the limit exists for DF implies that the limit exists for $f'(z)$ and equals w. □

Definition 5.23. If if $f = u + iv$ is analytic then u, v are called *harmonic conjugates.*

Proposition 5.47.
(i) If $\mathcal{U}$ is a simply connected domain and $u : \mathcal{U} \to \mathbb{R}$ is harmonic, then there exists a unique $v : \mathcal{U} \to \mathbb{R}$ such that (u, v) are harmonic conjugates.
(ii) The ordered pair (u, v) are harmonic conjugates iff the pair $(v, -u)$ are (so order matters here!)

Proof. (i): By the previous proposition it is enough to find v harmonic such that (u, v) satisfies the Cauchy-Riemann equations, so such that $v_y = u_x$ and $v_x = -u_y$. But this is just like the problem of finding a potential for a curl zero vector field!

Thus, we consider the vector field $F = (P, Q) = (-u_y, u_x)$.

Then $\text{curl}(F) \cdot \mathbf{k} = Q_x - P_y = -u_{xx} - u_{yy} = 0$.

By Theorem 5.22, there is a potential for F; we call this v. Thus $\nabla(v) = (v_x, v_y) = (P, Q) = (-u_y, u_x)$ so $v_{xx} + v_{yy} = -u_{xy} + u_{yx} = 0$ by the equality of mixed partials, whence v is a harmonic function such that the pair (u, v) are indeed harmonic conjugates.

In this proof of (i) we have followed Churchill [CB14]. □

Remark 5.15. Lang [Lan99] and Marsden-Hoffman [MH87] have nice treatments of this. Following Marsden and Hoffman, note that since $if = v - iu$ is analytic, then v and $-u$ are harmonic conjugates (but that the order is important!) A second, purely complex analytic, proof of (i) is given by Marsden [MH87]. See also Ahlfors [Ahl66].

Fig. 21 shows the harmonic conjugates for the function $f(z) = z^2$.

Corollary 5.48. *Given a harmonic function $u : \mathcal{U} \to \mathbb{R}$, where $\mathcal{U}$ is a simply connected domain in $\mathbb{R}^2$, then there exists a unique $v : \mathcal{U} \to \mathbb{R}$, which is harmonic such that (u, v) satisfy the Cauchy-Riemann equations. Also there exists a unique analytic f on $\mathcal{U}$ thought of as a subset of $\mathbb{C}$ such that $u = \Re(f)$. Moreover $f = u + iv$. Writing $\widetilde{f}$ for the second analytic function defined from the harmonic function v, then $\widetilde{f} = v - iu$ has harmonic conjugate pair $(v, -u)$. Furthermore $\widetilde{f}(z) = -if(z)$.*

Harmonic functions are characterized by the important *mean value property*: for a proof see e.g. [MH87].

Theorem 5.49. *A $\mathcal{C}^2$ function u is harmonic iff the value at a point $\mathbf{p}$ is equal to the average of the values on any circle about $\mathbf{p}$.*

Definition 5.24. A flow τ_t on $\mathbb{R}^n$ is a *gradient flow* iff there is a function $\varphi : \mathbb{R}^n \to \mathbb{R}$ such that for the field $F = \nabla\varphi$, then the flow orbits are tangent to the gradient vector field. That is, the orbits $\gamma(t) = \tau_t(x)$ for some initial point x satisfy the differential equation

$$\gamma'(t) = F(\gamma(t))$$

.

We conclude:

Theorem 5.50. *Let u be a harmonic function on $\mathcal{U} \subseteq \mathbb{R}^2$. Let v be its harmonic conjugate. Write $F = (u, v)$ and $\widetilde{F} = (-v, u)$, so $F = \nabla u$ and $\widetilde{F} = \nabla v$. Then the gradient flow of u is the flow of F, and the gradient flow of v is the flow of $\widetilde{F}$. The flow lines of F are the level curves of v and the flow lines of $\widetilde{F}$ are the level curves of u. The orbits of F and $\widetilde{F}$ are mutually orthogonal.*

Example 17. Consider $f(z) = z^2 = u + iv$. The gradient fields are $F(\mathbf{v}) = A\mathbf{v}$ and $\widetilde{F}(\mathbf{v}) = \widetilde{A}\mathbf{v}$ where

$$A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

and

$$\widetilde{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

for the potentials u and v respectively.

5.15. **Electrostatic and gravitational fields in the plane and in $\mathbb{R}^3$.** The same geometry (with dual, orthogonal families of level curves) happens for electrostatic fields: one family is the *equipotentials* (curves or surfaces, depending on the dimension) while the other depicts the *lines of force*: flow lines tangent to the force vector field. See the Figures.

When the opposite charges of Fig. 25 get closer and closer, the behavior approximates that of an *Electrostatic Dipole*; see Figs. 26, 30. The charges would cancel out, if we place one on top of the other, but if we take a limit of the fields as the distance d goes to 0 as charges c are balanced with this so that the product dc remains constant, then the limit of the fields (and potentials) exists. Note there is a limiting vector from plus to minus, along the x-axis. The picture is for the case of charges in the plane.

We note here that the *pictures* are unchanged by this sort of normalization, since:

Lemma 5.51.
(i) If F is a conservative field on $\mathbb{R}^n$ with potential function φ, then the collection of equipotential curves (or dimension $(n-1)$ submanifolds) is the same as for the field aF, $\mathbf{a} \neq 0$.
(ii) If γ is a line of force for F, then γ is orthogonal to each equipotential submanifold.

Proof. (i) We have: $\nabla\varphi = F$ iff $\nabla a\varphi = aF$, and the level curve of level c corresponds to that of level ac.
(ii) line of force for F is a curve γ with the property that $\mathcal{F}(\gamma(t) = \gamma'(t)$, i.e. γ is tangent to the field everywhere (is an orbit of the flow for the ODE). Then $\varphi(\gamma(t) = c$

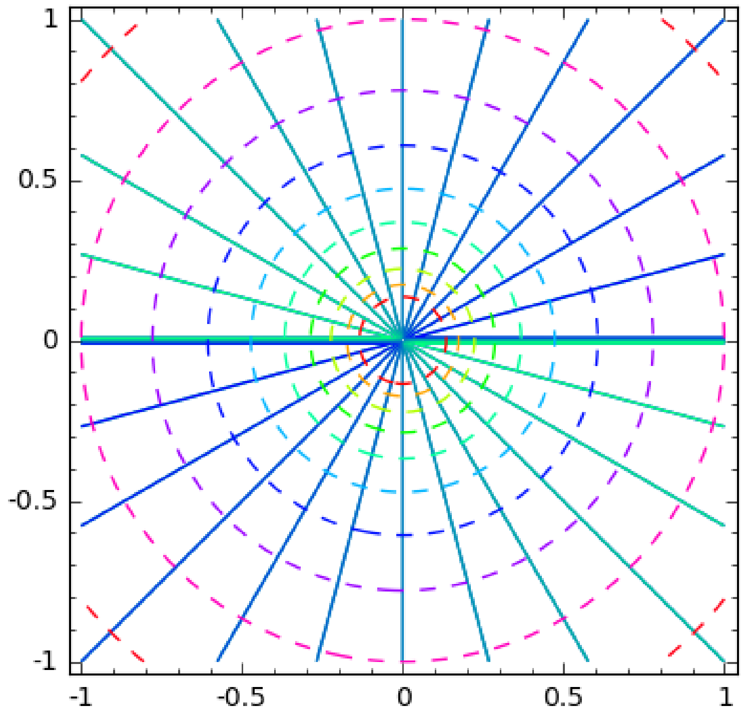


FIGURE 24. Equipotential curves and lines of force for the electrostatic field of a single charge in the plane. The equipotentials are level curves for the potential function φ and change color as the angle increases from 0 to π and again from π to 2π. This depends on the formula chosen for φ and the "color map" chosen for the graphics. In complex terms, the complex log function is $f(z) = \log(z)$ and for $z = re^{i\theta}$ with $\theta \in [0, 2\pi)$ then $f(z) = \log(re^{i\theta}) = \log(r) + \log(e^{i\theta}) = \log(r) + i\theta = u + iv$ with harmonic conjugates $u(x, y) = \log(r)$ and $v(x, y) = \theta$. We see the level curves in the Figure; they form a spiral staircase. See Fig. 20.

so $\varphi(\gamma(t))' = 0$ but by the Chain Rule this is $\varphi(\gamma(t))' = \nabla\varphi(\gamma(t)) \cdot \gamma'(t) = F(\gamma(t)) \cdot \gamma'(t)$.

$\square$

That the pictures converge (of both the equipotentials and field lines) looks clear from the figures, but to have the fields and potentials converge we need this normalization.

The potential function shown is

$$u(u, y) = \frac{1}{d} \log \frac{(x+d)^2 + y^2}{(x-d)^2 + y^2}$$

for $d = 1, .5, .05$.

Dipoles (both electric and magnetic) are useful in applications to electrical engineering and are itnriguing mathematically.

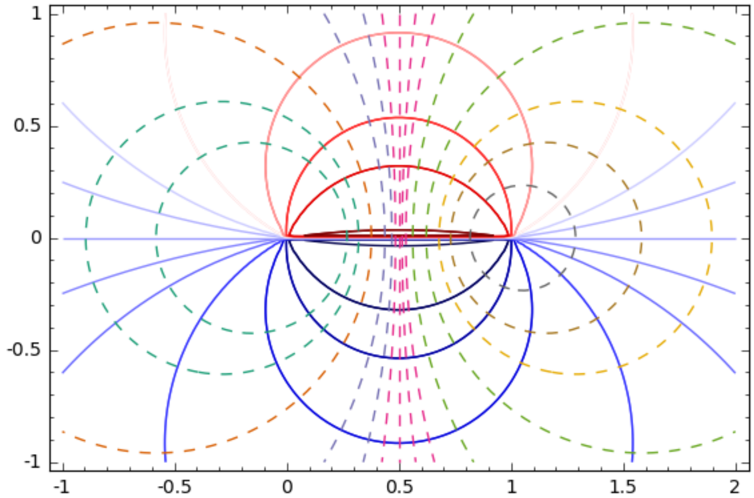


FIGURE 25. Equipotential curves and lines of force for the electrostatic field of two opposite charges in the plane. Colors indicate different levels of the potential and dual potential, where these are the harmonic conjugates coming from the associated complex function $g(z) = f(z) - f(z-1) = \log(z) - \log(z-1)$. These harmonic functions are $u(x, y) - u(x-1, y)$ and $v(x, y) - v(x-1, y)$.

We mention that the geometry of fields in two-dimensional space has practical relevance: for example, the magnetic field generated by electric current passing throung a wire (in the form of a line) decreases like $1/r$, as we can think of the field as being in the plane perpendicular to the wire. For fascinating related material see the Wikipedia article on *Ampere's circuital law.*

Experiments show that the force between two charged particles with charges $q_1, q_2 \in \mathbb{R}$ with position difference given by a vector $\mathbf{v} \in \mathbb{R}^3$ is

$$\frac{q_1 q_2}{r^2} \cdot \frac{\mathbf{v}}{||\mathbf{v}||}, r = ||\mathbf{v}||$$

(so it is positive hence repulsive if the charges have the same sign).

An intuitive explanation for the factor of $1/r^2$ is this: suppose we have a light bulb at the origin and we want to calculate the light density at distance r; the light consists of photons, and the number emitted per second is the same as the number that pass through a sphere of radius r, which is proportional to the area $4\pi r^2$. Another way to say this that we are counting the number of field lines per unit area. Both the electrostatic field of a single charge and gravity (which is more simple as the is no negative gravity) are mediated by radiating particles and so should decrease in the same way.

We claim that the attractive potential φ of a single charge in $\mathbb{R}^3$ is

$$\varphi = 1/r = (x^2 + y^2 + z^2)^{-1/2}$$

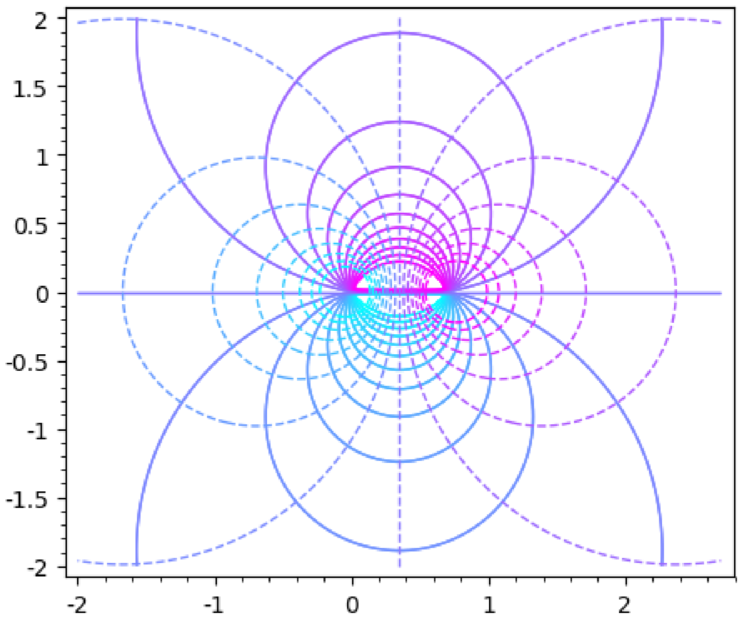


FIGURE 26. Equipotential curves and lines of force for the electrostatic field of two unlike charges, now closer together.

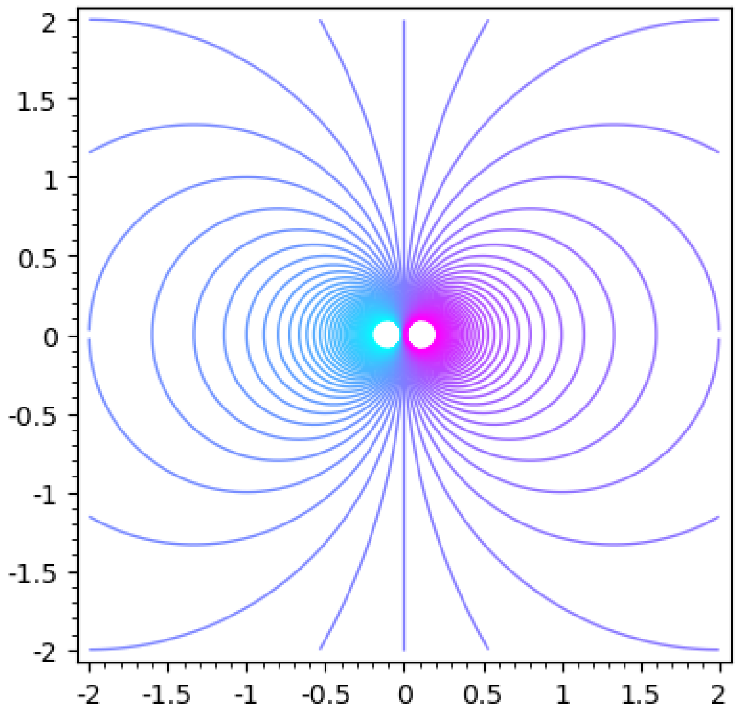


FIGURE 27. Equipotential curves for the electrostatic field of a planar dipole: two unlike charges close together.

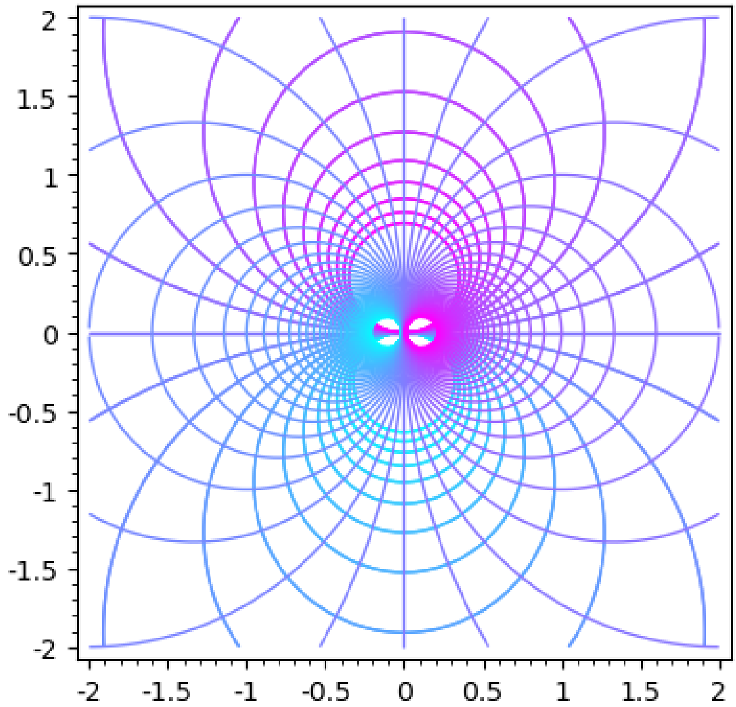


FIGURE 28. Equipotential curves and lines of force for the electrostatic field of a planar dipole: two unlike charges very close together. The potential is $1/d\log(((x-d)^2+(x+d)^2)$ for $d=0.5$.

Since the force field is then $F=\nabla\varphi$ we have $F=(P,Q,R)$ where

$$P=\frac{-x}{(x^2+y^2+z^2)^{3/2}}$$

and similarly for Q,R. The field strength at (x,y,z) is then

$$F(x,y,z)=||(x,y,z)||/||(x,y,z)||^3=1/r^2$$

as we wanted.

We are thinking of a single large charge being tested by a small charge; we are not yet calculating the resulting field of two equal charges (or the gravitational field of two equal mass objects).

In two dimensions, the math is very different, as the field strength now should be proportional to $1/r$ as it is inversely proportional to the circumference of a circle, $2\pi r$.

Thus in $\mathbb{R}^2$, for a single unit charge particle at the origin, we claim that the potential is

$$\varphi(x,y)=\frac{1}{2}\log(x^2+y^2)$$

for then the force field is

$$F=(P,Q)=\nabla\varphi=\left(\frac{x}{x^2+y^2},\frac{y}{x^2+y^2}\right)$$

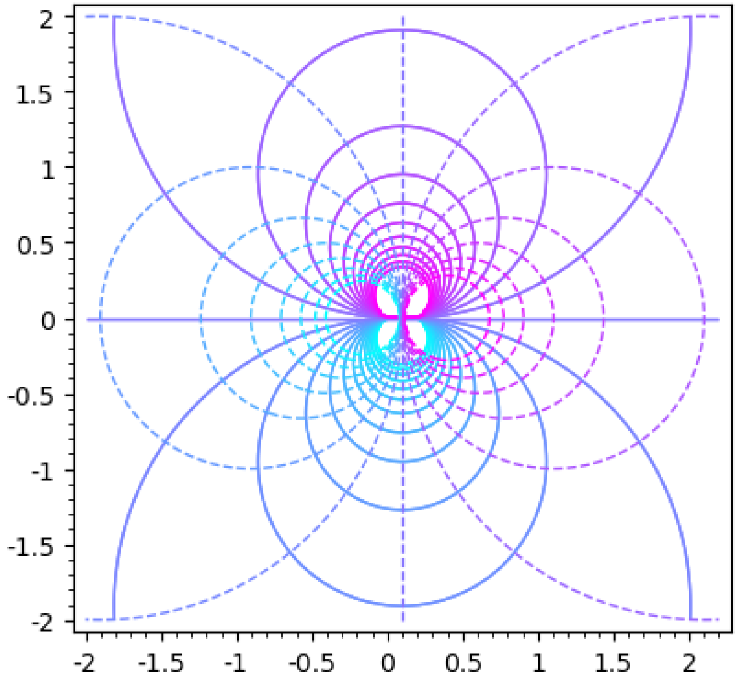


FIGURE 29. Equipotential curves and lines of force for the electrostatic field of an approximate planar dipole: two unlike charges close together.

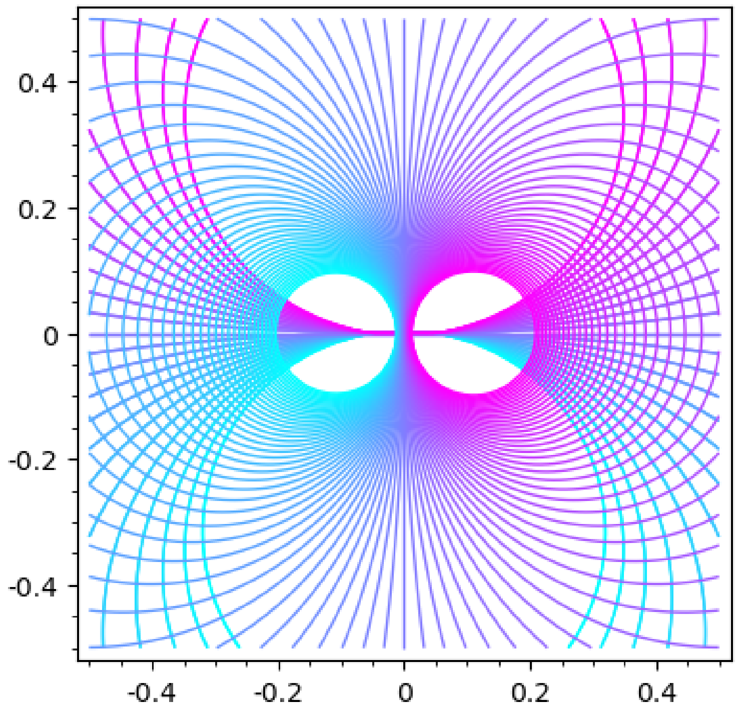


FIGURE 30. Equipotential curves and lines of force for the electrostatic field of an approximate planar dipole: two unlike charges close together.

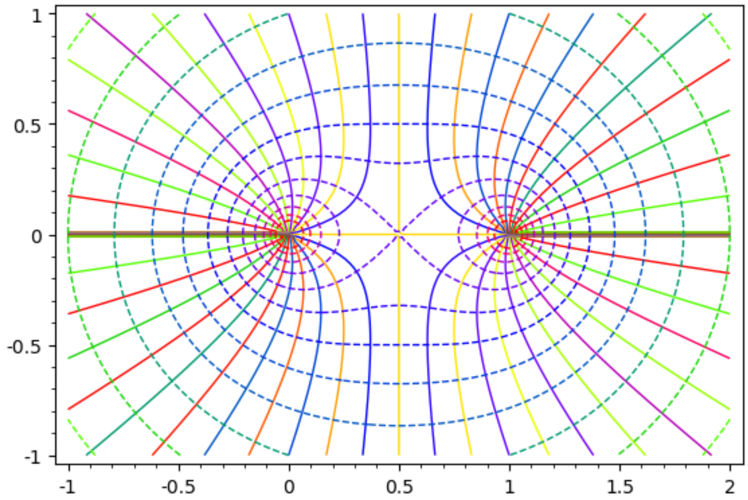


FIGURE 31. Equipotential curves and lines of force for the electrostatic field of two like charges in the plane. Since for one charge at $\mathbf{0}$ the associated complex function is $f(z) = \log(z) = u + iv$, here it is $g(z) = f(z) + f(z-1) = \log(z) + \log(z-1)$. The equipotentials and field lines are respectively the level curves for the harmonic conjugates $u(x,y) + u(x-1,y)$ and $v(x,y) + v(x-1,y)$.

which has norm

$$||F|| = ||(x,y)||/||(x,y)||^2 = 1/r,$$

as we wished.

The dual field is

$$F^* = (-Q, P) = \left(\frac{-y}{x^2+y^2}, \frac{x}{x^2+y^2}\right)$$

which as we have seen in §5.4 has potential the angle function Θ of Fig. 20, given by $\psi(x,y) = \arctan(y/x)$ or $\psi(x,y) = \text{arccot}(x/y)$ depending on the location (since $\mathbb{R}^2 \setminus \mathbf{0}$ is not simply connected).

The corresponding analytic function is

$$f(z) = \log(z)$$

and for $z = re^{i\theta}$ with $\theta \in [0, 2\pi)$ then $f(z) = \log(re^{i\theta}) = \log(r) + \log(e^{i\theta}) = \log(r) + i\theta = u + iv$ giving the harmonic conjugates $u(x,y) = \log(r)$ and $v(x,y) = \theta$, whose level curves we see in Fig. 24.

This is the case of a single charge. In fact, when combining objects all we have to do is add the two potentials, $\varphi = \varphi_1 + \varphi_2$, and then the gradient will give the field. See Figs. 25, 31 for the cases of two oppositely, and equally, charged particles.

That we sum the potentials means in two dimensions that we sum the associated complex functions as well; for opposite charges we change one of the signs.

In this figure, we have depicted two sets of curves: the level curves of the field φ (the equipotentials), and the flow lines of the gradient field $F = \nabla\varphi$ (the lines of force).

We can formulate this as a theorem; compare to Theorem 5.45 regarding analytic functions:

Theorem 5.52. *For an electrostatic field $F = (P, Q)$ on the plane, then P and Q are harmonic conjugates, whence*
(i)their gradient vector fields are orthogonal;
(ii) their families of level curves are orthogonal.

Further, the potential P and dual potential Q are (perhaps integral) linear combinations of the log *and argument (angle) functions on $\mathbb{R}^2$. The corresponding analytic functions are (integral) linear combinations of the complex log function.*

Proof. For a finite combination of point charges at points $\mathbf{p}_i \in \mathbb{R}^2$ with charges $q_i \in \mathbb{R}$, the associated analytic function on $\mathbb{C}$ is $f(z) = \sum q_i \log(z - z_i)$ where $\mathbf{p} = (x, y)$ corresponds to $z = x + iy$.

For a charge density given by a Riemann integrable real-valued function q, the associated analytic function is the vector-valued integral version of this (see §**??**):

$$f(z) = \int_{\mathbb{R}^2} q(w) \log(z - w) \mathrm{d}x\mathrm{d}y.$$

(The more general measure version of this also holds).

□

At first we may think that a potential such as the hyperbola shown in Fig. 21, cannot come from an electrostatic field. However as Feynman Vol II §7.3 [FLS64] points out, it can (in the limit): the field in the exact middle of two opposite charges of Fig. 25 looks just like this. See Figs. 32, 33.

To prove this rigorously, take instead the charges at $-1, 1$ so now $f(z) = \log(z + 1) + \log(z - 1)$. The Taylor expansion of this about the middle point 0 is $-z^2 + \dots$ and $g(z) = z^2$ is indeed the analytic function of Fig. 21. g determines harmonic conjugates which define the linear vector fields depicted there.

Theorem 5.53. *For gravitational fields in $\mathbb{R}^2$, we have the same statement as Theorem 5.52 except that now only positive values of the density function q can occur.*

In fact, according to Feynman, any harmonic function and hence any complex analytic function can occur for a physical electrostatic field in $\mathbb{R}^2$. One can prove this as follows.

From the mathematical point of view, there are two equivalent ways to characterize an electrostatic field (in $\mathbb{R}^2$ or in $\mathbb{R}^3$). The first is that the potential of the field is a solution of *Poisson's equation*,

$$\nabla^2(\varphi) = \rho$$

where ρ is a signed measure describing the distribution of charge. From this point of view one can then go about solving this linear partial differential equation. The second is to describe the *fundamental solution* , which is the single point charge, with its associated (gradient) field, and then define an electric field to be a (vector-valued integral) linear combination of such fundamental solutions, integrated with respect to the charge density.

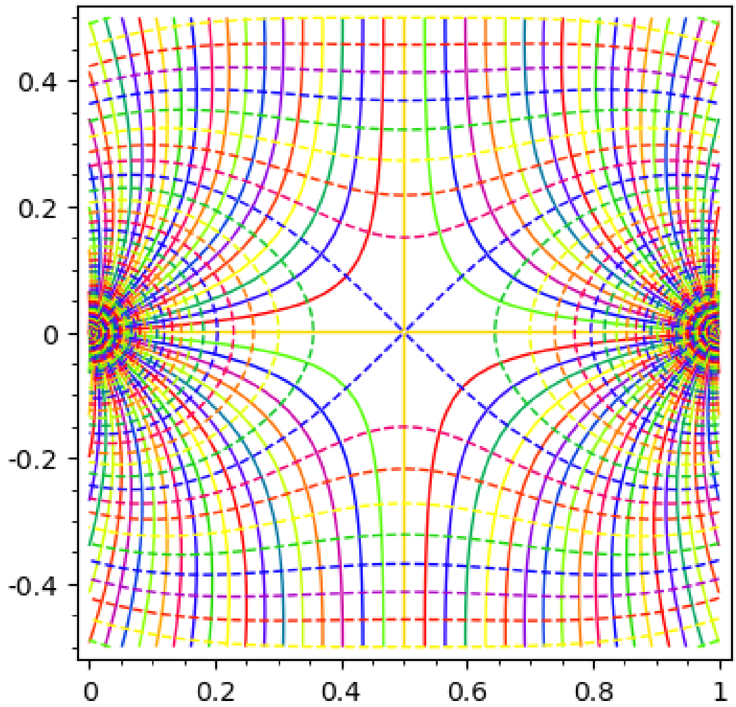


FIGURE 32. Equipotential curves and lines of force for the electrostatic field of two like charges in the plane. Close to the center, $(1/2, 0)$, the potential and its dual start to approximate the dual hyperbolas of Fig. 21.

From the first point of view what is fundamental is the PDE, from the second what is most basic is the fundamental solution (this is Coulomb's law!). What bridges the two is the *superposition principle*, which simply says the space of solutions is a vector space: we can take linear combinations.

In other words, for this linear equation knowing the fundamental solution characterizes the infinite-dimensional vector space of all solutions. And conversely, one of the methods for solving the PDE is to find its fundamental solution.

(For gravity the solution space is not all of the vector space but rather the positive cone inside of it).

Now $\nabla(\varphi) = F$ is the field, so Poisson's equation states that

$$\nabla^2(\varphi) = \text{div}(F) = \rho.$$

Thus from the field or the potential we can determine the charge distribution. Applying the operator ∇ is a type of derivative; the opposite procedure is a type of integration. Thus given the charge density ρ we find the field by solving the (partial) differential equation $\text{div}F = \rho$, and given the field we find the potential by solving the PDE

$$\nabla\varphi = F.$$

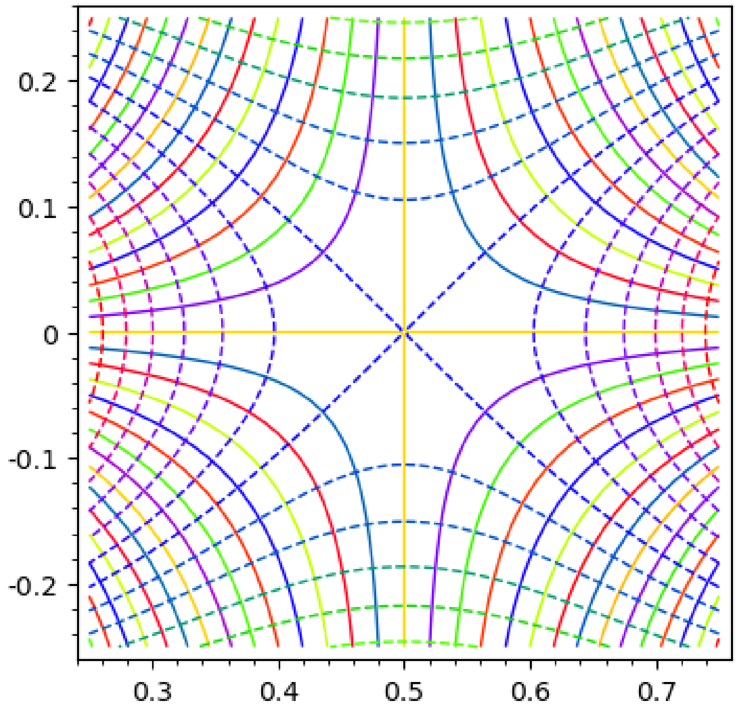


FIGURE 33. Equipotential curves and lines of force for the electrostatic field of two like charges in the plane. Close to the center, $(1/2, 0)$, the potential and its dual start to approximate the dual hyperbolas of Fig. 21.

Combining these, given ρ we can find φ by solving the PDE

$$\nabla^2\varphi = \rho,$$

which is now a *second order* PDE as it involves second order partials.

The general operation of solving a DE is referred to as *integration*. As always, differention is automatic, while integration can be hard! Mathematically speaking, the first task is to prove that under certain circumstances a solution exists, and conversely trying to identify any *obstructions* to having a solution. Such obstructions are often especially interesting because they are topological; e.g. the equation $\nabla\varphi =$ only has a solution on a simply connected $\mathcal{U} \subseteq \mathbb{R}^2 \setminus \{\mathbf{0}\}$.

If there is no charge in a region $\mathcal{U}$, then from Poisson's equation

$$\nabla^2(\varphi) = \rho = 0$$

and the potential function φ is harmonic. Thus for Figs. 25, 25, the potential is 0 everywhere except exacly at those two points. At those points themselves the potential is infinite and the field is not only infinite but points in all directions, so neither is defined. When we have a continuous charge density, however, these are defined everywhere. In that case, by Poisson's equation the potential is not harmonic as $\nabla^2(\varphi) = \rho \neq 0$. When the charge density is continuous but nonzero, the field and potential make perfect sense mathematically being continuous functions, but the potential is no longer a harmonic function, so it certainly cannot (in R^2) have a harmonic conjugate and does not extend to a comlpex analytic function. Hence the tools of Complex Analysis are not as applicable. Nevertheless, there is still a dual potential, whose level sets are orthogonal to those of φ, similar to the harmonic case.

To prove this, (I believe and would like to work this out!) we can again refer to the fundamental solution; since it hold there it must extend to all densities ρ.

But what "is" a point charge? From the mathematical point of view it is a point mass, simply a measure concentrated at a point. In physics this is called a *Dirac delta function*, which is the viewpoint of Riemann-Stieltjes integration. From the standpoint of Lebesgue integration, it is a measure and not a function at all.

Then we know how to rigorously treat two cases: point masses and continuous densities. Similarly, one can include densities given by any other Borel measures.

I say "density" rather than "distribution" here because that word will immediately get used in a very different way! That is the yet more sophisticated viewpoint of Laurent Schwartz' theory of distributions, see e.g. [Rud73]. Roughly speaking a Schwartz distribution is a continuous linear functional defined on a carefully chosen space of *test functions* which are smooth and rapidly decreasing. This enables one to define derivatives, by duality. Thus the advantage of Schwartz distributions is that they can be differentiated and also can be convolved. Thus if one finds a fundamental solution to be a Schwartz distribution, the general solution is found by convolving this over the density. This is exactly what we have described above.

For the simplest case of the fields described above we can get away with point masses, but for more sophisticated examples we really do need Schwartz distributions. This is the case when we consider *dipoles*, but that is beyond the present scope.

For a clear overview of the physics, see the beginning of Jackson's text [Jac99]; this however goes quickly into much (much) deeper material, including boundary values, dipoles, Green's functions, and magnetism, dynamics and the connections with Special Relativity.

For a remarkable mathematical treatment see Arnold's book on PDEs: [Arn04].

Now to sketch a proof of Feynman's claim, given a harmonic function, we define a field to be the gradient of this potential. Given a field, we find such a potential. ...

Finding a potential

Next we see (by working out some examples) how to find the potential of a conservative vector field.

We know that given $F : \mathbb{R}^n \to \mathbb{R}$, the gradient vector field is orthogonal to the level hypersurfaces (submanifolds of dimension $n-1$) of F, so level surfaces in $\mathbb{R}^3$ and level curves in $\mathbb{R}^2$.

We know that a vector field is conservative iff it is the gradient of a function. There are two ways that this can fail to be the case: locally or globally.

We want to first examine the local problem: when is a vector field locally conservative?

Switching equivalently to the language of differential forms, the vector field V is conservative iff the associated $1-$form η is *exact*. We know that a necessary condition for this to occur is that the form be *closed*, i.e. that $d(\eta) = \mathbf{0}$. In $\mathbb{R}^2$ or $\mathbb{R}^3$ this is the same as curl $= \mathbf{0}$.

Poincaré's Lemma tells us that locally, the converse holds: any closed form is exact. A basic counterexample for the global exactness is the angle function on the plane: there is a local potential (the infinite spiral staircase) but this is a multivalued function, so not a potential in the usual sense.

Here is a method to try to find a potential for any vector field in the plane. Given a nowhere-$\mathbf{0}$ vector field V, we want to find a potential φ, that is a function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ such that $\nabla\varphi = V$. In this case, its level curves are orthogonal to V. So, let us consider the orthogonal vector field W to V, say at angle $+\pi/2$. Then, using the Fundamental Theorem of ODEs, draw the integral curves. These are unique hence do not intersect. Globally, they might say be spirals, we can define a function φ with different values on each. Thus, φ is a candidate for a potential.

We can see an example in the illustrations of the electrostatic potentials Fig. 31, 25.

There are two families of curves: the equipotentials and the lines of force.

The lines of force are tangent to the gradient vector field. For opposite charges, we can picture the gradient flow as flowing from the positive to the negative charge. In fact, we can interpret this as a gravitational field, with a mountain at the positive and a valley at the negative charge. For like charges, we can picture two mountains.

It is important to remember that there are two quite different interpretations, as force fields or as velocity fields. The gradient flow refers to the velocity field, and a particle moves along the curve with that tangent vector. For the force field interpretation, the particle accelerates and may go off the curve because of the acceleration due to the curvature.

In any case, we can try to imagine switching roles, so the equipotential curves become the orbits of a gradient flow and vice-versa.

If this works, we will have succeeded in constructing a potential for our vector field. ....

5.16. **The role of differential forms.** (To DO)......

6. Ordinary differential equations

6.1. **The classical one-dimensional case. 1. Introduction:**

Our notes for this section are based in part on course notes by Marina Talet, Université Aix-Marseille, 2021.

Consider an integration exercise from Calculus such as: given $f(x) = 1/x$, defined on $\mathbb{R} \setminus \{0\}$, find

$$F(x) = \int f(x)\mathrm{d}x$$

with the solution

$$F(x) = \log|x| + c$$

The function F has several names: the *antiderivative*, *integral*, or *primitive* or *indefinite integral* of f, and the operation of finding F is termed *integration* of the function f.

We can rewrite this as: find $y = y(x)$ where

$$y' = f$$

which can be considered as the simplest type of differential equation. If we specify that $y(1) = 0$, this *initial condition* fixes the solution on the interval $(0, +\infty)$, as then $c = 0$.

The explanation for the integral being *indefinite* in that it is defined up to a constant c is that the derivative map $D : \mathbb{C}^{k+1} \to \mathbb{C}^k$ is a linear transformation with one-dimensional kernel the constant functions.

The general concept of *integrating* or *finding a primitive* goes far beyond this.

We find further examples in what follows.

Exponential growth.

Let us recall the two main rules for exponents:

(i)$a^{b+c} = a^b a^c$

and

(ii)$a^{bc} = (a^b)^c$.

An *exponential function* is of the form $f(x) = a^x$ for $a > 0$ and $x \in \mathbb{R}$. The number a is termed the *base* and x the *exponent*. By the above rules, $1/a = a^{-1}$ and $(a^{\frac{1}{2}})^2 = a$ whence $\sqrt{a} = a^{\frac{1}{2}}$. This makes it easy to define a^x for exponent rational.

Thus exponentiation turns adddition into multiplication, multiplication into taking powers. Mathematically speaking, the fact that $a^{x+y} = a^x a^y$ and $a^0 = 1$ tells us that writing $\Phi_a(x) = a^x$, the map Φ is an isomorphism from the additive group of real numbers $(\mathbb{R}, +)$ to the multiplicative group of positive real numbers $(\mathbb{R}^> \cdot)$ where $\mathbb{R}^> \equiv (0, +\infty)$.

We write $\ln_a(x)$ for the inverse function of base a, that is, for $f(x) = a^x$ and $g = \ln_a(x)$ then $f \circ g(x) = x$ and $g \circ f(x) = x$ wherever these are defined: the first for $x > 0$ and the second for all $x \in \mathbb{R}$.

This function does the opposite of the exponential: it maps $\mathbb{R}^> \equiv (0, +\infty)$ to $(\mathbb{R}, +)$ and converts multiplication to addition and powers to products, thus

$$\ln_a(xy) = \ln_a(x) + \ln_a(y)$$

and

$$\ln_a(x^y) = y \ln_a(x).$$

The most practical base in many applications is 10 or 2, but in pure mathematics by far the most important base is the irrational number $e = 2.1714...$ The reason is that the formula for the derivative is much simpler for base e, as we shall see shortly.

But first, to define e^x for real, non-rational exponents, we note that there are several approaches one can take.

(1)First, we can use continuity: it can be proved that there is a unique continuous way to extend this function to the reals. That is, we can approximate x by rational numbers and take the limit.

We can also give more explicit definitions:

(2) From the Fundamental Theorem of Ordinary Differential Equations, there exists a unique solution to the *differential equation* (or DE) $y' = y$ satisfying $y(0) = 1$ (this is called an *initial condition* for the DE). We define this function to be $\exp(t) = y(t)$, and then and define the number e to be $\exp(1)$. This is also the slope of e^x at $x = 0$.

(3) We define e^x to be the function with the following series expansion:

$$\exp(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots + \frac{x^n}{n!} + \dots \tag{38}$$

where the factorial is defined by $0! = 1$, $k! = 1 \cdot 2 \cdot \dots \cdot k$.

This expresses e^x as a limit of rational numbers. In particular, the number $e = e^1$ can be approximated as a decimal using this series.

We need:

Convergence of the Tayor series for the exponential function.

Theorem 6.1. *The series for e^x in* (38) *converges for all $x \in \mathbb{R}$. The corresponding series where x is replaced by a complex number $z \in \mathbb{C}$ convergse, as does the series where x is replaced by a square matrix.*

Proof. For x fixed, let $m > 2x$ so $x/m < 1/2$. Then for any $n > 0$,

$$\frac{x^{n+m}}{(n+m)!} \leq \frac{x^m}{m!} \cdot \left(\frac{1}{2}\right)^n$$

which gives a geometric series hence converges. Thus the sequence of partial sums is an increasing bounded sequence hence converges by the completeness property of the real numbers.

Similarly, using the fact that $|zw| = |z||w|$, the series

$$\exp(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \cdots + \frac{z^n}{n!} + \dots$$

comverges for all $z \in \mathbb{C}$.

For square matrices we need the follwing: We define a norm on the linear space $L(V, W)$ of all linear operators from one normed space V to another W by $||A|| = \sup_{\mathbf{v} \in V} ||A\mathbf{v}||/||\mathbf{v}|| = \sup_{||\mathbf{v}||=1} ||A\mathbf{v}||$. This is called the **operator norm**. One of its main useful properties is that it (clearly) behaves nicely under composition: $||AB|| \leq ||A||\,||B||$. This is called **submultiplicativity**.

In particular this holds for squares matrices A, B. Using this, just as for complex numbers, the series

$$\exp(A) = 1 + A + \frac{A^2}{2} + \frac{A^3}{6} + \cdots + \frac{A^n}{n!} + \dots$$

comverges for all $A \in \mathcal{M}_{(d\times d)}$, the collection of square matrices (with entries in $\mathbb{R}$ or $\mathbb{C}$.)

□

Note that the derivative of the series for $f(x) = e^x$ taken term-by-term, does satisfy the DE $y' = y$, $y(0) = 1$, so (3) yields (2). Conversely, knowing the derivative from (2) gives (3) as the Taylor series: recall that for an infinitely differentiable function f the Taylor series about 0 is

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(0)x^k}{k!}.$$

We let $\ln(x)$ denote the *natural logarithm*, the inverse function $g(x)$ of $f(x) = e^x = \exp(x)$, so $\ln = \ln_e$.

Another way to define ln is via integration: for $x > 0$,

$$\ln(x) = \int_1^x \frac{1}{x} \mathrm{d}x$$

Another possible definition for ln is via its Taylor series, calculated around the value $x = 1$, in other words, find the series for $\ln(x + 1)$ around 0.)

This leads to a third definition of exp:

(4) First we define ln, in one of these ways; then exp is defined to be its inverse function.

Next we define, for any base $a > 0$:

$$a^x = e^{(\ln a)x}.$$

Hence the derivative is $(a^x)' = \ln(a)a^x$. Note that indeed $(e^x)' = e^x$, and the number e is the only base such that a^x is its own derivative.

We then denote by $\ln_a$ its inverse function.

Exponential growth and doubling times; exponential decay and half-life

Suppose a quantity $f(n)$ doubles every day, starting at 1 at time $n = 0$. Then we have $f(n) = 2^n$. (You should draw the graph, for say $n = -3, \dots 3$). Here the *doubling time* is 1.

When we first see this equation, we naturally wonder why not to use base 2 (or perhaps 10!) instead of the irrational number $e = 2.1714\dots$. The reason is because base e has the simplest expressions for the derivative, hence also for the series. In fact, for both calculations and theory, for exponential and also for logs, it is generally easier to first change to base e.

Nevertheless the concept of doubling time is intuitively very useful.

When a^t for $a < 1$ we call this *exponential decay*, for example the decay of radioactivity of a substance.

When we know the doubling time of for instance a pandemic or a bank account, we can easily make rough estimates in our heads, and similarly for exponential decay of a radioactive substance.

These can vary considerably, ranging from 4.4 billion years for Uranium-238 to 10^{-24} seconds for Hydrogen-7. Plutonium-239 has a half-life $24,110$ years, indicating its danger when in radioactive waste, while Carbon-14 which is so useful in the radiocarbon dating process used by archeologists has a half-life of $5,730$ years.

Exercises:
(1) Show that $\ln_a(x) = \ln(x)/\ln(a)$.
(2) Find the Taylor's series for $\ln(x+1)$ about $x = 0$.
(3) Prove that the series for e^x converges for all $x \in \mathbb{R}$.
(4) Find a formula for the doubling time t_d for $f(t) = e^{at}$, for $a > 0$.
(5) Find the half-life t_h half-life for $f(t) = e^{at}$, for $a < 0$.

Solving the equation $y' = ay$, $a \in \mathbb{R}$.

Exponential growth $y(t) = A^t$ for $A > 1$ grows at a rate proportional to the quantity at time t. Thus for example for $A = 2$, $2^{n+1} - 2^n = 2^n(2-1) = 2^n$, while for a bank account growing at 10% per year, $a = 1.10$ and $y(t+1) - y(t) = A^{n+1} - A^n = A^n(A-1) = c \cdot y(n)$ for the constant $c = A - 1$.

As noted, this includes both exponential growth or decay, and also the constant case $y' = 0$.

We simplify to the special case $a = 1$ and recall that $y(t) = e^t$ solves this. Then we see that $y(t) = Ke^t$ for $K \in \mathbb{R}$ also works. Lastly we note that $y(t) = Ke^{at}$ will provide a solution of the DE $y' = ay$, for any $a \in \mathbb{R}$.

This is valid for any $a, K \in \mathbb{R}$.

But are these all possible solutions? To answer this let $v(t) = e^{at}$ and suppose that u is another solution, so $u' = au$. Now since $v(t) = e^{at}$, $v^{-1} = e^{-at}$ whence $(v^{-1})' = -av^{-1}$.

We guess that $u = Kv$ is the only possibility, thus that u/v will be a constant. Equivalently, its derivative is 0. We compute:

$$(u/v)' = (u \cdot v^{-1})' = u'(v^{-1}) + u \cdot (v^{-1})' = auv^{-1} - auv^{-1} = 0$$

as we guessed, so $u/v = K$ and

$$u = Kv = Ke^{at}.$$

1.b. Analytic definition of a differential equation.

It is time for a definition! Basically, differential equation (denoted DE) is a relation between an unknown function y (to be determined) and a certain number of its derivatives. More precisely,

Definition 6.1. A *differential equation in one dimension* is an equation of the form

$$(*) \qquad F(t, y(t), y'(t), \cdots, y^{(n)}(t)) = 0,$$

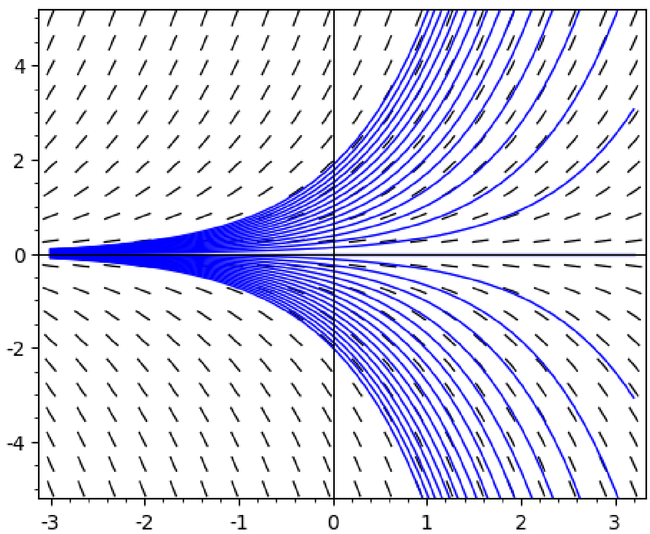


FIGURE 34. Slope field and solution curves for exponential growth $y' = cy$. The equation is in one dimension, and its flow is along the real line; these curves are the *graphs* of those solutions, so including the time variable. This can also be viewed as solutions to a vector ODE in the plane, where the curves are tangent to the vector field $V(x, y) = (1, y)$. These solution curves are $\gamma(t) = (t, y(t))$ so $\gamma'(t) = (1, y'(t)) = V(\gamma(t)) = (1, y(t))$. The difference between a *slope field* and a vector field is this: segments in the slope field are parallel to the vector field but meet the curves in their midpoint. The picture of the slope field is often easier to understand, as it is much less cluttered since all the segments are all of the same maneagable length.

where F is a given real-valued continuous function of $n + 2$ variables (so at least one derivative is involved!) and $t \mapsto y(t)$ is a *unknown* function, that we are trying to find. We denote by $y^{(n)}$ its derivative of order n, a strictly positive integer.

The above DE (*) is then said to be **order** n.

It is a *linear DE* iff the function $F : \mathbb{R}^{n+2} \to \mathbb{R}$ is affine (unfortunately, this is the standard terminology!). Thus the equation $y' = -y + c$ is called linear.

It is a *autonomous* or *stationary DE* if there is no variable t for F, so F is of the form

$$(*) \qquad F(y, y', \cdots, y^{(n)}) = 0$$

Otherwise it is a *non-autonomous* or *non-stationary* or *time-varying* DE.

The DE is said to be **explicit** of order n, or **in normal form**, if it can be solved for the highest-order derivative, in other words if can be written in the form

$$(*) \qquad y^{(n)}(t) = F(t, y(t), y'(t), \cdots, y^{(n-1)}(t))$$

where now F is a given continuous function of $n + 1$ variables. Otherwise it is an **implicit** DE.

To motivate this terminology, recall that in Calculus the equation $x^2 + y^2 = 1$ is equivalent to the four equations $y = \pm\sqrt{1-x^2}$, $x = \pm\sqrt{1-y^2}$ and we can say that the first equation is an *implicit* equation in that it "implies" the other "explicit" equations where we have solved for one variable as a function of the other. For $\mathbb{R}^n$, the Implicit Function Theorem can help us determine when this can be done. In much the same way, we can have an implicit DE, for example $(y')^2 + y^2 = 1$ which implies the explicit equations $y' = \pm(1-y^2)$.

Here are some examples:

(1)For $a : \mathbb{R} \to \mathbb{R}$ continuous, $y' = a(t)$ is an explicit equation, nonautonomous unless $a(t) = a$ is constant. The solution is just the antiderivative of $a(t)$.

(2)$F(r,s) = s$, so $y' = F(t,y) = y$: $y' = y$. This is the autonomous, linear, first order equation we encountered above.

(3)For $a, b \in \mathbb{R}$, $F(r,s) = a \cdot s + b$, so $y' = ay + b$. This is an *autonomous linear* DE.

(4) For $a, b : \mathbb{R} \to \mathbb{R}$ continuous, $F(r,s) = a(t) \cdot s + b(t)$, so $y' = a(t)y + b(t)$. This is a linear first-order nonautonomous DE.

More examples:
$y'(t) + y(t) = 1$: is an DE of first order
$\sin(y'(t)) + y^3(t) = t$: an implicit DE of first order
$y^{(7)}(t) + y^9(t) = t + \sin(5t)$: an DE of order 7
$y(t) + y^2(t) = t$ is not an DE (as there is no derivative involved!).

A **solution** of (*) over an interval I of $\mathbb{R}$ is a function $t \mapsto y(t)$ which is n times derivable on I and which satisfies (*).

The interval I on which we solve a DE is very important, as changing the interval may allow for different solutions.

To **solve** (*) means to find **all** the solutions of (*).
There can be zero, one, several or an infinite number of solutions.

Examples:
- The DE $y'(t) = 0$ admits an infinite number of solutions. Indeed, $y(t) = c$ for any $c \in \mathbb{R}$ is a solution.
- The implicit DE $y'^2(t) + 1 = 0$ does not admit any real solution.

6.2. **Flows, systems of DEs and vector differential equations.** At this point it will actually be better to consider an apparently more difficult problem: that of DEs in higher dimensions, or *vector differential equations.*

Definition 6.2. Given a vector field V on $\mathbb{R}^n$, a *vector differential equation* of first order with *initial condition* $\mathbf{x} \in \mathbb{R}^n$ is:$\gamma(0) = \mathbf{x}$. is:

$$\gamma'(t) = V(\gamma(t);\ \gamma(0) = \mathbf{x}$$

This is equivalent to a *system* of n first order DEs. For example, taking $n = 2$ so $V = (P, Q)$, then for $\gamma = (x, y)$,

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} P(x,y) \\ Q(x,y) \end{bmatrix}$$

or equivalently

$$\begin{cases} x' = \\ y' = \end{cases} \tag{39}$$

In §§**??** and 5.4 we have already given an introduction to this topic.

We summarize what was said there.

Definition 6.3. Given a set X and a function $\tau : X \times \mathbb{R} \to X$, note that fixing $t \in \mathbb{R}$ then $\tau_t(x) = \tau(x,t)$ defines a map $\tau_t : X \to X$. Thus $\{\tau_t\}_{t\in\mathbb{R}}$ is a collection of maps on X.

We say τ defines a *flow* on X iff
(i) τ_0 is the identity map and
(ii) τ_t satisfies the *flow property*

$$\tau_{t+s} = \tau_s \circ \tau_t.$$

A flow is also known as a *one-parameter group of transformations*.

We think of the variable t as *time*; then τ_t is called the *time-t map* of the flow.

The *orbit* of a point $x \in X$ is $\{\tau_t(x) : t \in \mathbb{R}\}$.

A flow is a *continuous-time dynamical system* as it describes the time evolution of a point $x \in X$ as it moves along its orbit. The *future* orbit is the collection of points $\tau_t(x)$ for $t > 0$, the *past* orbit for $t < 0$, and the present moment is $\tau_0(x) = x$, where we are right now. It is assumed that the system has only one past and future; there is no randomness here. In other words, two orbits are either disjoint or identical. (This is the case for any physical system except possibly quantum mechanics, for example one obeying Newton's laws; however, even in such a "predictable" system, randomness can enter in a different way because of complicated dynamics: *chaos* or *sensitive dependence on initial conditions*.)

If X is a vector space V, then τ_t is a *linear flow* iff each map τ_t is linear. Note that by the flow property plus the fact that τ_0 is the identity, τ_t is bijective as its inverse is τ_{-t}. Thus each τ_t is a linear isomorphism of V.

Example 18. **(Rotation Flow** 1**)**

Consider the group of rotations of the plane, setting $a = \cos(2\pi t), b = \sin(2\pi t)$ and $R_t = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$ and defining $G = \{R_t : t \in \mathbb{R}\} \cong \mathbb{T}^1 = \mathbb{R}/\mathbb{Z}$.

Noting that $R_{t+s} = R_s \circ R_t = R_t \circ R_s$, this is a flow.

The next result will show that these maps are of the form e^{tA}, for a certain matrix A, making the connection with Lie algebras. As we shall see, in fact all linear flows arise in this way.

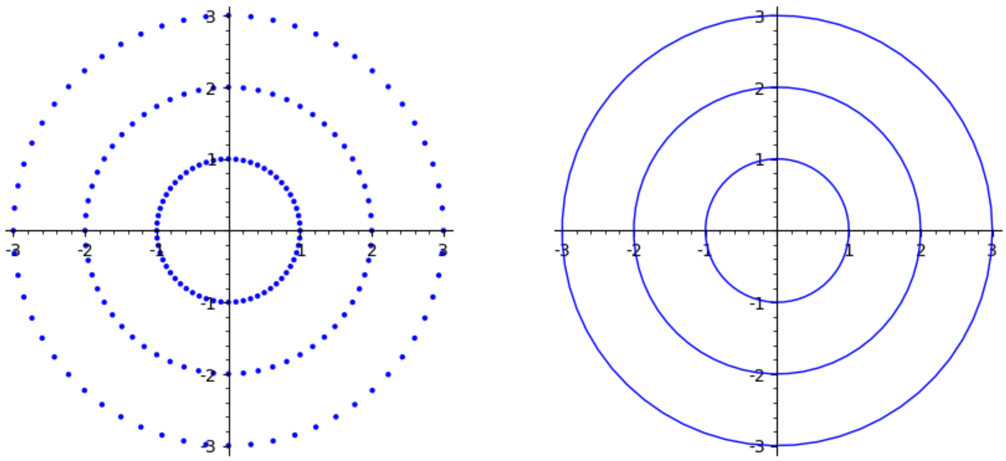


FIGURE 35. Some discrete and continuous-time orbits of the rotation flow, Example 6.2.

Proposition 6.2. *(Rotation flow) For* $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, *then*

$$e^{tA} = R_t = \begin{bmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{bmatrix}.$$

Proof. (First proof) For $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ we note that the powers of A have period 4, with $(A^0, A^1, A^2, A^3, \dots) = (I, A, -I, -A, \dots)$. We separate the Taylor series into even and odd terms. Writing $c = \cos t$ and $s = \sin t$, this gives:

$$\begin{aligned}
&\exp(tA) = \sum_{k=0}^{\infty} (tA)^k/k! = \\
&I + tA + (tA)^2/2 + (tA)^3/3! + (tA)^4/4! + \cdots = \\
&(I + (tA)^2/2 + (tA)^4/4! + \dots) + (tA + (tA)^3/3! + (tA)^5/5! + \dots) = \\
&I(1 - t^2/2 + t^4/4! - t^6/6! + \dots) + A(t - t^3/3! + t^5/5! - \dots) = \\
&\begin{bmatrix} c & 0 \\ 0 & c \end{bmatrix} + A \begin{bmatrix} s & 0 \\ 0 & s \end{bmatrix} = \begin{bmatrix} c & 0 \\ 0 & c \end{bmatrix} + \begin{bmatrix} 0 & -s \\ s & 0 \end{bmatrix} = \begin{bmatrix} c & -s \\ s & c \end{bmatrix}
\end{aligned} \tag{40}$$

as claimed. □

Exercise 6.1. (Harmonic oscillator)

Perhaps the second most important example is that of the the harmonic oscillator, with the DE

$$y'' = -y.$$

The idea is that a spring is attached to a wall and the other end to an object; when this is pulled out to a distance y the force felt is approximately $-cy$, where $c > 0$ is a constant called the *spring constant*. By Newton's Law $F = ma$, this gives (taking

$m = 1$) the above equation. The reason for the $-$ sign is that it is being pulled back toward its rest position 0: in the negative direction if if $y > 0$, in the positive direction if $y < 0$, changing as it oscillates.

Ir is clear that (taking for simplicity $c = 1$) $y(t) = \sin t$ and $\cos t$ are solutions.

To find all the solutions, we show how this second-order equation in one dimension can be rewritten as a system of two one-dimensional first-order equations, and equivalently as a vector DE in $\mathbb{R}^2$, given by a linear vector field. *This is always possible:* a single higher-order DE in one dimension can always be rewritten as an equivalent order-one vector DE (also in the nonautonomous case, by adding one more dimension). Geometrically, this means we have a velocity vector field, with an integral curve exactly corresponding to a solution!

To carry this out in this case, we set $w_1 = -y, w_2 = y'$. We thus have the pair of equations $w_2' = y'' = -y = w_1,$, $w_1' = -y' = -w_2$ giving the system

$$\begin{cases} w_1' = -w_2 \\ w_2' = w_1 \end{cases}$$

This can be written in matrix form, where $\mathbf{w} = (w_1, w_2)$, as

$$\mathbf{w}' = A\mathbf{w} \tag{41}$$

so

$$\begin{bmatrix} w_1' \\ w_2' \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}.$$

Definition 6.4. Equation (41), $\mathbf{w}' = A\mathbf{w}$ where A is an $(n \times n)$ matrix, and $\mathbf{w} = \mathbf{w}(t) \in \mathbb{R}^n$ is called a *vector DE*. Note that A defines a linear vector field on $\mathbb{R}^n$, and $\mathbf{w}(t)$ is a curve in $\mathbb{R}^n$; it is a curve which is tangent to the vector field, as its tangent vector at each point is the vector field. If the starting point of the curve $\mathbf{w}(0) = \mathbf{w}_0$ is specified, then we have a vector DE with this *initial condition*.

Remark 6.1. Now A defines a linear vector field on $\mathbb{R}^2$. Since the variable for time does not occur here, we had for $y'' = -y$ ian autonomous second-order equation, and now we have an autonomous vector DE, equivalently an autonomous system of two first-order equations. Physically, the variables $(w_1, w_2) = (-y, y')$ represent, for the oscillator, $-$position and velocity (or momentum, since momentum$= mV$). The vector solution $\mathbf{w}(t) = (\cos t, \sin, t)$ gives the one-dimensional solution $y(t) = \cos t$ for the original equation; the graph of the curve $\mathbf{w}(t)$ is the helix $(t, \cos t, \sin t)$ which projects to the position $y(t) = \cos t$ and velocity $y'(t) = \sin t$.

In Physics, we often pass to (position, momentum) soace which is called *phase space.*

Exercise 6.2. Solve the second order linear equation $y'' = -y$ by the following strategy: we define $y' = x$ and $x' = -y$, giving a system of two equations of first order. Then we rewrite the system in vector form $\mathbf{x}' = A\mathbf{x}$, for $\mathbf{x} = (x, y)$ as above.

Now solve this vector DE explicitly for initial condition $\mathbf{x}_0 = (a, b)$ and sketch the solutions. Lastly, returning to the original equation $y'' = -y$, what are the solutions $y(t)$?

Solution.

The matrix is $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. We have the solution

$$e^{tA}\mathbf{x}_0 = R_t\mathbf{x}_0 = \begin{bmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} a\cos t - b\sin t \\ a\sin t + b\cos t \end{bmatrix} = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}$$

for the vector DE with initial condition $\mathbf{x}_0 = (a, b)$. The general solution for the one-dimensional second order equation $y'' = -y$ is therefore

$$y(t) = a\sin t + b\cos t. \tag{42}$$

Note that $x(0) = a = y'(0)$, so the initial condition is $y(0) = b, y'(0) = a$. Physically, this corresponds to a harmonic oscillator with mass and spring constant 1, and with initial position $y(0) = b$, initial velocity $y'(0) = a$.

Fixing $y(0) = b$, we see all the circles which meet the line $y = b$ in the plane, each corresponding to a different initial velocity.

The general linear higher-order linear case.

Any higher-order linear DE can be handled similarly. Thus, the matrix for the general n^{th}-order linear case

$$y^{(n)} = a_1 y + a_2 y' + \cdots + a_n y^{(n-1)}$$

then has the nice form

$$A = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & \\ \vdots & & & & \\ 0 & 0 & 0 & \dots & 1 \\ a_1 & a_2 & a_3 & \dots & a_n \end{bmatrix}.$$

References

[Ahl66] Lars V. Ahlfors. *Complex Analysis*. McGraw-Hill, second edition, 1966.

[Arm83] M.A. Armstrong. *Basic Topology*, volume 8 of *Undergraduate Texts in Mathematics*. Springer, 1983.

[Arn04] Vladimir Igorevc Arnold. *Lectures on partial differential equations*. Springer, 2004.

[Axl97] Sheldon Jay Axler. *Linear Algebra Done Right*, volume 2. Springer, 1997.

[CB14] Ruel Churchill and James Brown. *Ebook: Complex Variables and Applications*. McGraw Hill, 2014.

[DC16] Manfredo P Do Carmo. *Differential geometry of curves and surfaces: revised and updated second edition*. Courier Dover Publications, 2016.

[FLS64] RP Feynman, RB Leighton, and M Sands. The feynman lectures on physics, ii, addisonr-wesley. *Reading, Massachusetts*, 1964.

[GP74] Victor Guillemin and Alan Pollack. *Differential topology*. Prentice-Hall, 1974.

[Gui02] Hamilton Luiz Guidorizzi. Um Curso de Cálculo, Vols I- III, ltc, 5a, 2002.

[HH15] John H Hubbard and Barbara Burke Hubbard. *Vector Calculus, Linear Algebra, and Differential Forms: a unified approach*. Matrix Editions, 2015.

[Jac99] John David Jackson. Classical electrodynamics, 1999.

[Lan99] Serge Lang. *Complex Analysis*, volume 103 of *Graduate Texts in Mathematics*. Spriger, 1999.

[Mar74] Jerrold E. Marsden. *Elementary Classical Analysis*. W. H. Freeman, 1974.

[MH87] Jerrold E. Marsden and Michael J. Hoffman. *Basic Complex Analysis*. W. H. Freeman, second edition, 1987.

[MH98] JE Marsden and JM Hoffman. Basic complex analysis, 3ª edição, 1998.

[Mil16] John Milnor. Morse theory.(am-51), volume 51. In *Morse Theory.(AM-51), Volume 51*. Princeton university press, 2016.

[MW97] John Milnor and David W Weaver. *Topology from the differentiable viewpoint*, volume 21. Princeton university press, 1997.

[O'N06] Barrett O'Neill. *Elementary differential geometry*. Elsevier, 2006.

[Ros68] Maxwell Rosenlicht. Liouvilles theorem on functions with elementary integrals. *Pacific Journal of Mathematics*, 24(1):153–161, 1968.

[Rud73] W. Rudin. *Functional Analysis*. McGraw-Hill, New York, 1973.

[Spi65] Michael Spivak. *Calculus on Manifolds*, volume 1. WA Benjamin New York, 1965.

[Str12] Gilbert Strang. *Linear algebra and its applications*. 2012.

[War71] Frank W Warner. *Foundations of differentiable manifolds and Lie groups*, volume 94 of *Graduate Texts in Mathematics*. Springer Verlag, 1971.

Albert M. Fisher, Dept Mat IME-USP, Caixa Postal 66281, CEP 05315-970 São Paulo, Brazil

URL: http://ime.usp.br/∼afisher

E-mail address: afisher@ime.usp.br